Leveraging CRISPR/Cas Genome Editing Technology to Identify and Characterize Causal GWAS Variants for Blood Lipids

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Dr. Kiran Musunuru, for supporting my research and for giving me an invaluable opportunity to learn about human genetics and cardiovascular disease during my time in his laboratory. I am grateful for his scientific and professional guidance. I would also like to extend my gratitude to the members of the Musunuru laboratory for their outstanding mentorship and guidance, particularly Qiurong Ding, Xiao Wang, Nicolas Kuperwasser, Derek Peters, and Tao Chen.

I want to acknowledge our collaborators at the Broad Institute, led by Dr. Tarjei Mikkelsen, whose work was the foundation for many of the experiments performed in this thesis. I would also like to recognize the support of the HST faculty, especially Patty Cunningham and Rick Mitchell. Additionally, I would like to acknowledge funding support from the HST Research Assistantship Program and the Howard Hughes Medical Institute (HHMI) Medical Fellows Program.

ABSTRACT

Genome-wide association studies (GWAS) have identified a number of novel genetic loci linked to serum cholesterol and triglyceride levels. The causal DNA variants at these loci and the mechanism by which they influence phenotype and disease risk remain largely unexplored. Expression quantitative trait locus (eQTL) analyses of patient liver and adipose biopsies indicate that many lipid-associated variants influence gene expression in a *cis*-regulatory manner. However, linkage disequilibrium (LD) among neighboring single nucleotide polymorphisms (SNPs) at a GWAS-implicated locus makes it challenging to pinpoint the actual variant underlying an association signal. Here we performed high-throughput identification of putative disease-causal loci through a functional reporter-based screen, the massively parallel reporter assay (MPRA). We then validated prioritized variants using a combination of genome edited stem cells, clustered regularly interspaced short palindromic repeats (CRISPR) interference, and *in vivo* genome edited humanized mouse models to establish rs2277862-*CPNE1*, rs10889356-*ANGPTL3*, and rs12740374-*SORT1* as causal SNP gene sets. These results highlight a novel experimental framework to discover causal genes and variants contributing to complex human traits.

GLOSSARY

ABCA1	ATP binding cassette transporter A1
BAC	Bacterial artificial chromosome
CETP	Cholesteryl ester transfer protein
CHD	Coronary heart disease
ChIP	Chromatin immunoprecipitation
CRISPR	Clustered regularly interspaced short palindromic repeats
dCas9	Catalytically dead Cas9
EL	Endothelial lipase
EMSA	Electrophoretic mobility shift assay
eQTL	Expression quantitative trait locus
FACS	Fluorescence activated cell sorting
FH	Familial hypercholesterolemia
GFP	Green fluorescent protein
GLGC	Global Lipids Genetics Consortium
GWAS	Genome-wide association study
HDL-C	High density lipoprotein cholesterol
HDR	Homology-directed repair
HL	Hepatic lipase
HLC	Hepatocyte-like cell
hPSC	Human pluripotent stem cell
IDL	Intermediate density lipoprotein
Indel	Insertion-deletion mutation
iPSC	Induced pluripotent stem cell
LCAT	Lecithin:cholesterol acyltransferase
LD	Linkage disequilibrium
LDL-C	Low density lipoprotein cholesterol
LDLR	Low density lipoprotein receptor
LpL	Lipoprotein lipase
MAF	Minor allele frequency
MPC	Mesenchymal progenitor cell
MPRA	Massively parallel reporter assay
MTP	Microsomal transfer protein
NHEJ	Non-homologous end joining
PAM	Protospacer-adjacent motif
PCR	Polymerase chain reaction
PLTP	Phospholipid transfer protein
qPCR	Quantitative polymerase chain reaction
sgRNA	Single guide RNA
SNP	Single nucleotide polymorphism
SR-BI	Scavenger receptor class B type I
ssODN	Single-stranded oligonucleotide
TALEN	Transcription activator-like effector nuclease
TC	Total cholesterol
TG	Triglyceride
UTR	Untranslated region
VLDL	Very low density lipoprotein
ZFN	Zinc finger nuclease

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

Coronary heart disease (CHD) is the principal cause of morbidity and mortality worldwide. About one-half of men and one-third of women are expected to experience a coronary event in their lifetimes.[1] CHD arises secondary to the development of an atherosclerotic lesion in the coronary circulation. Partially obstructive plaques impede blood flow to myocardial tissue, producing clinical symptoms of angina pectoris that may variably manifest at rest or upon exertion. In some cases, sudden plaque rupture precipitates thrombosis and complete arterial occlusion, resulting in myocardial infarction and its pursuant complications including heart failure and arrhythmia.[2,3]

Population-based epidemiologic studies, like the Framingham Heart Study, provide important insights into risk factors for the development and progression of CHD.[4] Abnormal plasma concentrations of low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG) are significant risk factors for CHD. Epidemiologic data consistently point to a positive correlation between LDL-C levels and cardiovascular disease, earning LDL-C the nickname of "bad cholesterol." The success of the statin drugs in decreasing LDL-C levels and diminishing cardiovascular risk support a causal role for LDL-C in CHD pathogenesis.[5] Conversely, HDL-C levels demonstrate an inverse relationship with cardiovascular risk, independent of LDL-C levels, suggesting that HDL-C functions as the "good cholesterol." However, the role of steady-state HDL-C levels in CHD progression is unclear, in light of recent Mendelian randomization studies that have failed to demonstrate a compelling causative link between genetic variants that affect HDL-C and myocardial infarction risk.[6] Moreover, the failure of cholesteryl ester transfer protein (CETP) inhibitors to improve cardiovascular outcomes despite elevating HDL-C levels[7,8] suggests that the relationship between HDL-C and CHD may be more complex than previously believed.[9,10] Like LDL-C, elevated TG levels are associated with increased CHD risk, but whether TG is causal in the pathogenesis of CHD or simply a biomarker of disease coincident with other causal risk factors remains uncertain.[11] In individuals with low LDL-C secondary to statin therapy, TG levels are strongly associated with incident coronary events, suggesting they are an independent predictor of CHD risk.[12]

1.2 Overview of lipoprotein metabolism

Lipoproteins are water-soluble complexes that carry lipids throughout the bloodstream. Circulating lipoproteins are composed of a hydrophobic core with TGs and esterified cholesterol, an outer shell enriched in polar lipids such as phospholipids and free cholesterol, and a characteristic repertoire of apolipoproteins that influence the functional properties of the particle. The major lipoprotein subclasses – chylomicrons, very low-density lipoprotein (VLDL), intermediate density lipoprotein (IDL), LDL and HDL – exhibit characteristic buoyant densities due to variable lipid compositions and lipid to protein ratios.[3]

Apolipoprotein B (apoB) is the principal protein component of chylomicrons, VLDL, IDL and LDL. ApoB-containing lipoproteins participate in lipid delivery, and can be sub-classified into an intestinal apoB-48 lineage (chylomicrons) and a hepatic apoB-100 lineage (VLDL, IDL and LDL). The liver secretes endogenously synthesized TGs as VLDL for distribution throughout the body. In the initial stage of VLDL assembly, full-length apoB-100 is translocated to the lumen of the rough endoplasmic reticulum, where it becomes lipidated by microsomal triglyceride transfer protein (MTP). The pre-VLDL particle is then directed to the Golgi, where it undergoes further lipidation and post-translational modification. Upon secretion, VLDL acquires apoE and apoC-II from HDL. VLDL-TGs are rapidly hydrolyzed by lipoprotein lipase (LpL) in the capillary endothelium to generate free fatty acids that can be utilized by skeletal muscle and adipose for energy and storage, respectively. The interaction of VLDL with LpL is regulated by different components of the VLDL proteome, including apoC-II, which activates LpL, and apoC-III, which inhibits LpL. VLDL is further remodeled by CETP-mediated exchange of VLDL-TGs for cholesteryl ester from HDL. Throughout this remodeling process, apoE, apoC-II and apoC-III are variably shed and eventually re-associate with HDL. The remaining IDL can either be internalized by hepatic receptors in an apoE-dependent process, or undergo further lipolysis by hepatic lipase (HL) to form the TG-poor, cholesterol-enriched LDL particle, which lacks apoE. LDL catabolism is primarily mediated by the hepatic LDL receptor (LDLR), which binds apoB-100 and internalizes LDL via clathrin-dependent endocytosis. Peripheral cells can also internalize LDL in an LDLR-dependent process.[3] Finally, oxidized LDL can infiltrate the arterial intima, where it is phagocytosed by macrophage scavenger receptors to form cholesterol-laden foam cells in the earliest pathologic event of atherogenesis.[2]

In the post-prandial state, dietary TGs are absorbed by intestinal epithelial cells and loaded onto apoB-48, a truncated form of apoB that lacks the C-terminal LDLR-binding domain present on apoB-100. Assembled chylomicrons are secreted into the lymphatics and enter the peripheral circulation via the lymphatic duct. Through LpL and HL-mediated lipolysis, chylomicrons are catabolized to remnant particles that are cleared from the circulation via apoB-independent mechanisms.[3]

Whereas apoB-containing lipoproteins export lipids to peripheral tissues, apoA-I-containing lipoproteins (namely HDL) participate in reverse cholesterol transport, the process by which cholesterol is effluxed from the periphery and returned to the liver for biliary excretion. Lipid-poor apoA-I is secreted by the liver and intestine into the bloodstream, where it serves as an acceptor for ATP binding cassette transporter A1 (ABCA1)-mediated free cholesterol efflux from peripheral tissues. Cholesterol efflux from arterial wall macrophages, although relatively minimal in terms of its percent contribution to total-body cholesterol efflux, is considered a key mechanism by which HDL exerts its atheroprotective effect.[13] Free cholesterol on the surface of the nascent HDL particle is then esterified by lecithin:cholesterol acyltransferase (LCAT). The mature, cholesteryl-ester enriched HDL particle can engage in exchange with TG-rich lipoproteins via CETP. Additionally, through the action of phospholipid transfer protein (PLTP), HDL acquires phospholipids derived from lipolysis of TG-rich lipoproteins. The HDL particle is remodeled by the TG lipase and phospholipase activities of HL and endothelial lipase (EL). Finally, HDL cholesterol is selectively internalized by hepatic scavenger receptor class B type I (SR-BI), while apoA-I catabolism is mediated by the liver and kidneys.[14,15]

1.3 Genetic insights into lipoprotein metabolism

Although plasma lipid levels are influenced by many environmental variables –including age, gender, diet, smoking, exercise and alcohol consumption – up to 50% of inter-individual variation in lipid levels is attributable to genetic factors.[16,17] Genetic inheritance may be Mendelian or multifactorial in nature. In Mendelian inheritance patterns, the transmission of a particular genetic variant is both necessary and sufficient to confer a given phenotype. Multifactorial inheritance relies on the contributions of many genetic loci, and displays far less predictable intergenerational transmission patterns. Both of these mechanisms contribute, at some level, to heritability in plasma lipid levels.

Study of Mendelian dyslipidemias has enabled identification of rare genetic variants of large effect that regulate key aspects of lipoprotein production and turnover. The oft-cited example of a monogenic dyslipidemia is the autosomal co-dominant disorder familial hypercholesterolemia (FH), caused by rare but highly penetrant mutations in *LDLR*, which encodes the LDL receptor. Due to impaired clearance of circulating LDL-C particles from the blood, FH patients develop a clinical syndrome characterized by abnormally high LDL-C levels, tendon xanthomata and premature atherosclerotic disease.[18] Rarer forms of FH are caused by other genetic mutations that similarly affect LDL catabolism. For example, loss-of-function mutations in *APOB* interfere with binding of the apoB-100 ligand to the LDL receptor, preventing hepatic uptake of LDL particles.[19] Autosomal dominant gain-of-function mutations in proprotein convertase subtilisin kexin 9 (*PCSK9*), a liver-secreted serine protease that degrades the LDL receptor, result in decreased LDL receptor occupancy at the hepatocyte surface and thus delayed LDL clearance.[20] To date, these three genes and over 20 others have been implicated in syndromic dyslipidemias, characterized by plasma levels of LDL-C, HDL-C and/or TG that display marked deviations from the population norm.[3]

However, the deleterious rare variants underlying monogenic dyslipidemias primarily explain lipid level variation at the phenotypic extremes of a population. Like other complex traits, plasma lipid levels are not dichotomous but display a continuously graded distribution. For the most part, genetic heterogeneity in lipid levels reflects the aggregate impact of numerous variants of modest effect size. Although individual alleles may confer relatively incremental risk, the cumulative impact of many such variants throughout a patient's genome can significantly influence the overall lipid profile.[3,21] While family-based linkage studies have successfully identified the genetic underpinnings of many Mendelian disorders, they have largely failed to elucidate the genetic architecture of complex traits, due to the modest effect sizes of the implicated variants. Thus, discovery efforts in complex traits genetics have generally taken one of two forms: candidate gene studies and genome-wide association studies (GWAS).

Candidate gene analyses are hypothesis-driven experiments that use association or resequencing to identify putative causal alleles within a pre-selected panel of genes that are suspected to influence the phenotype of interest. Genes may be chosen based on data from prior linkage or association studies or knowledge of biological pathways. In a candidate gene association study, single nucleotide polymorphism (SNP) genotypes are compared between cases

and controls to uncover statistical correlations between a genetic variant and the phenotype under investigation. By profiling variants with high population prevalence, the candidate gene association study assumes that common genetic variation underlies susceptibility to complex traits. A key factor influencing the success of this type of study is the identity and number of SNPs included in the genotyping panel and the extent to which they affect gene function or exist in linkage disequilibrium (LD) with a functional variant. By and large, candidate gene association studies have yielded irreproducible associations, presumably due to limited gene and variant selection, as well as small sample sizes.[22,23]

In a candidate gene resequencing study, genes are sequenced in cases and controls to identify variants that segregate primarily within one group, implying a connection between the gene's function and the phenotype under investigation. Unlike association studies, resequencing studies have the power to detect the entire spectrum of common and rare allelic variation at a locus. Resequencing studies are based on the notion that low frequency alleles of intermediate penetrance are the primary drivers of complex trait susceptibility.[24] In one of the earliest applications of this approach, Cohen et al resequenced the HDL-C candidate genes *LCAT*, *ABCA1* and *APOA1* in individuals with low ($<5^{th}$ percentile) and high ($>95^{th}$ percentile) HDL-C levels and found that nonsynonymous sequence variants were highly enriched in the former group, suggesting that these alleles were causal for the low HDL-C phenotype. Through computational and biochemical analysis, variants from the low HDL-C cohort were found to have a deleterious effect on protein structure and function. Although these variants were individually rare, Cohen et al concluded that their aggregate prevalence was high enough to substantially contribute to population-level variation in HDL-C.[25] However, subsequent extension of the candidate gene resequencing approach has demonstrated that not all genes necessarily harbor an excess of low frequency variants, even at the population extremes.

More recently, GWAS have emerged as a powerful unbiased tool to identify SNPs associated with incidence of a particular phenotype or disease. The development of GWAS marks an important conceptual advance from candidate gene studies that profile a limited number of variants based on incomplete understanding of disease etiology. Instead, cases and controls are genotyped at a set of 100,000-2,000,000 SNPs that tag haplotype blocks spanning the entire genome. Because GWASs do not make *a priori* assumptions about the involvement of certain genes in a disease process, they have unprecedented power to identify novel loci that

have never previously been implicated in disease pathogenesis. Due to the modest effect sizes of most common variants, GWASs require large participant cohorts in order to reliably detect a *bona fide* association. GWASs are held to rigorous thresholds for statistical significance ($P<5x10^{-8}$), based on a Bonferroni correction to minimize false positive discovery when conducting a large number of independent tests. Further statistical confidence can be obtained by genotyping GWAS-implicated SNPs in an independent population (a replication study) to filter out spurious associations secondary to population stratification and other confounders.[26,27]

Interestingly, only a small fraction of GWAS-identified variants lie within coding sequence, thus directly implicating a causal gene at that locus. The vast majority of implicated SNPs fall in noncoding sequence, including introns and gene deserts, suggesting they may play a regulatory role in gene expression. Moreover, many of these SNPs are not themselves causal but exist in LD with the true functional variant. The causal gene driving an association signal is often not immediately apparent, unless the locus harbors a gene with a known connection to the phenotype of interest. Although GWASs typically label each associated SNP with the name of the nearest annotated gene or most plausible biological candidate at that locus, experiments in biological models are necessary to identify the true causal gene at the locus.[26,27]

Reassuringly, GWASs for lipid traits have identified variants in loci harboring genes that have previously been implicated in Mendelian disorders of lipoprotein metabolism (i.e. *LDLR*, *PCSK9*, *ABCA1*, etc). This finding suggests that common variants in these genes may actually contribute to phenotypic heterogeneity in the general population. Moreover, GWAS has uncovered a plethora of loci with no prior connection to lipid metabolism. In one of the most extensive GWASs for blood lipids to date, the Global Lipids Genetics Consortium (GLGC) conducted a meta-analysis of 46 prior lipid GWASs comprising >100,000 individuals of European descent, and identified 95 loci associated with total cholesterol (TC), LDL-C, HDL-C and/or TG. Of these loci, 36 had previously been reported by smaller-scale lipid GWASs at genome-wide significance, while the other 59 were previously unpublished.[28] Once characterized in biological systems, these novel loci may offer new insights into lipoprotein metabolism and promising targets for therapeutic intervention.

1.4 Genetics in the post-GWAS era

Although GWASs have identified a host of disease susceptibility loci, the pace of functional validation has lagged far behind. For the vast majority of GWAS loci, the causal DNA

variants and genes remain unexplored, largely due to the difficult and time-intensive nature of functional follow-up.

To identify a putative causal variant, researchers typically begin by resequencing the implicated interval in cases and controls to identify variants that demonstrate the most compelling association with the phenotype of interest. The recent publication of the 1000 Genomes Project has greatly facilitated fine-mapping efforts by providing a comprehensive catalog of low frequency variants, from which promising candidates can be selected for genotyping in cases and controls. By discriminating those variants that show the strongest statistical correlation with the phenotype, one can rationalize that the causal variant lies within the minimal interval defined by those variants and thereby exclude the remainder of the locus from further consideration. The probability of capturing the causal SNP through this approach depends on the effect size of the causal SNP, depth of sequence coverage, number of variants included in the genotyping platform, sample size, and how the boundaries of the resequenced interval are chosen. In some cases, trans-ethnic differences in LD structure may inform causal variant discovery. Since individuals of African and Asian descent typically have shorter LD block structure than European individuals, targeted resequencing of these populations (provided that the association signal also occurs in those populations) may help further refine the list of candidate variants, as long as the LD structure at the locus is favorable. However, even after fine mapping, as a result of LD tens to hundreds of variants can demonstrate indistinguishably strong associations with the phenotype, suggesting that genetic epidemiology alone is an insufficient means of causal variant discovery. Because many disease-associated variants are believed to modulate gene expression, further insight can be gleaned by integrating risk-associated variants with annotated maps of regulatory elements, such as DNase I hypersensitivity sites, chromatin immunoprecipitation sequencing (ChIP-seq) peaks, and histone modifications. Candidate SNPs that fall in transcriptionally active regions can then be prioritized for functional investigation. Experimental approaches such as reporter assays, electrophoretic mobility shift assays (EMSA) and ChIP can be employed to investigate allele-specific regulatory activity at each variant, as well as determinants of differential transcription factor binding and function.[24,29,30]

For some loci, expression quantitative trait locus (eQTL) studies may illuminate potential downstream targets of the risk variant. An eQTL is a genomic region that influences gene expression either in *cis* or *trans*; however, GWAS-implicated variants are predominantly

believed to function in *cis*-acting manner. As eQTLs tend to account for a greater fraction of trait variance than genetic risk variants, eQTL analyses can be carried out with smaller sample sizes than would be needed, for example, to detect association between a risk variant and a clinical phenotype. Additionally, the identity of the disease causal variant at a particular locus does not need to be known in order to perform an eQTL association study. Although early eQTL studies relied on lymphoblastoid cell lines, more recent analyses have been performed using human primary tissues obtained either post-mortem or during surgical resection. The availability of appropriate primary tissues is critical, since 50-90% of eQTLs are estimated to display tissue-specificity. Importantly, eQTL studies have identified differentially regulated transcripts hundreds of kilobases away from the genotyped variant, implicating long-range chromatin looping interactions in the mechanism of some eQTLs. These differentially regulated genes then become candidates for experimental manipulation (for example, through overexpression and knockout of the orthologous genes in murine models) to ascertain their relevance to the phenotype of interest.[31,32]

While a few seminal examples of genotype-to-phenotype connections at GWAS risk loci have been reported, most efforts have yielded preliminary and oftentimes contradictory insights into disease biology.[33,34] What factors have complicated the discovery of causal alleles in the post-GWAS era? The first issue is the haplotype structure of the genome and the sheer preponderance of highly correlated proxy SNPs at most disease susceptibility loci. The second potential issue is that a common variant may be insufficient to explain the association of a locus with the phenotype of interest. Hypothetically, one or more rare variants in LD with the implicated common variant may drive a "synthetic" association signal at that locus. Dickson et al have claimed that large-effect rare variants would explain a much higher proportion of phenotypic variance than the associated SNP itself, thereby accounting for the missing heritability of GWAS[35] – that is, the inability of GWAS-implicated common variation to fully account for the heritability of complex traits.[36] While this theory may be applicable to a select number of loci, resequencing of individual GWAS loci by and large has not supported the existence of "synthetic" associations. If "synthetic" associations were the norm, their aggregate effect sizes would account for more phenotypic variation than actually exists in the population. More likely, complex trait susceptibility is polygenic in nature, reflecting the composite

influence of hundreds of causal variants with variably low, intermediate or high allelic frequencies.[37]

1.5 A novel paradigm for functional genetic studies

Several recent technologies make it feasible to interrogate risk-associated variants at eQTL loci in a high-throughput fashion. The massively parallel reporter assay (MPRA) allows investigators to generate high-complexity pools of reporter constructs where each regulatory element of interest is linked to a synthetic reporter gene with a unique barcode identifier. The reporter construct pool is transfected into a relevant cultured cell type, and the relative transcriptional activity of each regulatory element is assessed by quantifying the abundance of each barcode in the transcribed reporter mRNA (**Figure 1**).[38] MPRA can thus be used to rapidly profile the regulatory activity of thousands of variants linked to eQTL SNPs, prioritizing candidate causal variants for further investigation.

Advances in genome editing technologies – from first-generation zinc finger nucleases (ZFNs) to, more recently, transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems – have opened up unprecedented avenues by which to rigorously assess the functional impact of novel genetic variants.[39] All three of these genome-editing tools can be used to introduce targeted alterations into mammalian cells and model organisms. However, CRISPR/Cas offers an optimal combination of high targeting efficiency, ease of use and scalability. The system uses the *Streptococcus pyogenes* Cas9 nuclease, which complexes with a synthetic guide RNA (gRNA) encoding a site-specific 20-nt protospacer sequence that hybridizes a $GN_{19}NGG$ target DNA sequence. Once Cas9 induces a double-strand break three nucleotides upstream of the NGG sequence, or protospacer adjacent motif (PAM), the cell employs error-prone non-homologous end joining (NHEJ) to repair the break, often leading to the introduction of an insertion or deletion that may disrupt gene function. If a single-stranded oligonucleotide (ssODN) is introduced, the cell can utilize it as a donor template for homology-directed repair (HDR), enabling knock-in of specific mutations.[40,41]

In addition to targeted genome editing, CRISPR/Cas9 has been co-opted for a wide variety of purposes in biological systems.[42] For example, genome-wide gRNA libraries have been developed to conduct powerful loss-of-function screens in human and mouse cells, thereby enabling hypothesis-free interrogation of biological processes.[43,44] As well as creating total loss-

of-function alleles, CRISPR/Cas can be used to activate or repress gene expression by attaching a catalytically inactive variant of Cas9 (dCas9) to different effector domains, which are recruited to a regulatory site via the gRNA. Cas9-based activators and repressors have also been employed on a genome scale to conduct gain and loss-of-function screens in which gene expression can be modulated over a wide dynamic range.[45,46] Furthermore, recent discovery and characterization of smaller Cas9 orthologues with greater targeting specificity has greatly expanded the therapeutic potential of CRISPR/Cas9.[47]

1.6 Rationale

Although GWASs have discovered a wealth of novel loci associated with blood lipid levels, the mechanisms by which they influence phenotype and disease risk remain poorly understood. eQTL analysis of patient liver and adipose biopsies indicates that many lipid-associated tag SNPs influence gene expression in a *cis*-regulatory manner. These eQTL associations may highlight candidate lipid-modulating genes, sometimes located hundreds of kilobases away from the eQTL tag SNP, which underlie the GWAS association signals at these loci. In a GWAS of 100,000 individuals of European descent, the Global Lipids Genetics Consortium (GLGC) interrogated tag SNPs at 95 loci for blood lipids against transcript abundance of local genes in samples of human liver, subcutaneous fat and omental fat. This analysis identified 57 liver and adipose eQTLs.[28]

One of the most robust eQTL associations from the GLGC study was a risk variant on chromosome 1p13 associated with both LDL-C and CHD. Within this locus, *SORT1*, whose gene product sortilin 1 regulates hepatic VLDL secretion and LDL-C clearance, is the causal gene responsible for the GWAS association. In the earliest example of mechanistic validation at a GWAS-implicated lipid locus, Musunuru et al used a combination of fine mapping and luciferase expression experiments to indirectly demonstrate that the noncoding variant rs12740374 was the likely candidate causal variant at this locus, and that the risk allele of this SNP disrupts hepatic transcription of *SORT1*.[33] However, similar efforts to map eQTL associations at other GWAS lipid loci using these traditional experimental techniques have been largely unsuccessful. In particular, LD among hundreds of neighboring SNPs at a GWAS-implicated locus makes it challenging to pinpoint the causal variant underlying an association signal.

To address this issue, our collaborators at the Broad Institute performed an MPRA experiment to rapidly profile the regulatory activity of the eQTL lead SNPs identified by the

GLGC study. We used a pool of reporter constructs in which every plausible regulatory variant (that is, all SNPs with $r^2 \geq 0.5$ relative to the eQTL tag SNPs) was embedded within a 145-bp tile in six versions (major or minor allele in the center, near the 5' end, or near the 3' end) to accurately capture as much of the surrounding genomic context as possible. Each regulatory element was coupled to a reporter gene with a unique barcode identifier in the 3' UTR. The construct pool was transfected into murine 3T3-L1 adipocytes, and the copy number of each barcode was quantified by RNAseq and normalized to the amount of corresponding reporter DNA plasmid that entered the cells (**Figure 1**). MPRA variants were prioritized according to the magnitude of allele-specific regulatory activity, as measured by reporter expression, in mouse 3T3-L1 adipocytes (**Table 1**).

In this thesis, I sought to leverage CRISPR/Cas technology to provide functional evidence of causality at several high-priority GWAS eQTL loci for serum lipids. I selected the two top-ranked MPRA variants, rs2277862 and rs10889356 (**Table 1**), as well as the 1p13 SNP, rs12740374, as candidates for further investigation. I hypothesize that each SNP is causal for its respective eQTL and lies within a transcriptional regulatory element to influence nearby gene expression in an allele-specific manner. To accurately model human genetic regulation, I employ a combination of genome edited human pluripotent stem cells (hPSCs) and primary hepatocytes, CRISPR interference, and humanized mouse models. These approaches and results highlight a novel experimental framework to discover causal genes and variants contributing to complex human traits.

2. MATERIALS AND METHODS

2.1 Plasmid construction

Guide RNAs were designed by manual inspection of the genomic sequence flanking rs2277862 and rs10889356, and evaluated for potential off-target activity using the CRISPR design tool at http://crispr.mit.edu. Protospacers were cloned into the BbsI site of pGuide (Addgene plasmid #64711) via the oligonucleotide annealing method, and, if not already present, a G was added to the 5' end to facilitate U6 polymerase transcription. Genome editing was performed using pCas9_GFP (Addgene plasmid #44719), which co-expresses a human codon-optimized Cas9 nuclease and GFP via a viral 2A sequence.

For CRISPR interference studies, pAC154-dual-dCas9VP160-sgExpression (Dr. Rudolph Jaenisch, Addgene plasmid #48240), a dual expression construct that expresses dCas9-VP160 and sgRNA from separate promoters, was modified by PCR-based methods to include a viral 2A sequence and GFP after dCas9-VP160. Additionally, the gRNA sequence was modified to include a 5 bp hairpin extension, which improves Cas9-gRNA interaction, and a single base pair substitution (A-U flip) that removes a putative Pol III terminator sequence, as described previously.[48] Where indicated, the VP160 transactivation domain was removed from the construct by PCR-based methods.

2.2 Cultured cell line maintenance and transfection

All cell lines were maintained in a humidified 37°C incubator with 5% CO_2. HEK293T, HepG2 and 3T3-L1 cells were cultured in high glucose DMEM supplemented with 10% FBS and 1% penicillin/streptomycin. For CRISPRi experiments, HEK293T and HepG2 cells were seeded into 6-well plates and transfected 24 hours later using Lipofectamine 3000 (Life Technologies) according to the manufacturer's instructions.

2.3 Human pluripotent stem cell (hPSC) culture and CRISPR targeting

HUES 8 (Harvard University) and H7 cells (WiCell Research Institute) were grown under feeder-free conditions on Geltrex (Life Technologies)-coated plates in chemically defined mTeSR1 medium (STEMCELL Technologies), supplemented with 1% penicillin/streptomycin and 5 μg/mL Plasmocin (InvivoGen). Medium was changed every 24 hours. For electroporation, cells in a 60-70% confluent 10-cm plate were dissociated into single cells with Accutase (Life Technologies), resuspended in PBS, and combined with 25 μg pCas9 and 25 μg gRNA plasmid (or 12.5 μg of two different gRNA plasmids, for multiplexed targeting) in a 0.4 cm cuvette. For

knock-in, 15 μg pCas9_GFP, 15 μg gRNA plasmid, and 30 μg ssODN (5'-GGTCGTCAGAACCCACGAGGTCATGATCAAATAT**T**GGCGACCGTCAGCTCCGTCTCAGCTGGGAGAGA-3') were used instead. A single pulse was delivered at 250 V/500 μF (Bio-Rad Gene Pulser), and the cells were recovered and plated in mTeSR1 with 0.4 μM ROCK inhibitor (Y-27632, Cayman Chemical). Cells were dissociated with Accutase 48 hours post-electroporation, and GFP positive cells were isolated by FACS (FACSAriaII, BD Biosciences) and replated onto 10-cm Geltrex-coated plates (15,000 cells/plate) with conditioned medium and 0.4 μM ROCK inhibitor to facilitate recovery.

2.4 Isolation and screening of clonal hPSC populations

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

hPSC clones were screened by PCR amplification of a small region surrounding the targeted site using BioReady rTaq DNA Polymerase (Bulldog Bio) and the following cycling conditions: 94°C 5 min, [94°C 30 s, 54-56.5°C 30 s, 72°C 30 s] x 40 cycles, 72°C 5 min. The following primer pairs were used: for rs2277862, F: 5'- TGCTGGACCCACACTTCATA-3' and R: 5'-CTCAGTCCCTCTCCCTCCTT-3'; for rs10889356, F: 5'-CCATTAGGTCACTTGCCAGA-3' and R: 5'- ACAGGGGGATTCTGTCTAAAA-3'. PCR amplicons were separated on a high-percentage agarose gel and clones with indels were identified based on size shifts relative to the wild-type band. Suspected mutant clones were confirmed by Sanger sequencing of the PCR products.

Multiple mutant clones were retrieved from the frozen stock, or if possible, from the second working stock and expanded for experiments. Additionally, several clones that underwent

the targeting procedure but remained genetically wild-type at the intended site were expanded as controls.

2.5 Differentiation of hPSCs into white adipocytes

Differentiation of hPSCs cells to white adipocytes was performed according to a published protocol.[49] To induce embryoid body formation, wild-type and mutant hPSCs were pre-treated overnight with 2% DMSO; dissociated into small clumps with Accutase; resuspended in growth medium containing DMEM, 10% knockout serum replacement (Life Technologies), 2 mM GlutaMAX (Life Technologies), 1% non-essential amino acids, 1% penicillin/streptomycin, and 0.1 mM beta-mercaptoethanol; and transferred to low-attachment 6-well plates (Costar Ultra Low Attachment; Corning Life Sciences). After one week in culture, embryoid bodies were collected and replated onto gelatin-coated plates in MPC medium containing DMEM, 10% FBS, 1% penicillin/streptomycin, and 2.5 ng/mL bFGF (Aldevron), Cells were serially passaged at a 1:3 ratio to obtain a homogenous population of MPCs by passage 3-4.

Recombinant lentivirus was produced using a third-generation, Tat-free packaging system. Lentiviral vectors encoding either doxycycline-inducible *PPARG2* or rtTA were transfected into HEK293T cells by the calcium phosphate method, along with the packaging plasmids pMDL and pREV and a capsid plasmid encoding VSV-G. Viral supernatant was harvested at 48 and 72 hours post-transfection and filtered through a 0.45 μm membrane. One day before transduction, MPCs were plated at $1x10^6$ cells per 10-cm plate. The following day, MPCs were transduced with 5 mL lenti-*PPARG2* and 5 mL lenti-rtTA and incubated at 37°C for 16 hours. After the viral supernatant was aspirated, the cells were washed with PBS and allowed to grow to confluence. Transduced MPCs were split into 6-well dishes prior to initiating white adipocyte differentiation.

Differentiation was induced by the addition of adipogenic media containing DMEM, 7.5% knockout serum replacement, 7.5% human plasmanate (Grifols), 0.5% non-essential amino acids, 1% penicillin/streptomycin, 0.1 μM dexamethasone (Sigma), 10 μg/mL insulin (Sigma), and 0.5 μM rosiglitazone (Santa Cruz). The differentiation medium was supplemented with 700 ng/mL doxycycline from day 0 to 16. Doxycycline was then removed from the culture medium until day 21, at which point the differentiated cells were harvested for gene expression experiments.

2.6 Differentiation of hPSCs into hepatocyte-like cells

Differentiation of hPSCs into HLCs was performed according to a published protocol.[50] One day before differentiation, hPSCs at 60% confluence were split at a 1:3 ratio into 6-well dishes with mTeSR1 plus 0.4 μM ROCK inhibitor. Cells were serially cultured in (1) RPMI-B27 (RPMI-1640 from Sigma; B27 supplement minus Vitamin A from Life Technologies) supplemented with 100 ng/mL Activin A (PeproTech) and 3 μM CHIR99021 (Cayman Chemical), a glycogen synthase kinase 3 inhibitor, for 3 days to obtain definite endoderm, (2) RPMI-B27 supplemented with 5 ng/mL bFGF (Millipore), 20 ng/mL BMP4 (PeproTech), and 0.5% DMSO for 5 days to obtain hepatic endoderm, (3) RPMI-B27 supplemented with 20 ng/mL HGF (PeproTech) and 0.5% DMSO for 5 days to obtain immature hepatocytes, and (4) Hepatocyte Culture Medium (Lonza) supplemented with 20 ng/mL HGF, 20 ng/mL Oncostatin M (PeproTech), 100 nM dexamethasone (Sigma), and 0.5% DMSO for 10 to 12 days to obtain mature HLCs.

2.7 Adenovirus generation

The *Streptococcus pyogenes* CRISPR-Cas9 system and the guide RNA protospacers were inserted into the Adeno-X vector (Clontech) as previously described.[51] The protospacer to specifically target the rs12740374 minor allele sequence (5'- GTGCTTGATTGAGCAACCTC-3') was designed by manual inspection. Irrelevant protospacers were used as controls for the primary human hepatocyte experiments (protospacer 5'-TTTTTTGTTTTTTGTTTTTT-3') and the BAC transgenic mouse experiments (5'- GGTGCTAGCCTTGCGTTCCG-3'). The Penn Vector Core at the University of Pennsylvania used these vectors to generate recombinant adenoviral particles (designated CRISPR-SNP or CRISPR-control).

2.8 Primary human hepatocyte experiments

Single vials (each containing 5 to 15 million viable primary hepatocytes) derived from 20 individuals were obtained from Gibco, representing a combination of: Human Plateable Hepatocytes, Induction Qualified; Human Plateable Hepatocytes, Metabolism Qualified; and Human Plateable Hepatocytes, Transporter Qualified. All were rated to survive in culture for at least three days after replating. The cells from each vial were thawed into Cryopreserved Hepatocyte Recovery Medium (Gibco), gently spun down, and then resuspended in William's Medium E with added Hepatocyte Thawing & Plating Supplement (fetal bovine serum, dexamethasone, and a cocktail solution of penicillin-streptomycin, bovine insulin, GlutaMAX, and HEPES) (Gibco) and plated at 400,000 cells/well in collagen-coated 24-well plates,

according to the manufacturer's instructions. After 6 hours, each well was refed with 500 µL maintenance medium [William's Medium E with added Hepatocyte Maintenance Supplement (dexamethasone and a cocktail solution of penicillin-streptomycin, insulin, transferrin, selenium complex, BSA, linoleic acid, GlutaMAX and HEPES)] (Gibco). Each well was refed with 500 µL maintenance medium daily for the duration of the experiment.

Residual cells left over in the thawed vials were washed out with PBS and collected, and genomic DNA was isolated using the DNeasy Blood and Tissue Kit (QIAGEN) according to the manufacturer's instructions. A 423-bp region surrounding the rs12740374 SNP was PCR amplified using the following primers: F: 5'-AGGAACTGGAAAAGCCCTGT-3' and R: 5'-GAGGCCACAGCAGGTTAGAC-3'. PCR amplicons were subjected to Sanger sequencing to determine the rs12740374 genotypes for each lot.

One day after plating, one to four pairs of wells for each lot (depending on availability of cells) were treated with either CRISPR-1p13 adenovirus or CRISPR-control adenovirus. Each well received 1.5×10^7 viral particles mixed into 250 µL maintenance medium, for an estimated MOI of 37.5. The virus was aspirated after four hours and replaced with 500 µL maintenance medium. Two days later, both non-virus-treated and virus-treated wells were either (1) used for DNA isolation with QuickExtract DNA Extraction Solution (Epicentre) and PCR amplification of the rs12740374 SNP sequence (as described above) for deep sequencing or (2) lysed directly in TRIzol Reagent (Thermo Fisher Scientific) for gene expression studies as described below.

2.9 Guide RNA screening and generation of CRISPR knock-in mice

Four candidate guide RNAs with a cut site near rs27324996 were designed by manual inspection and the corresponding protospacers were cloned into the pGuide plasmid as described above. Each gRNA plasmid was co-transfected with pCas9_GFP into mouse 3T3-L1 cells using TransIT-2020 Reagent (Mirus Bio) according to the manufacturer's instructions. Two days post-transfection, GFP-positive cells were isolated by FACS and genomic DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen). The region flanking rs27324996 was PCR amplified (F: 5'- TGGGAATGGCTTCTTAGGGC-3' and R: 5'-CATCCCCAAGCAACTCAACC-3') using AccuPrime Taq DNA Polymerase (Life Technologies) with the following cycling conditions: 94°C 2 min, [94°C 30 s, 55°C 30 s, 68°C 30 s] x 40 cycles, 68°C 5 min. PCR products were purified using the DNA Clean and Concentrator kit (Zymo Research) and analyzed for the presence of indels using the Surveyor Mutation Detection Kit (IDT) according

to the manufacturer's instructions. CEL I nuclease-treated PCR products were resolved on a 1.5% agarose gel to detect mutagenesis activity. The gRNA sequence exhibiting the highest mutation rate was PCR amplified, and the purified PCR product was used as a template for *in vitro* transcription using the MEGAshortscript T7 kit (Life Technologies). The transcribed RNA was purified by phenol/chloroform extraction, ethanol precipitated, and resuspended in injection buffer (5 mM Tris-HCl pH 7.6, 0.1 mM EDTA).

All animal protocols described here were reviewed and approved by the Harvard University Institutional Animal Care and Use Committee. One-cell embryo injections were performed by the Genome Modification Facility at Harvard University. Superovulated C57BL/6J females were mated with C57BL/6J males and fertilized embryos were harvested from the oviducts. One-cell embryos were injected with a mixture of 100 ng/μL Cas9 mRNA (TriLink BioTechnologies), 50 ng/μL gRNA, and 100 ng/μL ssODN (5'-AGCCCACAGTTGGCTCTGTGGTGGCTATAGAATCTGTTTTCCAGGTCAATGTGGGTCTCCCCGATGAGGTCATCTGAACCCACGAGGTCATGATCAAATA**T**GGCGACCGTCAGCTCTGGCTGGGCTGGGAGGGAGACGCTCAGCTCCAGGACCCTGGGCAGGAAGGGAAATTGACTAACCACAGCTCCATGCCCTCAGAG-3'). Injected embryos were implanted into the uterus of pseudopregnant foster mothers.

DNA was prepared from tail biopsies of 3-week-old founder mice by the hot hydroxide method, and genotyping was performed with the same PCR primers and cycling conditions used for the Cel-I nuclease assay. Positive founders were identified by Sanger sequencing of PCR products. A single positive founder was bred to wild-type C57BL/6J mice (Jackson Laboratories) and the resulting progeny were intercrossed for one to two generations to breed the knock-in allele to homozygosity. Wild-type and homozygous knock-in littermates from several litters, approximately 12 weeks of age, were used for gene expression studies.

2.10 Generation of BAC transgenic mice and somatic in vivo genome editing experiments

BAC clone RP11-463O24 (with the insert hg18/chr1:109541561-109744884) from the human RPCI-11 Human Male BAC Library (BACPAC Resource Library, Children's Hospital Oakland Research Institute) was grown in DH10B *E. coli* cells and purified with the NucleoBond BAC 100 kit (MACHEREY-NAGEL). The BAC DNA was used for pronuclear injection into C57BL/6J embryos at the Harvard University Genome Modification Facility. Ten founder mice that were positive for the BAC transgene by PCR analysis were obtained. In each

of these mice, the integrity of the transgene was tested with PCR of 11 amplicons distributed throughout the BAC insert sequence. Quantitative PCR with TaqMan SNP Genotyping Assay, Human SM, Assay ID C__25753757_20 (Applied Biosystems) was used to assess relative copy numbers among the mice that were positive for all of the amplicons. The founder mouse with the lowest copy number of the complete transgene and its descendants were bred with wild-type C57BL/6J mice to the F2 generation, in which roughly 50% of the offspring were positive for the complete transgene, consistent with Mendelian transmission of a single transgene insertion site. Quantitative PCR confirmed a consistent copy number across all of the positive F2 mice. Mice that were three to four months of age were used for experiments. The mice were administered 1×10^{11} viral particles each of either CRISPR-SNP or CRISPR-control adenovirus via retro-orbital injection. As much as possible, the mice in the two groups were matched with respect to age and sex. After four days, the mice were sacrificed by carbon dioxide asphyxiation, and whole liver samples were harvested and snap-frozen in liquid nitrogen. Liver genomic DNA was subsequently isolated using the DNeasy Blood & Tissue Kit. A 423-bp region surrounding the rs12740374 SNP was PCR amplified (as described above). PCR products were purified, analyzed using the Surveyor Mutation Detection Kit according to the manufacturer's instructions, and resolved on 2.0% agarose gels. Liver samples were used for gene expression studies as described below.

2.11 Quantitative PCR

Cells were washed with ice-cold PBS and lysed in TRIzol (Life Technologies). Snap-frozen liver samples were homogenized in TRIzol reagent. RNA was isolated according to the manufacturer's instructions and reverse transcribed using SuperScript III Reverse Transcriptase (Life Technologies) with an equimolar mixture of random hexamers and oligo-dT. Gene expression was measured using the following TaqMan Gene Expression Assays along with TaqMan Gene Expression Master Mix (Applied Biosystems): Hs00898245_m1 for *CEP250,* Hs00537765_m1 for *CPNE1*, Hs00211070_m1 for *ERGIC3*, Hs00205581_m1 for *ANGPTL3*, Hs00290630_m1 for *DOCK7*, Hs00910225_m1 for *ALB*, Hs01097800_m1 for *SERPINA1*, Mm00623502_m1 for *Cep250*, Mm00467970_m1 for *Cpne1*, Mm00499400_m1 for *Ergic3*, Hs00361760_m1 for *SORT1*, Hs00934024_g1 for *PSRC1*, Hs00197856_m1 for *SARS*, and Hs00936004_m1 for *PSMA5*. Human B2M (Assay ID 4326319E) or mouse Actb (Assay ID 4352341E) was used as the reference gene as appropriate. Each 10 μL qPCR reaction contained

1 μL cDNA (diluted 1:3 with water) and was performed in technical duplicate or triplicate. Reactions were carried out on a ViiA 7 Real-Time PCR system (Applied Biosystems) and relative expression differences were quantitated by the $\Delta\Delta C_t$ method.

2.12 Deep Sequencing

PCR amplicons from adenovirus-treated primary human hepatocytes or BAC transgenic mice were subjected to next-generation DNA sequencing at the Massachusetts General Hospital CCIB DNA Core (CRISPR Sequencing service; https://dnacore.mgh.harvard.edu/new-cgi-bin/site/pages/crispr_sequencing_main.jsp). Sequencing data were processed according to standard Illumina sequencing analysis procedures. Processed reads were mapped to the expected PCR amplicon as reference sequences using a custom script; reads that did not map to reference were discarded. Indel frequencies were determined as follows. The reads were analyzed using a custom script to identify indels by matching reads against reference, with indels involving any portion of the sequence within 15 nt upstream or downstream of the predicted CRISPR-Cas9 cleavage site (3 nt upstream of the 3' end of the protospacer) considered to be possible CRISPR-Cas9-induced mutations. Reads for which there was any 18-nt sequence with more than 2 mismatches with the corresponding 18-nt portion of the reference sequence, either upstream or downstream of a candidate indel, were discarded as errors. For reads from CRISPR-1p13-treated primary human hepatocytes, a custom script was used to discriminate whenever possible between indels on minor allele-bearing chromosomes and indels on major allele-bearing chromosomes.

2.13 Statistical analysis

Data represent mean ± standard error of the mean. For hPSC and mouse experiments, average gene expression levels were compared between groups using the non-parametric Mann-Whitney *U* test. For CRISPRi experiments, average gene expression levels in the control and experimental groups were compared using the unpaired Student *t*-test. For experiments in non-transduced primary human hepatocytes, average gene expression levels between rs12740374 homozygous major lots (n = 14 lots, each plated in triplicate) and heterozygous lots (n = 6 lots, each plated in triplicate) were compared using the Mann-Whitney *U* test. For experiments in virally transduced primary human hepatocytes, average gene expression levels in the control and experimental groups were compared using the Wilcoxon signed-rank test (n = 19 paired sets with paired wells derived from 6 heterozygous lots; the total number of paired wells derived from

each lot varied based on the availability of viable cells in the original vial). Statistical analysis was carried out using GraphPad Prism.

3. RESULTS

3.1 Functional analysis of rs2277862 at the 20q11 locus

3.1.1 Functional studies of rs2277862 in hPSCs

MPRA identified rs2277862 as the top-ranked regulatory variant in 3T3-L1 adipocytes, as measured by allele-specific reporter expression (**Table 1**). Of note, in the GLGC study, rs2277862 was also the lead SNP for total cholesterol (TC) at the 20q11 locus (P=4x10^{-10}), with the minor allele associated with a 1.19 mg/dL decrease in TC. The 20q11 genetic locus harbors a number of genes with no prior connection to lipid metabolism, although only three – *CEP250*, *CPNE1* and *ERGIC3* – show evidence of differential regulation in eQTL studies from human tissue biopsies (**Table 2**).[28] The biology and relevant sites of action of all three of these genes is poorly understood.

I hypothesized that rs2277862 is causal for the eQTL at the 20q11 locus, and that it lies within a transcriptional regulatory site to influence nearby gene expression in an allele-specific manner. The MPRA data indicates that the minor allele (T) of rs2277862 (MAF = 0.15) increases transcriptional activity relative to the major allele (C) (**Table 1**). For this relationship to be true, either the minor allele functions as an enhancer, the major allele as a repressor, or both. However, the human eQTL data suggests that the causal SNP may variably regulate expression of different genes in the locus, as the presence of the minor allele is associated with increased levels of certain genes, but decreased levels of others (**Table 2**). Interestingly, rs2277862 is located over 50 kb away from two of the eQTL genes, *CEP250* and *CPNE1*, suggesting that it may function in long-range enhancer or repressor interactions (**Figure 2a**).

To functionally demonstrate causality for these two SNPs, I sought a system in which to accurately model human genetic regulation. Ideally, I would want to compare gene expression in human primary cells with different genotypes at these SNPs. However, obtaining biopsies from healthy donors is unethical; while surgical specimens are available, they are likely derived from non-healthy donors whose underlying disease processes may alter gene expression patterns in that tissue. An alternate method would be to generate induced pluripotent stem cell (iPSC) lines from patients that have different genotypes at the SNP of interest and reprogram the cells to a relevant cell type for gene expression studies. Although iPSCs have several advantages for genetic modeling – including normal karyotype, renewability and pluripotency – the principal disadvantage is that differences in genetic background between iPSC lines may confound gene

expression analyses, and a large number of cell lines would have to be compared to obtain a meaningful result.[39,52]

Instead, I reasoned that an even more rigorous approach would be to use genome editing in human pluripotent stem cells (hPSCs) to create isogenic cell lines that differ only at the SNP of interest. Unlike iPSCs, isogenic cell lines are similar in epigenetic state and derivation/culture conditions, and are more likely to display similar differentiation propensities. Expression of genes within several hundred kilobases of the causal variant can then be compared in differentiated cells of varying genotypes. Because the cell lines are derived from the same parental cells, I could, in theory, confidently attribute any observed change in transcript levels to the induced genetic alteration, and not confounding differences in genetic background between the lines.[39,52]

To determine if rs2277862 regulates expression of genes at the 20q11 locus, I sought to use CRISPR/Cas9 technology to create isogenic hPSC lines with alternate allelic variants at rs2277862. In the hPSC line HUES 8, which is homozygous major (C/C) at rs2277862, I used CRISPR/Cas9 with a single-stranded oligonucleotide (ssODN) repair template to knock in the minor allele via HDR. To generate isogenic hPSC lines, I used the CRISPR/Cas genome-editing platform developed by our laboratory.[53,54] In brief, the platform uses two plasmids, one co-expressing Cas9 and GFP from the CAG promoter, the other the gRNA from the U6 polymerase III promoter. The plasmids are co-electroporated into hPSCs, followed two days later by fluorescence-activated cell sorting (FACS) gated on GFP. Single cells are re-plated at low density and cultured for 1-2 weeks to establish distinct clonal populations. Colonies are then expanded, genomic DNA extracted, and PCR used to amplify a small area surrounding the SNP, followed by sequencing to identify genomic alterations. Out of the screened clones, several successfully targeted clones as well as clones that remain wild-type are used for experiments. This controls for the effects of experimental manipulation, as the wild-type clones have also been exposed to CRISPR/Cas9 and subjected to the same workflow as the mutant clones.

Using a single gRNA along with a 67-nucleotide ssODN in HUES 8 cells, I obtained a single recombinant heterozygote at a frequency of 0.15% (1 out of 672 clones screened) (**Figure 2b**), reflecting the low efficiency of HDR in hPSCs. Because rs2277862 has an eQTL in three developmentally distinct tissues – liver, subcutaneous fat and omental fat – I reasoned that it may function as a global, non-cell type restricted regulator of gene expression and thus modulate

transcription in undifferentiated hPSCs as well. By quantitative PCR (qPCR), I observed significantly decreased *CPNE1* expression in the undifferentiated knock-in hPSC clone (down 19%) compared to two matched wild-type clones, with a non-significant decrease in *ERGIC3* (*CEP250* expression was not assayed in this experiment) (**Figure 3a**).

I next sought to evaluate gene expression changes in rs2277862 knock-in hPSCs that had been differentiated into two cell types relevant to lipid metabolism, hepatocytes and adipocytes. One knock-in clone and two matched wild-type clones were differentiated to hepatocyte-like cells (HLCs) using a pre-validated virus-free protocol in which cells are serially cultured with various small molecules to obtain definitive endoderm, hepatic endoderm, immature hepatocytes, and finally mature HLCs.[50] As with the undifferentiated hPSCs, *CPNE1* expression was significantly diminished in the knock-in HLCs (down 10%), with a non-significant decrease in *ERGIC3* (**Figure 3b**) (*CEP250* expression was not assayed in this experiment). Expression of the liver-specific marker *ALB* was equivalent between groups.

To obtain differentiated adipocytes, I used an established protocol in which hPSCs are grown into embryoid bodies, replated to obtain fibroblast outgrowth, and serially passaged to obtain mesenchymal progenitor cells (MPCs), which are multipotent precursors to a wide array of cell types, including osteoblasts, chondrocytes, myocytes and adipocytes. MPCs are then transduced with a doxycycline-inducible *PPARG* construct and grown in adipogenic medium with doxycycline for 16 days, followed by withdrawal of doxycycline for 5 days before analysis of gene expression profiles by qPCR. I obtained near-complete reprogramming of hPSCs to adipocytes as evidenced by the presence of multilocular lipid droplets (which are the distinguishing hallmark of adipocytes) in the vast majority of cells, confirming prior estimates of differentiation efficiency using this gene transfer-based protocol.[49] Similar to undifferentiated hPSCs and HLCs, *CPNE1* expression was decreased in knock-in adipocytes (down 6.9%), although this data only trended towards statistical significance. No statistically significant changes in *ERGIC3* expression were observed (**Figure 3c**).

I surmised that the statistical power of these experiments was constrained both by the expected modest effect of the common variant on intra-locus gene expression, as well as the limited availability of knock-in clones, secondary to the poor efficiency of HDR in hPSCs. In light of these limitations, I pursued an alternative approach. Since transcription factor binding sites are typically 8-10 nucleotides long, I reasoned that even small deletions encompassing the

candidate SNP would have a deleterious effect on gene expression. Importantly, CRISPR/Cas9 can be used to generate deletion mutations via NHEJ at a much higher frequency than knock-in mutations, increasing the number of targeted clones available for experiments. However, when a single gRNA is used, NHEJ randomly creates a wide array of indel mutations. This is a disadvantage for studies involving undefined regulatory elements, as variable mutations may have differing effects on transcriptional activity. To circumvent this issue, I utilized dual gRNAs with cut sites flanking the SNP, 38-bp apart. This multiplexing strategy facilitated efficient generation of many hPSC clones harboring predictable, and often homozygous, microdeletions encompassing the candidate SNP (59% mutation frequency, 168 out of 285 clones screened) (**Figure 2c**).

To define the contribution of SNP allele to gene expression, I created rs2277862 "knockouts" in two hPSC lines with alternate genotypes: HUES 8 (homozygous major, C/C) and H7 (homozygous minor, T/T). I compared a large number of independent hPSC clones in order to maximize the chance of detecting even a small gene expression difference. In undifferentiated hPSCs, homozygous disruption of the major allele in HUES 8 cells (n=10 wild-type and 10 knockout clones) significantly decreased expression of *CEP250* (down 26%), *CPNE1* (down 31%) and *ERGIC3* (down 20%) (**Figure 4a**). Homozygous disruption of the minor allele in H7 cells (n=8 wild-type and 6 knockout clones) decreased expression of *CEP250* (down 8.8%) and *ERGIC3* (down 10%) to a lesser degree, with non-significant effects on *CPNE1* expression (**Figure 4b**).

In combination, these data suggest that the major allele (C) of rs2277862 has enhancer activity, given that both substitution of the major allele with the minor allele (in the knock-in) as well as homozygous deletion of the major allele (in the knockout) result in diminished expression of the 20q11 genes.

3.1.2 CRISPR interference at rs2277862

Next, as a complementary approach, I sought to use a CRISPR-based transcriptional modulator to validate rs2277862 as a causal SNP. This experiment harnesses the sequence-dependent targeting specificity of CRISPR/Cas9 to direct a catalytically dead Cas9 mutant (dCas9) fused to either a transactivation or repression domain to the SNP site.[45] The rationale is that if a variant is truly causal and lies within a transcriptional regulatory element, then artificially activating or repressing the site with an exogenous transcription factor will alter

expression of transcripts regulated by that site. Initially, I generated CRISPR-activation constructs that co-expressed dCas9 with a C-terminal VP160 domain and GFP, along with each of the three gRNAs shown in **Figure 5a**. Cheng et al have shown that clusters of 3-4 gRNAs, when targeted to a promoter, synergistically induce gene expression compared to individual gRNAs.[55] Originally, I introduced the dCas9-VP160-gRNA constructs into HUES 8 cells, either singly or in combination. By qPCR, I did not observe statistically significant gene expression changes relative to control cells, which received the dCas9-VP160 construct without a gRNA. Reasoning that the experiment was limited by low transfection or gRNA targeting efficiency in hPSCs, I repeated the experiment in HEK 293T cells, which are homozygous for the major allele at rs2277862 (C/C). Paradoxically, expression of *CEP250*, *CPNE1*, and *ERGIC3* was diminished despite the presence of the transactivation domain. I hypothesized that the transactivation domain was inadequately positioned to activate gene expression, and had instead sterically hindered recruitment of native transcriptional machinery to the regulatory site. I repeated the experiment in HEK 293T cells using constructs that lacked the VP160 domain, and noted that with at least two of the gRNAs, expression of *CEP250* and *CPNE1* were diminished relative to control cells (**Figure 5b**). This result suggests that the dCas9/gRNA complexes (with or without the VP160 domain) had similarly obstructed binding or function of a transcriptional enhancer at the major allele. Notably, this conclusion is directionally concordant with the hPSC-based genome editing experiments, which had also ascribed enhancer activity to the major allele of rs2277862.

3.1.3 Functional studies of rs2277862 in a locus-humanized mouse model

Although both hPSC-based genome editing and CRISPR interference experiments had provided important insights into rs2277862 biology, each system presented notable limitations. The hPSC-based experiments were constrained by the low efficiency of HDR and artificial differentiation protocols that may create inauthentic replicas of primary tissue. The CRISPR interference experiments, while relatively rapid and straightforward, could only provide indirect evidence of causality in cultured cell lines through transcriptional blockade in the general genomic area surrounding the SNP. I sought an alternative system by which to faithfully model the effect of allelic variation at rs2277862 in primary tissues of interest, namely liver and adipose. Remarkably, the noncoding region encompassing rs2277862 is well conserved in mouse, and the orthologous nucleotide in mouse, rs27324996, also displays naturally occurring variation with the same major (C) and minor (T) alleles as in humans. Per dbSNP, rs27324996,

on mouse chromosome 2, has a MAF of 43% based on genotyping analysis of 14 different inbred strains of mice. All three human eQTL genes have a murine homolog at this locus, and the orientation of these genes relative to the putative regulatory variant is conserved between mouse and human (**Figure 6a**).

Because MPRA identified a transcriptional role for rs2277862 in 3T3-L1 adipocytes, which are of murine origin, I reasoned that the regulatory machinery at this site is also conserved across species. Since HDR is far more efficient in mouse embryos than in hPSCs,[56] I sought to knock-in the minor allele onto the C57BL/6J background, which is homozygous major at rs27324996. Compared to genome editing in hPSCs, the obvious advantage of this strategy is that even if a single positive founder is obtained, the knock-in allele can be bred to homozygosity in a matter of months. Thus, a large number of wild-type and knock-in mice can be obtained for well-powered gene expression studies in liver, omental fat and subcutaneous fat. This study design thus enables elegant replication of the human eQTL association data in primary tissue, while avoiding the limitations of imprecise differentiation protocols or cultured cell lines.

I designed four candidate gRNAs targeting the sequence flanking rs27324996 and screened them for activity in mouse 3T3-L1 cells using the CEL I nuclease assay (**Figure 6b**). The most active gRNA was chosen for embryo injection. In addition to the minor SNP allele, I introduced four non-conserved nucleotides into the donor oligonucleotide that "humanized" the flanking regulatory sequence and altered both the gRNA protospacer and PAM, thereby preventing re-cleavage of the knock-in allele. One-cell mouse embryos were injected with Cas9 mRNA, a gRNA targeting rs27324996, and the donor oligonucleotide. Out of 37 founder mice, one heterozygous minor allele knock-in founder mouse was obtained (**Figure 6c**), and subsequently bred for multiple generations to obtain homozygous minor allele knock-in mice. (Of note, I did attempt in parallel to create a homozygous major allele knock-in mouse with the same four nucleotide substitutions as above, to serve as a perfect match for the humanized minor allele knock-in mouse; however, this effort was unsuccessful. Ideally, I would have liked to generate an allelic series of mice with humanized homozygous major, heterozygous, or homozygous minor SNP alleles, but settled upon intercrossing F1 descendants of the minor allele knock-in founder mouse.)

Expression levels of *Cep250*, *Cpne1*, and *Ergic3* were compared in primary liver, omental fat, and subcutaneous fat samples obtained from wild-type (C/C) and homozygous

knock-in (T/T) littermates (*n*=18 wild-type and 10 knock-in mice). There was significantly decreased expression of *Cpne1* (down 37%) in the liver of homozygous minor allele knock-in mice. *Cep250* (down 28%) and *Ergic3* (down 34.1%) also displayed consistent reductions with data trending toward statistical significance (**Figure 7a**). Analysis of gene expression in omental and subcutaneous fat did not display any noteworthy trends, owing to the large degree of variance in gene expression among individual mice (**Figures 7b and 7c**). Cholesterol levels were unchanged between wild-type and knock-in mice, as expected given the small effect size in humans. Interestingly, these data mirror the results obtained in rs2277862 minor allele knock-in hPSCs; a statistically significant reduction in *CPNE1* expression was observed in hPSCs and differentiated HLCs, while the results in differentiated adipocytes were non-significant. Importantly, the directionality of the hPSC, CRISPR interference and mouse experiments are concordant, suggesting that the major allele of rs2277862 functions as a transcriptional enhancer. Taken together, these results establish a causal role for rs2277862 in gene expression at the 20q11 cholesterol locus.

3.2 Functional analysis of rs10889356 at the 1p31 locus

3.2.1 Functional studies of rs10889356 in hPSCs

I next focused on the second-ranked MPRA-nominated variant, rs10889356 (**Table 1**). rs10889356 is tightly linked to rs2131925 (r^2=0.897), the lead SNP for TG (P=9x10^{-43}), TC and LDL-C in the GLGC study. Possession of the minor allele at rs2131925 (MAF=0.32) is associated with a 4.94 mg/dL decrease in TG.[28] rs10889356 is situated in the promoter of the *DOCK7* gene, which encodes a guanine nucleotide exchange factor that has not previously been implicated in lipid metabolism. *ANGPTL3*, the probable causal gene at this locus, lies within an intron of *DOCK7*, and encodes a liver-specific secreted protein that inhibits EL and LpL, thereby increasing circulating levels of TG and HDL-C (**Figure 8a**).[3]

I hypothesized that rs10889356 is causal for the eQTL at the 1p31 locus, and that it lies within a transcriptional regulatory element to modulate intra-locus gene expression in an allele-specific manner. The MPRA data indicates that the major allele (G) of rs10889356 has enhancer activity relative to the minor allele (A) (**Table 1**). Curiously, *cis*-eQTL data for rs2131925 suggest that expression levels of *DOCK7* and *ANGPTL3* are inversely related in human liver samples, implying that the causal variant variably up or downregulates transcription of different genes at this locus through an unknown mechanism (**Table 2**).

As with rs2277862, I adopted a similar multiplexing approach to efficiently generate homozygous deletion mutants for rs10889356 in H7 cells, which are homozygous major (G/G) at this SNP. I observed some heterogeneity in deletion size, presumably because one of the gRNAs did not always induce a double-stranded break at the predicted location 3-bp upstream from the PAM (**Figure 8b**). For gene expression studies, I utilized hPSC clones harboring a range of 36 to 39-bp homozygous deletions as I was not able to obtain a sufficient number of deletion mutants of one particular genotype. In undifferentiated H7 cells (n=12 wild-type and 8 knockout clones), disruption of the major allele significantly diminished expression of *DOCK7* (down 8.3%), suggesting that the major allele confers enhancer activity (**Figure 9a**). When differentiated to HLCs (n=4 wild-type and 4 knockout clones), rs10889356 major allele knockout cells displayed decreased *DOCK7* expression (down 32%) and increased *ANGPTL3* expression (up 67%), which trended towards but did not achieve statistical significance, presumably due to clone-to-clone variability induced by the differentiation protocol (**Figure 9b**). Average expression of the liver-specific markers *ALB* and *SERPINA1* were equivalent between wild-type and knockout cells.

3.2.2 CRISPR interference at rs10889356

I validated these results through a complementary CRISPR interference strategy in which dCas9 was co-expressed with various gRNAs targeting the rs10889356 locus in HepG2 cultured hepatoma cells (**Figure 10a**). HepG2 cells are homozygous major (G/G) at rs10889356 and are known to have only two copies of chromosome 1. I reasoned that use of a hepatoma cell line may circumvent some of the inconsistencies associated with the HLC differentiation protocol; moreover, *ANGPTL3* expression is liver-specific. Due to suboptimal transfection efficiency, positive transfectants were isolated by FACS prior to gene expression analysis. With three different gRNAs introduced either singly or in combination, expression of *DOCK7* was significantly decreased and expression of *ANGPTL3* was increased relative to control cells, which received the dCas9 construct without an accompanying gRNA (**Figures 9b**). This result is directionally consistent with the HLC experiment, which was also performed on the rs10889356 homozygous major background. Additionally, both experiments recapitulate the inverse relationship between *DOCK7* and *ANGPTL3* expression levels revealed by the human eQTL data for the lead SNP rs2131925. Collectively, these data support a causal role for rs10889356 in gene expression at the 1p31 locus.

3.3 Functional analysis of rs12740374 at the 1p13 locus

3.3.1 Functional studies of rs12740374 in primary human hepatocytes

One of the most robust eQTL associations from the GLGC study is a variant on chromosome 1p13 associated with LDL-C (P=1x10^{-170}) and CHD. Possession of the minor allele at the 1p13 lead SNP is associated with a 5.65 mg/dL decrease in LDL-C and protection against CHD and myocardial infarction. The strongest GWAS-implicated SNPs comprise a haplotype that maps to a noncoding region between the genes *CELSR2* and *PSRC1*, and the locus contains five other genes, which are *SORT1*, *SARS*, *MYBPHL*, *PSMA5*, and *SYPL2*. In eQTL studies using human liver biopsies, the minor allele haplotype is associated with increased expression of *CELSR2*, *SORT1*, and *PSRC1*, with expression of the latter two genes increasing approximately six-fold with each copy of the minor allele (note that the percent gene expression changes from the GLGC eQTL study shown in **Table 2** underestimate the actual effect of the minor allele). However, the eQTL relationship is specific to liver, and is not present in omental or subcutaneous adipose.[28,33,57]

Given the strong association with both LDL-C and cardiovascular disease risk, the 1p13 locus was the focus of early efforts to investigate genotype-to-phenotype connections at GWAS-implicated loci. In a seminal study, Musunuru et al used a combination of fine mapping and luciferase reporter experiments to provide indirect evidence that the noncoding variant rs12740374 was the likely causal DNA variant mediating the GWAS association. A series of experiments suggested that the minor allele of rs12740374 creates a binding site for the CCAAT/enhancer binding protein (C/EBP) transcription factors, resulting in increased liver-specific expression of the 1p13 genes, including *SORT1*. Increased expression of *SORT1*, which decreases hepatic VLDL secretion and promotes LDL particle clearance, diminishes circulating levels of LDL-C, thereby accounting for the cardioprotective effect of the 1p13 minor allele.[33,58]

The probable 1p13 causal variant, rs12740374, maps to the 3' UTR of *CELSR2* and lies approximately 120 kb away from the *SORT1* promoter (**Figure 11a**). In principle, direct demonstration of causality for rs12740374 would entail manipulating the SNP in its native genomic context and evaluating the effect on hepatic *SORT1* expression. However, several factors have complicated efforts to perform this ideal experiment. First, because the eQTL association is liver-specific, studies must be performed in an authentic hepatocyte model system to discern gene expression changes. Readily available human hepatoma cell lines (i.e. HepG2, HuH-7, and Hep3B) possess only the major allele at rs12740374, which has no inherent

transcriptional activity. Moreover, like all cultured cells, these lines are plagued by karyotypic aberrations that limit their utility for well-controlled genetic studies. Second, the sequence flanking rs12740374, including the C/EBP binding site, is poorly conserved in the orthologous DNA region in mouse, precluding the use of wild-type mice to model the eQTL association. Third, *SORT1* expression in undifferentiated hPSCs far exceeds that in differentiated HLCs. Therefore, gene expression studies in HLCs derived from hPSCs with alternate genotypes at rs12740374 have proven unsuccessful, as residual numbers of incompletely differentiated cells in the population will mask HLC-specific gene expression changes (Derek Peters, personal communication).

Because study of this locus is not amenable to the experimental approaches I previously used for two MPRA-nominated variants (i.e. hPSC genome editing, CRISPR interference, knock-in mouse model), I first sought to model the effect of the 1p13 SNP by CRISPR/Cas9 genome editing in primary human hepatocytes with the rs12740374 minor allele. I initially genotyped 20 individual lots of primary human hepatocytes at rs12740374; fourteen lots were homozygous for the major allele (G/G), six were heterozygous (G/T), and none were homozygous for the minor allele (T/T). These frequencies were roughly consistent with the expected genotypic distribution for this SNP based on its MAF of 0.22.

In human liver samples, the 1p13 minor haplotype is strongly associated with increased *SORT1* and *PSRC1* expression, moderately associated with increased *CELSR2* expression, and not associated with changes in *SARS*, *PSMA5*, *MYBPHL*, or S*YPL2* expression (while data from the GLGC study shown in **Table 2** suggests that *PSMA5* and *SYPL2* have weak eQTLs in human liver, this association has not borne out in subsequent eQTL analyses).[28,33] Three days after replating, primary human hepatocytes heterozygous for rs12740374 exhibited 10-fold higher *SORT1* expression and 2-fold higher *PSRC1* expression compared to homozygous major cells (**Figure 11b)**. *CELSR2* was not assayed due to low expression levels in primary human hepatocytes. While these data do suggest that one of the variants in the 1p13 minor haplotype is causal for the gene expression change, they do not directly implicate rs12740374 as the causal variant at this locus.

I next used CRISPR/Cas9 to disrupt the rs12740374 minor allele site in primary human hepatocytes heterozygous at this SNP. To minimize the likelihood of targeting the major allele, I selected a gRNA whose protospacer overlaps the rs12740374 minor allele (**Figure 12a**). Because

primary human hepatocytes exhibit low gene transfer efficiencies and survive for only several days in culture, I used adenovirus to facilitate rapid delivery and expression of Cas9. Given the technical challenges of genome editing in primary cells, I anticipated that indel rates would be low, but surmised that the low efficiency would be offset to some degree by the substantial effect size of the 1p13 SNP. One day after plating, six different lots of primary human hepatocytes heterozygous at rs12740374 were transduced with CRISPR-control or CRISPR-1p13. Two days later, cells were harvested for gene expression experiments. Deep sequencing of the rs12740374 site revealed a 6% indel rate on minor allele-containing chromosomes, with no discernable mutagenesis of the major allele.

Despite the low efficiency of NHEJ, I nonetheless observed reproducible and statistically significant reductions in *SORT1* and *PSRC1* expression in CRISPR-1p13-treated cells, relative to CRISPR-control-treated cells. This effect was consistent across each of the six heterozygous cell lines. When data from all six lots were aggregated (n=19 paired sets of CRISPR-control and CRISPR-1p13-treated cells; variable numbers of wells were plated from each lot depending on number of viable cells in the original vial), I observed a statistically significant 11% reduction in *SORT1* expression and 20% reduction in *PSRC1* expression, with non-significant changes in expression of *SARS* and *PSMA5*, neither of which exhibit an eQTL in human liver (**Figure 12b-d**).

3.3.2 Functional studies of rs12740374 in BAC transgenic mice

In a complementary approach, I developed a murine model for the 1p13 locus to overcome the gene delivery challenges associated with primary cells cultured *in vitro*. As noted previously, the genomic sequence surrounding rs12740374 is poorly conserved in the orthologous region in mouse. To overcome this issue, I generated BAC transgenic mice in which the human 1p13 locus was randomly integrated into the mouse genome. Of note, use of this locus-humanized mouse model assumes conservation of the 1p13 transcriptional machinery between mouse and human, which has not been previously documented in the literature. BAC clone RP11-463O24, which contains the rs12740374 minor allele as well as the entire coding sequences of the *SORT1*, *PSRC1*, *CELSR2*, *SARS*, and *MYBPHL* genes (**Figures 13a, b**), was injected into one-cell mouse embryos. Ten founder mice positive for the BAC transgene by PCR were obtained. Transgene stability was verified by PCR amplification of 11 sites distributed across the BAC insert. Relative copy number of the BAC insert was assessed by qPCR, and the

founder mouse with the lowest copy number of the full insert was selected for breeding. Littermates from the F2 generation were selected for experiments.

For somatic *in vivo* genome editing experiments, two to three month old mice were administered CRISPR-control (*n*=5) or CRISPR-1p13 (*n*=7) adenovirus. Mice were sacrificed four days following injection and liver was harvested for gene expression analysis. CEL I nuclease assay revealed significant levels of mutagenesis across all CRISPR-1p13-treated mice (**Figure 13c**); deep sequencing of a representative sample demonstrated a 33% indel rate. I observed a statistically significant 60% reduction in human *SORT1* expression and non-significant 30% reduction in *PSRC1* expression in CRISPR-1p13-treated mice relative to CRISPR-control-treated mice (*CELSR2* levels were again not measured due to low expression in mouse liver). Expression of the non-eQTL gene *SARS* was roughly equivalent between groups (**Figure 13d**). Collectively, these data support the association of the 1p13 minor haplotype with hepatic *SORT1* expression, and provide suggestive evidence that the minor allele of rs12740374 functions as a transcriptional enhancer at this locus.

4. DISCUSSION

4.1 Conclusions and Limitations

One of the principal challenges of the post-GWAS era has been cataloguing the allelic spectrum of causal variants underlying complex trait susceptibility. In this work, I describe a novel methodological framework for causal variant discovery that involves high-throughput identification of putative disease-causal loci through a functional reporter-based screen, MPRA. I then employ a combination of genome edited stem cells and primary cells, CRISPR interference, and locus-humanized mouse models to provide functional evidence of causality at three different eQTL loci for blood lipids.

I first focused on the MPRA-nominated variant rs2277862 at the 20q11 cholesterol locus, which harbors several genes with no prior connection to lipid metabolism. A series of four CRISPR-based experiments in knock-in hPSCs, knockout hPSCs, cultured cell lines, and locus-humanized mice consistently demonstrated that the major allele of rs2277862 has transcriptional enhancer activity. The gene expression effect was particularly notable for *CPNE1*, which exhibited altered expression in all four experiments, including those performed in hPSC-derived HLCs and mouse liver. Interestingly, the *CPNE1* promoter is situated ~100 kb downstream of the SNP, implicating long-range enhancer interactions in the mechanism of this eQTL. Of note, overexpression of *CPNE1* in mouse liver decreases serum HDL-C levels (Dr. Daniel Rader, personal communication), highlighting rs2277862-*CPNE1* as a novel causal SNP-gene set involved in the regulation of cholesterol metabolism. *CPNE1* encodes a calcium-dependent membrane binding protein, but its biological function is otherwise poorly understood.

I next evaluated another MRPA-implicated variant, rs10889356, at the 1p31 lipid locus. *ANGPTL3* is the probable causal gene at this locus, as humans with rare loss-of-functions mutations in this gene exhibit abnormally low levels of LDL-C, HDL-C and TG.[59] Genome editing and CRISPR interference-based experiments both ascribed regulatory activity to the major allele of rs10889356, which was found to enhance *DOCK7* expression but diminish *ANGPTL3* expression. This inverse relationship may be related to the relative positions of the two genes; because *ANGPTL3* is found within an intron of the *DOCK7* gene, increased *DOCK7* transcription could theoretically interfere with *ANGPTL3* transcription from the complementary DNA strand. While these results suggest rs10889356-*ANGPTL3* as a potential causal SNP-gene

set, further experiments in knock-in cell lines or mice are required to unequivocally assign causality to this SNP.

Lastly, I investigated a SNP on chromosome 1p13, rs12740374, which is believed to regulate liver-specific expression of *SORT1* and thus influence LDL-C metabolism.[33] Disrupting the putative regulatory site by genome editing of primary human hepatocytes *in vitro* and BAC transgenic mice *in vivo* supported a role for the rs12740374 minor allele as a transcriptional enhancer of *SORT1* expression in hepatocytes. Importantly, these two model systems mitigated several of the limitations associated with studying liver-specific eQTLs, and lent further credence to the hypothesized causal link between rs12740374 and *SORT1* previously uncovered by traditional reporter-based methods.

Collectively, these results highlight a number of important considerations for disease modeling in common variant genetics. First, MPRA provides a means for highly efficient and scalable causal variant discovery but must be paired with functional approaches to directly demonstrate causality. When MPRA contradicts findings in biological models (i.e. for rs2277862, MPRA suggested that the minor allele has enhancer activity relative to the major allele, whereas the opposite was true in cellular models), the results of functional studies should always take precedence. This is because MPRA artificially assesses regulatory elements outside of their native genomic context, versus genome editing or CRISPR interference approaches that are targeted to endogenous sequence. Furthermore, MPRA likely provides an incomplete view of genetic regulation, since the assay is not designed to identify causal haplotypes but rather only individual causal alleles, an assumption that oversimplifies the relationship between disease-causal variants and causal genes.[60]

For functional genetic studies, hPSCs are theoretically an attractive system in which to interrogate a putative disease-causal variant, though several limitations should be noted.[52] Although genetic knock-ins are the gold standard for ascertaining the regulatory significance of a variant, HDR efficiency in hPSCs is exceedingly low. However, given the modest effect sizes of most common variants, maximizing sample size is often critical to obtaining a statistically robust result. This leads to a reliance on cruder "knockout" models, which can efficiently be generated in large numbers through NHEJ but may provide imprecise information about the direct contribution of the SNP to gene expression, as the flanking regulatory sequence is modified along with the SNP itself. Inconsistent differentiation protocols may further confound efforts to

detect small gene expression changes by introducing a significant degree of clone-to-clone variability in gene expression. For example, in my experiments in rs10889356 knockout HLCs, a high level of inter-clonal heterogeneity likely concealed an effect on *DOCK7* expression that had been observed previously in undifferentiated cells.

This is not to say that hPSCs are a poor platform for disease modeling. hPSCs offer many attractive advantages over existing model systems, including genetic stability, multipotency and renewability.[52] However, it would be prudent to consider the estimated "signal" from a putative disease-causal variant relative to the estimated "noise" from a directed differentiation protocol prior to pursuing a disease modeling experiment in hPSCs. For highly penetrant variants, hPSCs have been shown to yield informative insights into human genetics. For example, Ding et al have characterized the rare gain-of-function variant E17K in the gene *AKT2*, which results in hypoglycemia, hypoinsulinemia and increased body fat secondary to impaired insulin signaling. hPSC-derived *AKT2* E17K knock-in HLCs displayed significantly decreased glucose production, while hPSC-derived adipocytes had increased TG content and increased glucose uptake compared to matched wild-type controls. Through disease modeling in hPSCs, Ding et al unequivocally established a dominant activating role for the *AKT2* E17K mutation.[53] However, GWAS-implicated common variants often do not display the disproportionately large effect sizes of rare variants, suggesting that careful pre-selection of variants with robust supportive evidence (i.e through eQTL association studies or MPRA) will maximize the likelihood of a successful experiment.

In some respects, primary cells are an appealing alternative to differentiated hPSCs, which can create immature and inauthentic replicas of human tissue. This is an important consideration for eQTL experiments, as a gene expression effect may not be apparent without the presence of the appropriate tissue-specific transcriptional machinery. The 1p13 locus was an attractive candidate for modeling in primary human hepatocytes, given the liver-specific association of rs12740374 with *SORT1* as well as the unique challenges that preclude modeling this SNP in other systems (i.e. lack of sequence conservation with the mouse genome and absence of cultured hepatoma lines with the rs12740374 minor allele). To improve the efficiency of gene transfer, I used adenovirus to deliver CRISPR/Cas9 into primary human hepatocytes (which, in theory, can infect up to 100% of primary hepatocytes *in vitro*),[61] but nonetheless observed only minimal levels of mutagenesis (~6%). Remarkably, this low level of genome

editing was sufficient to evoke highly reproducible decreases in *SORT1* expression, likely due to the substantial effect size of the SNP. However, for SNPs of low effect size, studies in primary cells may be futile unless the gene delivery strategy is optimized. Moreover, primary cells have a short lifespan in culture, which limits the length of exposure to CRISPR/Cas9 and further diminishes genome editing efficiency. Finally, this approach does not allow for clean substitution of one allelic variant for another, but rather relies on NHEJ to randomly generate indels at the regulatory site. Due to the additional mutations in the sequence flanking the SNP, it becomes more difficult to attribute the gene expression change directly to the SNP itself. In theory, complementary experiments using a gRNA targeted to the transcriptionally inactive alternate allele would control for the nonspecific effects of the indel on gene expression.

Locus-humanized mouse models offer yet another platform in which to interrogate human genetic variation. For rs2277862, I fortuitously noted that the human SNP has an orthologous counterpart in mouse, rs27324996, and that the overall genetic architecture of the human and mouse loci are conserved. As a result of this high degree of similarity, I was able to develop a humanized knock-in mouse model for the minor allele of rs2277862/rs27324996, and evaluate the effect of the SNP on murine gene expression. However, unlike coding sequence, regulatory sequence is rarely conserved between mouse and human, making this strategy fairly non-generalizable. Indeed, for rs12740374, the orthologous mouse locus lacked the C/EBP transcriptional binding site altogether. To overcome this limitation, I introduced the entire human 1p13 locus into wild-type mice via BAC transgenesis, and used CRISPR/Cas9 to disrupt the human SNP and evaluate the effect on human gene expression. Both of these experimental approaches – the knock-in mouse and the BAC transgenic mouse – yielded biologically informative insights that were concordant with the data obtained in cell-based experiments. Advantages of both systems included the large supply of primary tissues as well as the ability to increase sample size, and thus statistical power, by breeding large numbers of mice. However, use of a humanized mouse model assumes that transcriptional machinery is conserved between mouse and human – that is to say, murine transcription factors must be able to recognize and modulate gene expression from human regulatory sequence. Of note, I still observed a fair degree of variance in both *in vivo* experiments, particularly in adipose tissue, which may have confounded gene expression analysis. Disadvantages unique to the BAC transgenic mouse model include the random nature of BAC incorporation into the genome and susceptibility to position

effects – drawbacks that could potentially be mitigated by developing a mouse model in which the human locus replaces the orthologous mouse locus, although this remains a technically challenging task.[62,63]

In a parallel approach, I also explored the utility of CRISPR-based transcriptional modulation as an adjunct to genome editing experiments. If a variant is truly causal and lies within a transcriptional regulatory element, then artificially activating or repressing the site should, in theory, affect expression of downstream gene targets. Rather than exogenously modulating the site via a transactivator or repressor domain, I instead targeted a plain dCas9 protein to each SNP site to sterically interfere with recruitment of native transcriptional machinery to the locus and thereby antagonize the natural function of the SNP.[45] Although I obtained data that were directionally consistent with the genome editing studies for rs2277862 and rs10889356, this type of study on its own is insufficient to demonstrate causality for various reasons. Firstly, cultured cell lines are not an ideal system in which to study human genetic regulation due to aberrancies in karyotype and ploidy that may complicate interpretation of results. Secondly, absence of a transcriptional effect does not imply that a variant is not causal, but may simply reflect poor gRNA targeting or ineffective transcriptional activation or interference at that site due to spatial orientation of the dCas9:gRNA complex and its associated domains. Thirdly, the presence of a transcriptional effect also does not imply causality, since the transcriptional perturbation is not specific to the variant itself. Given the size of dCas9, its presence at a particular locus could obstruct binding of transcription factors not only at the gRNA recognition site, but also at more distal regulatory elements. Thus, a statistically significant gene expression difference in the context of CRISPR interference merely implies a regulatory role for the general genomic region flanking the SNP, but does not suggest that allelic variation at the SNP itself underlies this gene expression change. The CRISPR activation platform is a still more artificial means of showing causality as the strong transactivation domain may induce gene expression from sites that normally do not have regulatory function. Moreover, this approach may not permit discrimination between enhancer and repressors due to indiscriminate activation of all targeted sequence elements. Thus, while programmable dCas9-mediated activation and repression of endogenous regulatory elements can provide an interesting corollary to more rigorous genome-editing experiments, the latter remains a more reliable means of discerning causality.

Off-target activity is always a consideration in all CRISPR-based experiments. Reassuringly, my laboratory and others have performed whole-genome sequencing analysis of multiple CRISPR-targeted hPSC lines and found minimal evidence for off-target mutagenesis.[64,65] Because multiple lines are unlikely to harbor the same extraneous mutations, I likely counteracted the effect of any off-target mutations that did occur by using multiple wild-type and mutant clones for gene expression experiments. Moreover, since I am investigating a *cis*-regulatory eQTL, and thus intra-locus gene expression effects, distant off-target mutations, if they exist, should not significantly perturb local gene expression. In fact, rather than off-target mutations, accumulations of unique single nucleotide variants (SNVs) in each cell line are likely the main contributor to clonal heterogeneity. Whole-genome sequencing of genome edited hPSC lines indicates that each clone harbors hundreds of unique SNVs that likely arise secondary to the clonal selection procedure and prolonged maintenance in culture.[53,64,65] Thus, even in the setting of highly specific genome editing technologies, it may be impossible to derive fully isogenic hPSC lines, indicating that some level of clonal phenotypic variation is unavoidable.

CRISPR/Cas9 technology has greatly expanded the availability of model systems in which to rigorously model human eQTLs. Given the inherent advantages and disadvantages associated with each model system described above, selection of the most appropriate system depends on careful consideration of the factors (i.e. tissue specificity, magnitude of anticipated gene expression change) that may influence the ability to detect a statistically significant effect. Pursuing several complementary approaches, as illustrated in this thesis, allows the advantages and disadvantages of each system to offset one another, and increases confidence in the gene expression result.

Recently, two different groups have reported the use of MPRA to discover causal regulatory pathways underlying other complex traits in humans. In one study, Ulirsch et al employed MPRA to rapidly screen all GWAS-identified variants for red blood cell traits, and discovered numerous regulatory variants. By complementing these findings with CRISPR/Cas9 genome editing of putative regulatory elements in erythroid cells, Ulirsch et al uncovered several novel transcriptional pathways underlying erythrocyte biology.[66] In another study, Tewhey et al applied MPRA to tens of thousands of variants associated with eQTLs in lymphoblastoid cell lines, and found hundreds of putative causal alleles, many of which overlapped known GWAS loci. In CRISPR/Cas9 genome editing experiments, a noncoding risk allele associated with

ankylosing spondylitis was shown to directly influence expression of the prostaglandin receptor.[67] Both papers highlight the broad applicability of high-throughput reporter screens for identifying putative causal alleles relevant to human disease, and, like this study, provide a methodological framework that will likely accelerate the pace of causal variant discovery.

4.2 Future directions

Areas for further investigation include (1) ongoing investigation of causality at MPRA-nominated loci (2) defining mechanisms for allele-specific transcriptional activity at causal SNPs, and (3) improving the ability of MPRA to detect *bona fide* regulatory variants.

First, for loci investigated in this study, further experimentation focused on knock-in models (whether cells or mice) will provide the clearest evidence for causality. For rs10889356, evaluation of a knock-in hPSC line may provide more precise insights than the knockout model. For rs12740374, germline editing of the BAC transgenic mouse to knock-in the alternate 1p13 allele may provide more robust data than somatic genome editing, which targets a smaller proportion of cells by NHEJ. Similar methods can be applied to other high-priority MPRA-nominated loci.

Second, to elucidate mechanisms for allele-specific transcriptional activity, motif prediction programs can be used to identify transcription factors that differentially recognize regulatory sequence depending on the allelic variant of the SNP. For example, the minor allele of rs2277862 is predicted to create a binding site for the transcription factor Yin-Yang-1 (YYI) (consensus binding sequence 5'-(C/g/a)(G/t)(C/t/a)CATN(T/a)(T/g/c)-3'),[68] while the major allele abolishes it. Electrophoretic mobility shift assays (EMSA) and ChIP can then be used to validate this prediction. These methods can then be extended to other eQTL loci to uncover the transcriptional mechanisms mediating the eQTL associations.

Third, technological improvements such as including more barcodes, using longer genomic tiles, or switching promoters will likely improve the sensitivity of MPRA for detecting true regulatory variants.[66] Additionally, redesigning the MPRA to interrogate variants within their native genomic context may improve the ability of the assay to capture long-range transcriptional interactions that otherwise would be missed. In fact, several groups have reported the development of CRISPR-based screens targeted to non-coding DNA elements, and have successfully identified novel proximal and distal regulatory elements.[69-72] One approach is to perform saturation mutagenesis of the *cis*-regulatory genomic region associated with a gene of

interest, which is tagged with a fluorescent reporter as a proxy for gene activity.[72] In the future, similar techniques could be applied to GWAS-implicated genes to facilitate discovery of causal *cis*-regulatory transcriptional elements.

4.3 Summary

Genome-wide association studies (GWASs) have identified a number of novel genetic loci linked to serum cholesterol and triglyceride levels. However, the majority of implicated variants are non-coding, requiring a combination of fine mapping, chromatin state mapping and luciferase-based reporter assays to ascertain their influence on expression of disease-causal genes. In this work, I sought to develop an alternate methodology for causal variant discovery, in which putative disease-causal loci were first rapidly identified through a functional reporter-based screen, termed MPRA. I then utilized CRISPR/Cas9 technology to perform a series of experiments in cultured cell lines, primary cells, stem cells and humanized mice to rigorously validate prioritized variants. Collectively, I was able to define causal pathways at three different lipid-associated loci, in which I confirmed the causal variant, the transcriptional effect of each allele, and the target genes regulated by the variant. This work offers fresh insight into the mechanism by which GWAS-implicated SNPs influence expression of causal genes for lipid metabolism. More broadly, these results highlight a novel experimental framework to discover causal genes and variants contributing to complex human traits.

AUTHOR CONTRIBUTIONS

I performed the work described in this thesis under the guidance of my thesis advisor, Dr. Kiran Musunuru. Dr. Musunuru initially conceived the research project, and supervised experimental design, data analysis, and drafting of this thesis. The MPRA experiments, which formed the basis for many of the functional experiments described in this thesis, was designed, performed and analyzed by Alexandre Melnikov, Peter Rogov, Li Wang, and Xiaolan Zhang under the supervision of Dr. Tarjei Mikkelsen at the Broad Institute.

My role in this project involved designing experiments, generating cellular and mouse models, performing and trouble-shooting experiments, and analyzing data. Derek Peters and Qiurong Ding oriented me to the laboratory when I first joined and taught me how to maintain and differentiate hPSC cultures. I independently generated all four genome edited hPSC lines described in this thesis. I performed the differentiation and gene expression experiments described for those lines with the exception of those depicted in **Figures 3a** and **3b**, which were performed by Xiao Wang. I performed and analyzed the CRISPR interference experiments. I generated the rs2277862 knock-in mouse with the assistance of the Harvard University Genome Modification Facility. Tao Chen provided assistance with mouse breeding and colony maintenance. I collected the samples used for the experiment shown in **Figure 7**; Xiao Wang performed the gene expression experiment. I performed and analyzed all experiments done in human primary hepatocytes. Derek Peters generated the 1p13 BAC transgenic mouse with the help of the Harvard University Genome Modification Facility and screened the founder mice; I subsequently expanded the colony with assistance from Tao Chen. Xiao Wang assisted with adenoviral injections and sample collection for the BAC transgenic mouse experiments. Kiran Musunuru performed the deep sequencing analysis for the 1p13 experiments and kindly provided the schematics shown in Figures 2(a), 6(a), 8(a), 11(a), and 13(a). These individual contributions have also been noted in the text of each figure legend as appropriate.

REFERENCES

1. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. Lancet. 1999 Jan 9;353(9147):89-92.
2. Rader DJ, Daugherty A. Translating molecular discoveries into new therapies for atherosclerosis. Nature. 2008 Feb 21;451(7181):904-13.
3. Ballantyne CM. *Clinical Lipidology*. 1st ed. China: Elsevier Health Sciences;2009.608p.
4. Castelli WP, Garrison RJ, Wilson PW, Abbott RD, Kalousdian S, et al. Incidence of coronary heart disease and lipoprotein cholesterol levels The Framingham Study. JAMA. 1986 Nov 28;256(20):2835-8.
5. Brown MS, Goldstein JL. Heart attacks: gone with the century?. Science. 1996 May 3;272(5262):629.
6. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet. 2012 Aug 11;380(9841):572-80.
7. Barter PJ, Caulfield M, Eriksson M, Grundy SM, Kastelein JJ, et al. Effects of torcetrapib in patients at high risk for coronary events. N Engl J Med. 2007 Nov 22;357(21):2109-22.
8. Rader DJ, deGoma EM. Future of cholesteryl ester transfer protein inhibitors. Annu Rev Med. 2014;65:385-403.
9. Khera AV, Cuchel M, de la Llera-Moya M, Rodrigues A, Burke MF, et al. Cholesterol efflux capacity, high-density lipoprotein function, and atherosclerosis. N Engl J Med. 2011 Jan 13;364(2):127-35.
10. deGoma EM, deGoma RL, Rader DJ. Beyond high-density lipoprotein cholesterol levels evaluating high-density lipoprotein function as influenced by novel therapeutic approaches. J Am Coll Cardiol. 2008 Jun 10;51(23):2199-211.
11. Kohli P, Cannon CP. Triglycerides: how much credit do they deserve?. Med Clin North Am. 2012 Jan;96(1):39-55.
12. Miller M, Cannon CP, Murphy SA, Qin J, Ray KK, et al. Impact of triglyceride levels beyond low-density lipoprotein cholesterol after acute coronary syndrome in the PROVE IT-TIMI 22 trial. J Am Coll Cardiol. 2008 Feb 19;51(7):724-30.
13. Cuchel M, Rader DJ. Macrophage reverse cholesterol transport: key to the regression of atherosclerosis?. Circulation. 2006 May 30;113(21):2548-55.
14. Rader DJ. Molecular regulation of HDL metabolism and function: implications for novel therapies. J Clin Invest. 2006 Dec;116(12):3090-100.
15. Zannis VI, Chroni A, Krieger M. Role of apoA-I, ABCA1, LCAT, and SR-BI in the biogenesis of HDL. J Mol Med (Berl). 2006 Apr;84(4):276-94.
16. Heller DA, de Faire U, Pedersen NL, Dahlén G, McClearn GE. Genetic and environmental influences on serum lipid levels in twins. N Engl J Med. 1993 Apr 22;328(16):1150-6.
17. Rao DC, Laskarzewski PM, Morrison JA, Khoury P, Kelly K, et al. The Cincinnati Lipid Research Clinic family study: cultural and biological determinants of lipids and lipoprotein concentrations. Am J Hum Genet. 1982 Nov;34(6):888-903.
18. Goldstein JL, Brown MS. The LDL receptor. Arterioscler Thromb Vasc Biol. 2009 Apr;29(4):431-8.

19. Defesche JC, Pricker KL, Hayden MR, van der Ende BE, Kastelein JJ. Familial defective apolipoprotein B-100 is clinically indistinguishable from familial hypercholesterolemia. Arch Intern Med. 1993 Oct 25;153(20):2349-56.
20. Abifadel M, Varret M, Rabès JP, Allard D, Ouguerram K, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003 Jun;34(2):154-6.
21. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008 Jun;40(6):695-701.
22. Weissglas-Volkov D, Pajukanta P. Genetic causes of high and low serum HDL-cholesterol. J Lipid Res. 2010 Aug;51(8):2032-57.
23. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet. 2002 May;3(5):391-7.
24. Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. Cell. 2011 Sep 30;147(1):57-69.
25. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004 Aug 6;305(5685):869-72.
26. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010 Jul 8;363(2):166-76.
27. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet. 2005 Feb;6(2):109-18.
28. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010 Aug 5;466(7307):707-13.
29. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013 Nov 7;93(5):779-97.
30. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008 Oct 15;17(R2):R156-65.
31. Battle A, Montgomery SB. Determining causality and consequence of expression quantitative trait loci. Hum Genet. 2014 Jun;133(6):727-35.
32. Gupta RM, Musunuru K. Mapping Novel Pathways in Cardiovascular Disease Using eQTL Data: The Past, Present, and Future of Gene Expression Analysis. Front Genet. 2012;3:232.
33. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010 Aug 5;466(7307):714-9.
34. Holdt LM, Teupser D. From genotype to phenotype in human atherosclerosis--recent findings. Curr Opin Lipidol. 2013 Oct;24(5):410-8.
35. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 2010 Jan 26;8(1):e1000294.
36. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53.
37. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012 Jan 13;90(1):7-24.
38. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012 Feb 26;30(3):271-7.

39. Peters DT, Musunuru K. Functional evaluation of genetic variation in complex human traits. Hum Mol Genet. 2012 Oct 15;21(R1):R18-23.
40. Mali P, Yang L, Esvelt KM, Aach J, Guell M, et al. RNA-guided human genome engineering via Cas9. Science. 2013 Feb 15;339(6121):823-6.
41. Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013 Feb 15;339(6121):819-23.
42. Sternberg SH, Doudna JA. Expanding the Biologist's Toolkit with CRISPR-Cas9. Mol Cell. 2015 May 21;58(4):568-574.
43. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014 Jan 3;343(6166):80-4.
44. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014 Jan 3;343(6166):84-7.
45. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013 Feb 28;152(5):1173-83.
46. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014 Oct 23;159(3):647-61.
47. Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature. 2015 Apr 9;520(7546):186-91.
48. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell. 2013 Dec 19;155(7):1479-91.
49. Ahfeldt T, Schinzel RT, Lee YK, Hendrickson D, Kaplan A, et al. Programming human pluripotent stem cells into white and brown adipocytes. Nat Cell Biol. 2012 Jan 15;14(2):209-19.
50. Si-Tayeb K, Noto FK, Nagaoka M, Li J, Battle MA, et al. Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. Hepatology. 2010 Jan;51(1):297-305.
51. Ding Q, Strong A, Patel KM, Ng SL, Gosis BS, et al. Permanent alteration of PCSK9 with in vivo CRISPR-Cas9 genome editing. Circ Res. 2014 Aug 15;115(5):488-92.
52. Musunuru K. Genome editing of human pluripotent stem cells to generate human cellular disease models. Dis Model Mech. 2013 Jul;6(4):896-904.
53. Ding Q, Lee YK, Schaefer EA, Peters DT, Veres A, et al. A TALEN genome-editing system for generating human stem cell-based disease models. Cell Stem Cell. 2013 Feb 7;12(2):238-51.
54. Ding Q, Regan SN, Xia Y, Oostrom LA, Cowan CA, et al. Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. Cell Stem Cell. 2013 Apr 4;12(4):393-4.
55. Cheng AW, Wang H, Yang H, Shi L, Katz Y, et al. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. Cell Res. 2013 Oct;23(10):1163-71.
56. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell. 2013 May 9;153(4):910-8.

57. Strong A, Patel K, Rader DJ. Sortilin and lipoprotein metabolism: making sense out of complexity. Curr Opin Lipidol. 2014 Oct;25(5):350-7.
58. Strong A, Ding Q, Edmondson AC, Millar JS, Sachs KV, et al. Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. J Clin Invest. 2012 Aug;122(8):2807-16.
59. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med. 2010 Dec 2;363(23):2220-7.
60. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 2014 Jan;24(1):1-13.
61. Morsy MA, Alford EL, Bett A, Graham FL, Caskey CT. Efficient adenoviral-mediated ornithine transcarbamylase expression in deficient mouse and human hepatocytes. J Clin Invest. 1993 Sep;92(3):1580-6.
62. Schmouth JF, Bonaguro RJ, Corso-Diaz X, Simpson EM. Modelling human regulatory variation in mouse: finding the function in genome-wide association studies and whole-genome sequencing. PLoS Genet. 2012;8(3):e1002544.
63. Lee EC, Liang Q, Ali H, Bayliss L, Beasley A, et al. Complete humanization of the mouse immunoglobulin loci enables efficient therapeutic antibody discovery. Nat Biotechnol. 2014 Apr;32(4):356-63.
64. Veres A, Gosis BS, Ding Q, Collins R, Ragavendran A, et al. Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. Cell Stem Cell. 2014 Jul 3;15(1):27-30.
65. Suzuki K, Yu C, Qu J, Li M, Yao X, et al. Targeted gene correction minimally impacts whole-genome mutational load in human-disease-specific induced pluripotent stem cell clones. Cell Stem Cell. 2014 Jul 3;15(1):31-6.
66. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, et al. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell. 2016 Jun 2;165(6):1530-45.
67. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016 Jun 2;165(6):1519-29.
68. Houbaviy HB, Usheva A, Shenk T, Burley SK. Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. Proc Natl Acad Sci U S A. 1996 Nov 26;93(24):13577-82.
69. Canver MC, Smith EC, Sher F, Pinello L, Sanjana NE, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. 2015 Nov 12;527(7577):192-7.
70. Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, et al. High-resolution interrogation of functional elements in the noncoding genome. Science. 2016 Sep 30;353(6307):1545-1549.
71. Diao Y, Li B, Meng Z, Jung I, Lee AY, et al. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Res. 2016 Mar;26(3):397-405.
72. Rajagopal N, Srinivasan S, Kooshesh K, Guo Y, Edwards MD, et al. High-throughput mapping of regulatory DNA. Nat Biotechnol. 2016 Feb;34(2):167-74.

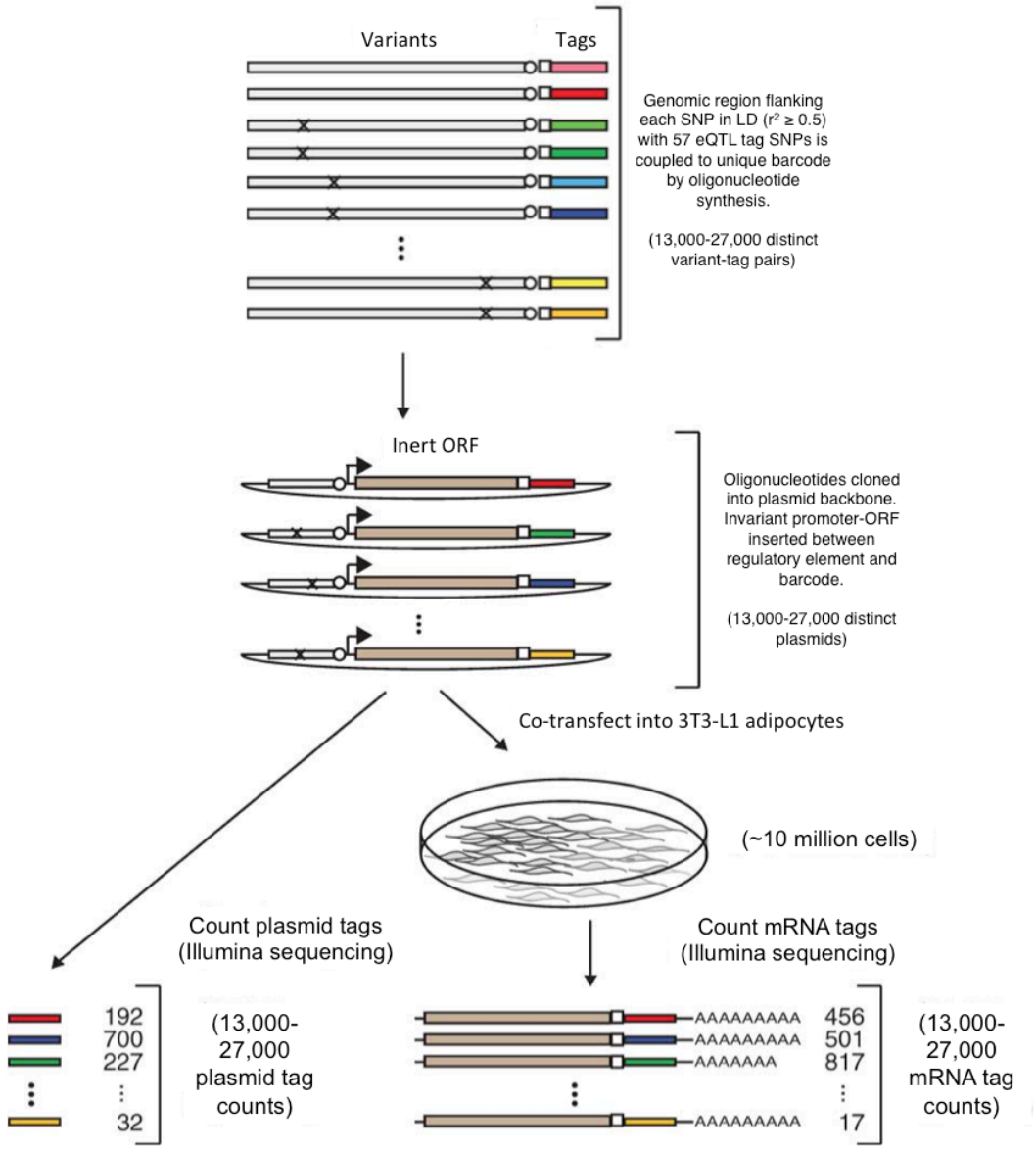


Figure 1. Massively parallel reporter assay (MPRA) identifies putative causal SNPs at lipid-associated eQTL loci. MPRA uses a pool of reporter constructs in which every candidate SNP is coupled to a reporter gene with a unique barcode identifier in the 3' UTR. The construct pool is transfected *en masse* into hepatocyte or adipocyte cultured cell lines, and barcode expression is quantified by RNAseq and normalized to the corresponding level of transfected plasmid DNA. MPRA-identified variants are then prioritized according to allele-specific regulatory activity (figure adapted from Melnikov et al, *Nat Biotechnol* 2012).

		Allele 1 Signal 1	Allele 2 Signal 1	Log ratio 1	*P*-value 1	Allele 1 Signal 2	Allele 2 Signal 2	Log ratio 2	*P*-value 2
rs2277862	*left*	0.12	0.99	-0.87	2E-06	-0.08	1.06	-1.15	3E-06
	center	0.20	1.01	-0.82	2E-04	0.22	1.16	-0.94	2E-05
	right	0.10	1.39	-1.29	2E-07	-0.11	1.32	-1.40	2E-07
rs10889356	*left*	1.90	0.65	1.26	8E-08	1.95	0.40	1.55	6E-08
	center	1.66	0.33	1.33	2E-07	1.70	0.43	1.27	1E-07
	right	0.47	0.18	0.29	0.17	0.45	0.17	0.28	0.27

Table 1. Prioritized variants from MPRA in 3T3-L1 cells. MPRAs identified rs2277862 and rs10889356 as the SNPs with highest allele-specific regulatory activity in mouse 3T3-L1 adipocytes. For this experiment, each candidate SNP was represented on a 145-bp tile that was either centered, left-shifted or right-shifted relative to the SNP, in order to increase the probability of capturing the correct regulatory context for that SNP. For each tile, the individual signals for the two alleles are shown for two independent experiments (where signal refers to the log of median barcode counts for the given tile divided by median barcode counts for all tiles). A positive signal implies enhancer activity, while a negative signal implies repressor activity. In the final two columns for each experiment, a log-ratio of the signals of the two alleles is calculated, along with a *P*-value (by Mann-Whitney *U* test) for the null hypothesis that the two alleles generate equal signals. *This experiment was performed and analyzed by the following individuals in Tarjei Mikkelsen's laboratory at the Broad Institute: Alexandre Melnikov, Peter Rogov, Li Wang, and Xiaolan Zhang.*

Cis-eQTL data for rs2277862

Tissue	Gene	# of samples	Major, minor alleles	Change in gene expression (Relative to C/C)		*P*-value
				C/T	T/T	
Liver	*CEP250*	949	C, T	+16.7%	+20.5%	3E-8
	CPNE1	954	C, T	-12.3%	-20.2%	7E-41
Omental fat	*CEP250*	737	C, T	+17.5%	+34.9%	8E-37
	CPNE1	732	C, T	-17.0%	-28.2%	1E-73
	ERGIC3	731	C, T	-6.5%	-16.1%	2E-11
Subcutaneous fat	*CEP250*	609	C, T	+19.1%	+46.9%	6E-31
	CPNE1	607	C, T	-17.0%	-32.2%	1E-47
	ERGIC3	583	C, T	-7.3%	-10.7%	2E-8

Cis-eQTL data for rs2131925

Tissue	Gene	# of samples	Major, minor alleles	Change in gene expression (Relative to T/T)		*P*-value
				T/G	G/G	
Liver	*ANGPTL3*	924	T, G	-13.7%	-27.1%	1E-13
	DOCK7	952	T, G	+8.64%	+21.9%	1E-22
Omental fat	*DOCK7*	738	T, G	-19.5%	-36.2%	3E-91
Subcutaneous fat	*DOCK7*	608	T, G	-26.5%	-49.6%	6E-80

Cis-eQTL data for rs629301

Tissue	Gene	# of samples	Major, minor alleles	Change in gene expression (Relative to T/T)		*P*-value
				T/G	G/G	
Liver	*CELSR2*	951	T, G	+46.9%	+209%	5E-94
	PSMA5	955	T, G	+11.9%	+15.3%	9E-17
	PSRC1	949	T, G	+289%	+502%	2E-271
	SORT1	951	T, G	+294%	+524%	2E-300
	SYPL2	955	T, G	+20.2%	+47.2%	1E-23

Table 2. *Cis*-acting associations of rs2277862, rs2131925, and rs629301 with transcript levels in human liver, subcutaneous fat and/or omental fat. To determine if lipid-associated variants influence gene expression in a *cis*-regulatory manner, the Global Lipids Genetics Consortium performed *cis*-eQTL analysis in human liver, omental fat and subcutaneous fat biopsies (obtained post-mortem or during surgical resection). Expression of all genes within 500 kb of the lead SNP at each GWAS-implicated lipid locus was profiled, and significant associations between SNP genotype and transcript levels were identified. Lipid-associated eQTLs identified for subcutaneous and/or omental fat were subsequently interrogated by MPRA. eQTL data for rs2277862, which is the lead SNP at the 20q11 locus; rs2131925, which is tightly linked to the MPRA-implicated SNP rs10889356 at the 1p31 locus; and rs629301, which is tightly linked to the plausible causal SNP rs12740374 at the 1p13 locus, is shown above. Gene expression changes are shown relative to major allele homozygotes (data from Teslovich et al, *Nature* 2010).

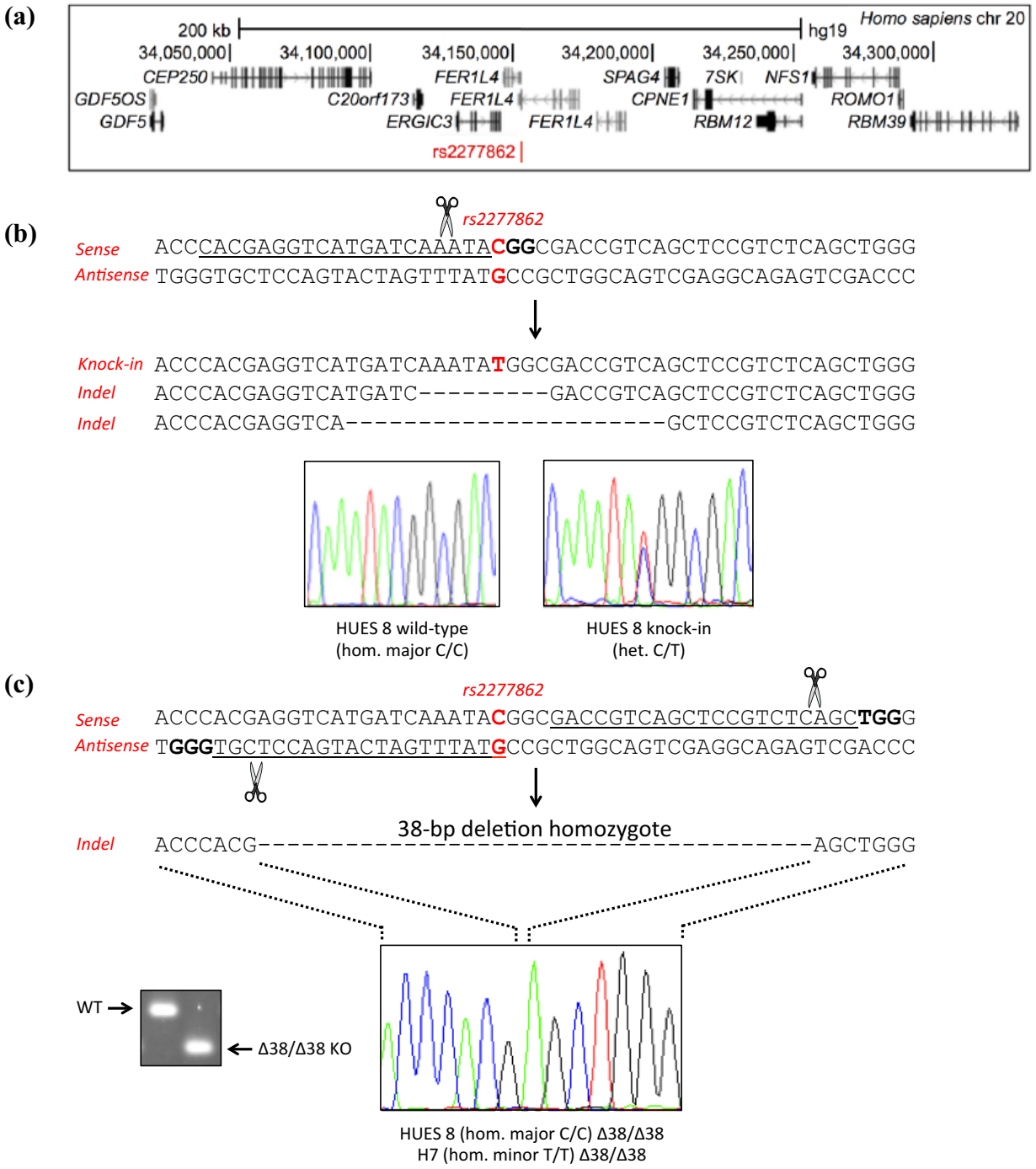


Figure 2. CRISPR/Cas genome editing at the rs2277862 locus in hPSCs. (a) Schematic of the human 20q11 locus (b) Heterozygous rs2277862 minor allele knock-in was generated on the HUES 8 background (homozygous major at rs2277862) using an exogenous ssODN. Representative indels from non-knock-in clones are also shown. (b) Homozygous 38-bp deletions encompassing rs2277862 were generated on the HUES 8 (homozygous major) and H7 (homozygous minor) backgrounds with a dual gRNA approach. Representative agarose gel of PCR amplicons is shown. Guide RNA protospacers are underlined and PAM is bolded. *Cell lines generated by Avanthi Raghavan.*

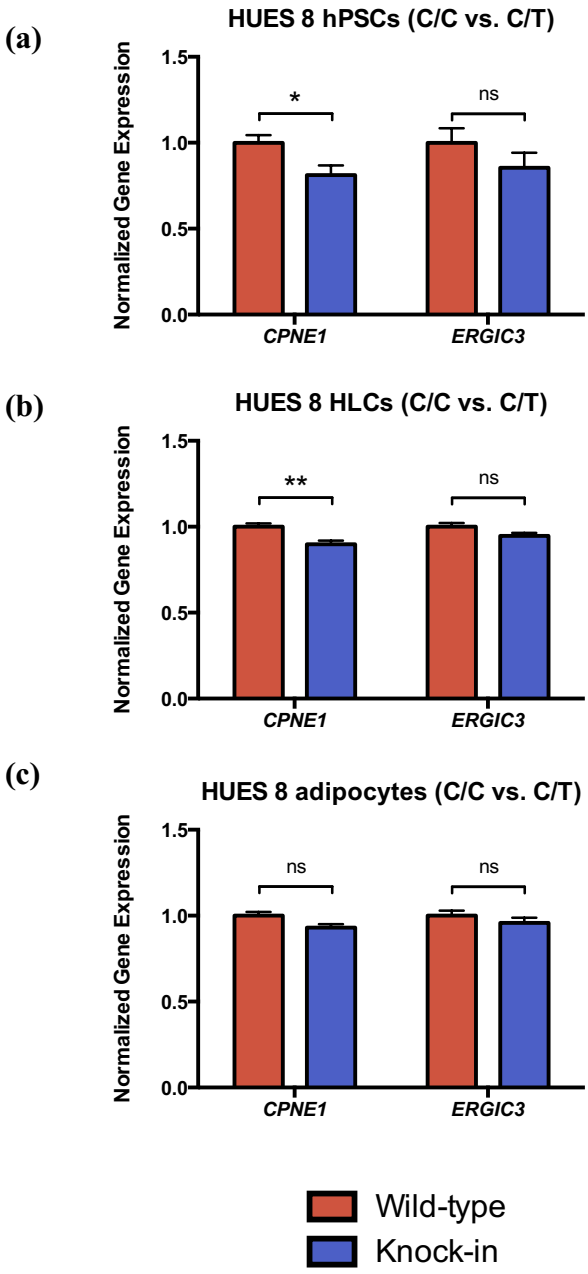


Figure 3. Gene expression analysis in rs2277862 knock-in hPSCs, differentiated HLCs, and white adipocytes. A heterozygous rs2277862 minor allele knock-in clone was generated on the HUES 8 background by homology-directed repair (HDR). (a) Gene expression in undifferentiated HUES 8 cells (n=2 wild-type clones and 1 knock-in clone; 6 wells per clone) (b) Gene expression in differentiated HUES 8 HLCs (n=2 wild-type clones and 1 knock-in clone; 6 wells per clone) (c) Gene expression in differentiated HUES 8 white adipocytes (n=2 wild-type clones and 1 knock-in clone; 6 wells per clone). Values are normalized to mean expression levels in wild-type clones. Statistical analysis by Mann-Whitney U test. Data represented as mean ± SEM (ns non-significant, *P<0.05, **P<0.01, ***P<0.001). *Experiments in Figures 3a and 3b performed by Xiao Wang; experiment in Figure 3c performed by Avanthi Raghavan.*

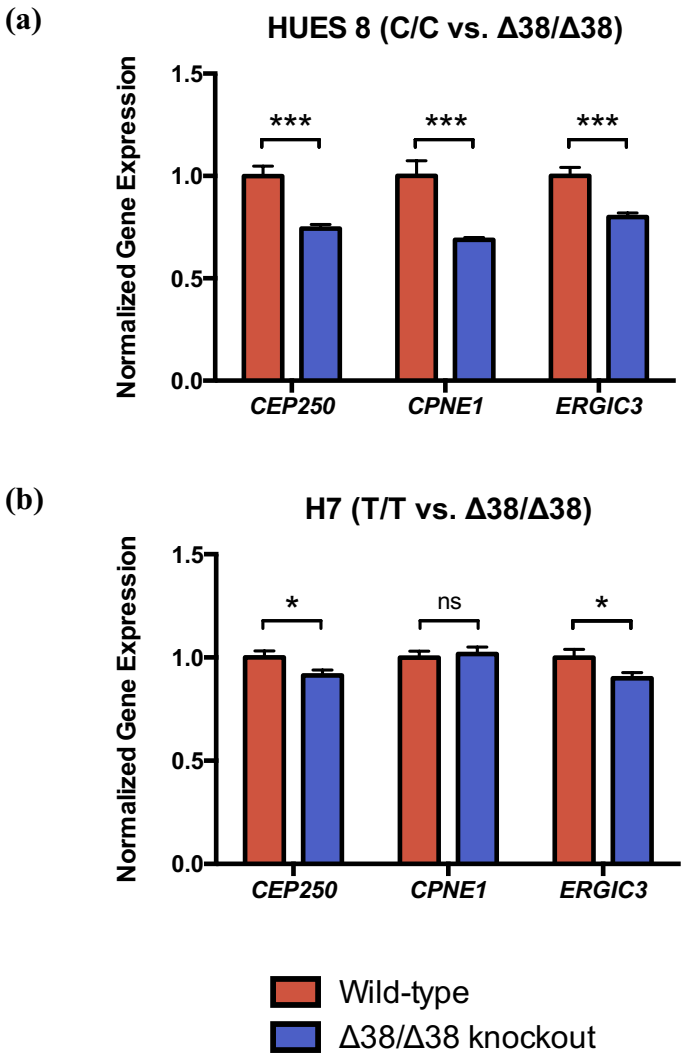


Figure 4. Gene expression analysis in rs2277862 knockout hPSCs. Homozygous 38-bp deletion mutants for rs2277862 were generated on the HUES 8 (homozygous major at rs2277862) and H7 (homozygous minor at rs2277862) backgrounds. Expression of the 20q11 genes *CEP250*, *CPNE1*, and *ERGIC3* in undifferentiated hPSCs was analyzed by qPCR. (a) Gene expression in undifferentiated HUES 8 cells (n=10 wild-type and 10 knockout clones, 3 wells per clone) (b) Gene expression in undifferentiated H7 cells (n=8 wild-type and 6 knockout clones, 3 wells per clone). Values are normalized to mean expression levels in wild-type clones. Statistical analysis by Mann-Whitney U test. Data represented as mean ± SEM (ns non-significant, *P<0.05, **P<0.01, ***P<0.001). *Experiments performed by Avanthi Raghavan.*

(a)

(b)

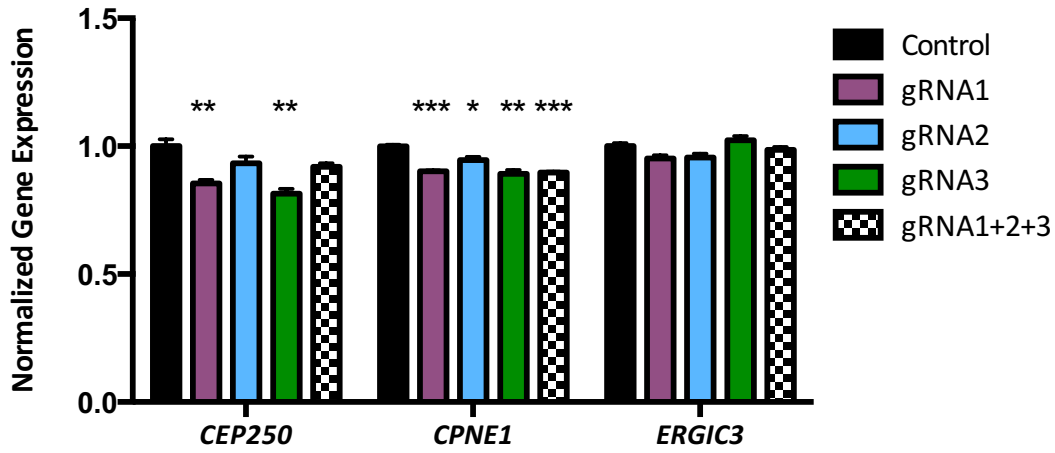


Figure 5. CRISPR interference enables modulation of gene expression from the rs2277862 locus. (a) gRNAs for the rs2277862 locus. Guide RNA protospacers are underlined and PAM is bolded. (b) Gene expression in HEK 293T cells (homozygous major at rs2277862) transfected with dCas9 and various gRNAs targeting the rs2277862 locus, either singly or in combination (n=3 wells per group). Control cells received the dCas9 construct without an accompanying gRNA. Values are normalized to mean expression levels in control cells. Statistical analysis by unpaired Student t-test. Data represented as mean ± SEM (*P<0.05, **P<0.01, ***P<0.001). *Experiment performed by Avanthi Raghavan.*

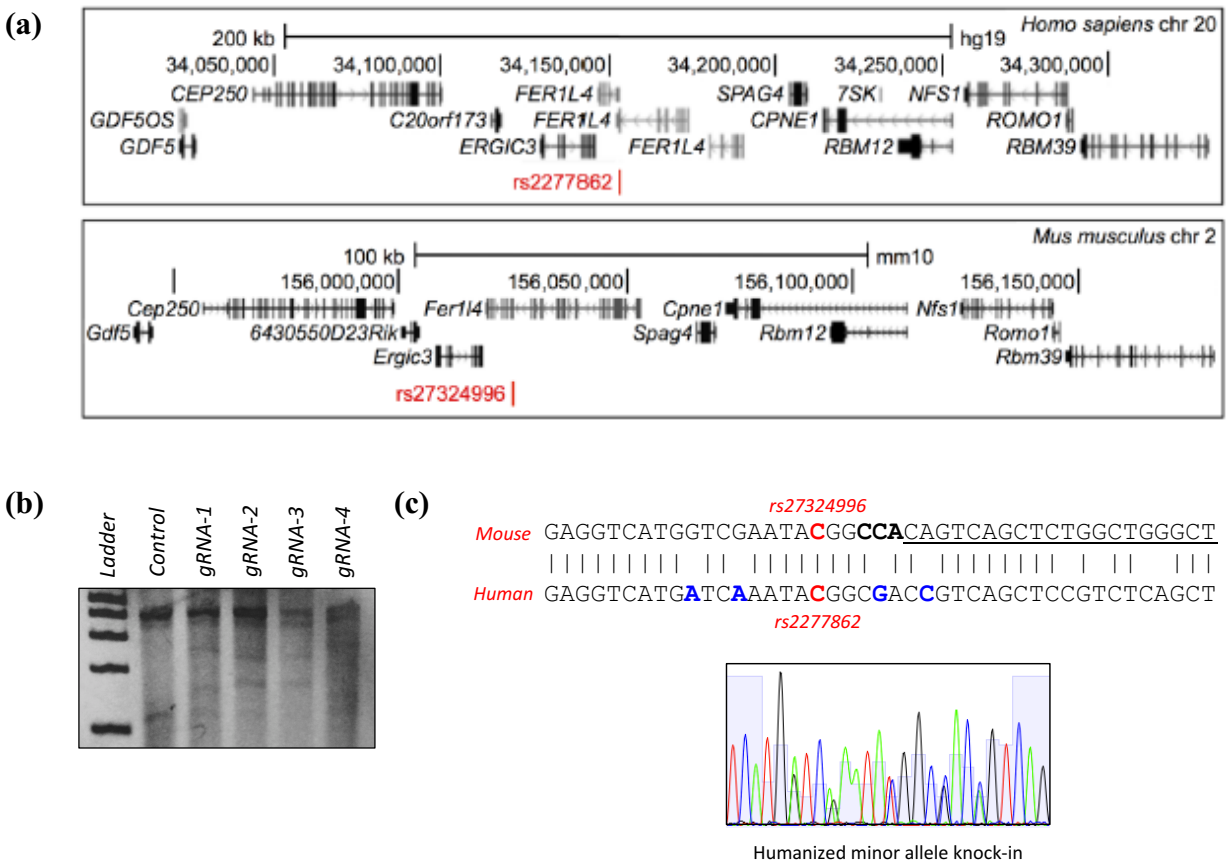


Figure 6. Generation of locus-humanized mice for rs2277862/rs27324996. (a) Schematics of the human rs2277862 locus (top) and the orthologous locus in mouse (bottom). The genetic architecture is well conserved in mouse, including allelic variants of the SNP itself (the murine equivalent is rs27324996). (b) CEL I assay to assess cleavage activity of candidate gRNAs in mouse 3T3-L1 cells. (c) Schematic of targeting strategy in one-cell mouse embryos and Sanger sequencing traces from the positive founder mouse, into which the minor allele of rs2277862/rs27324996 as well as four additional non-conserved nucleotides were knocked into one chromosome to humanize the site. Guide RNA protospacers are underlined and PAM is bolded. *Mouse model generated by Avanthi Raghavan.*

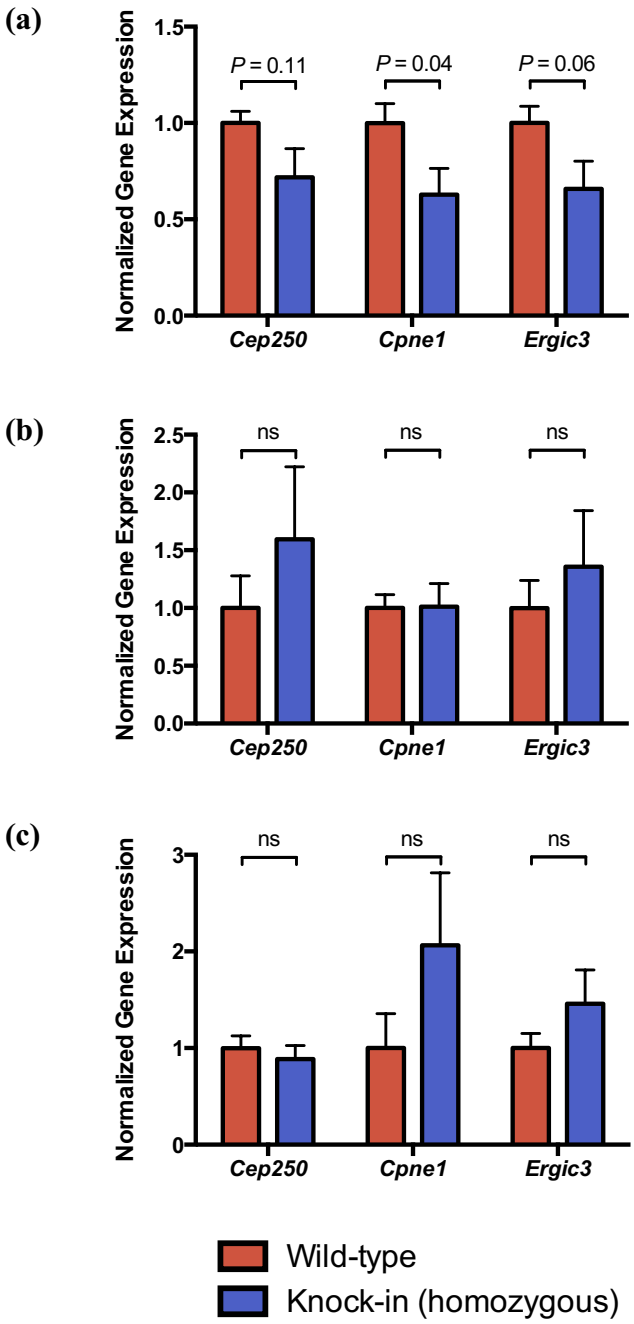


Figure 7. Gene expression analysis in rs2277862/rs27324996 locus-humanized mice. Gene expression was compared in wild-type (n = 18) and homozygous minor allele knock-in (n = 10) littermate mice in (a) liver, (b) omental fat, and (c) subcutaneous fat. Values are normalized to mean expression levels in wild-type clones. Statistical analysis by Mann-Whitney U test. Data represented as mean ± SEM (ns non-significant, *P<0.05, **P<0.01, ***P<0.001). *Mice were bred, genotyped and sacrificed by Avanthi Raghavan; experiment was performed by Xiao Wang.*

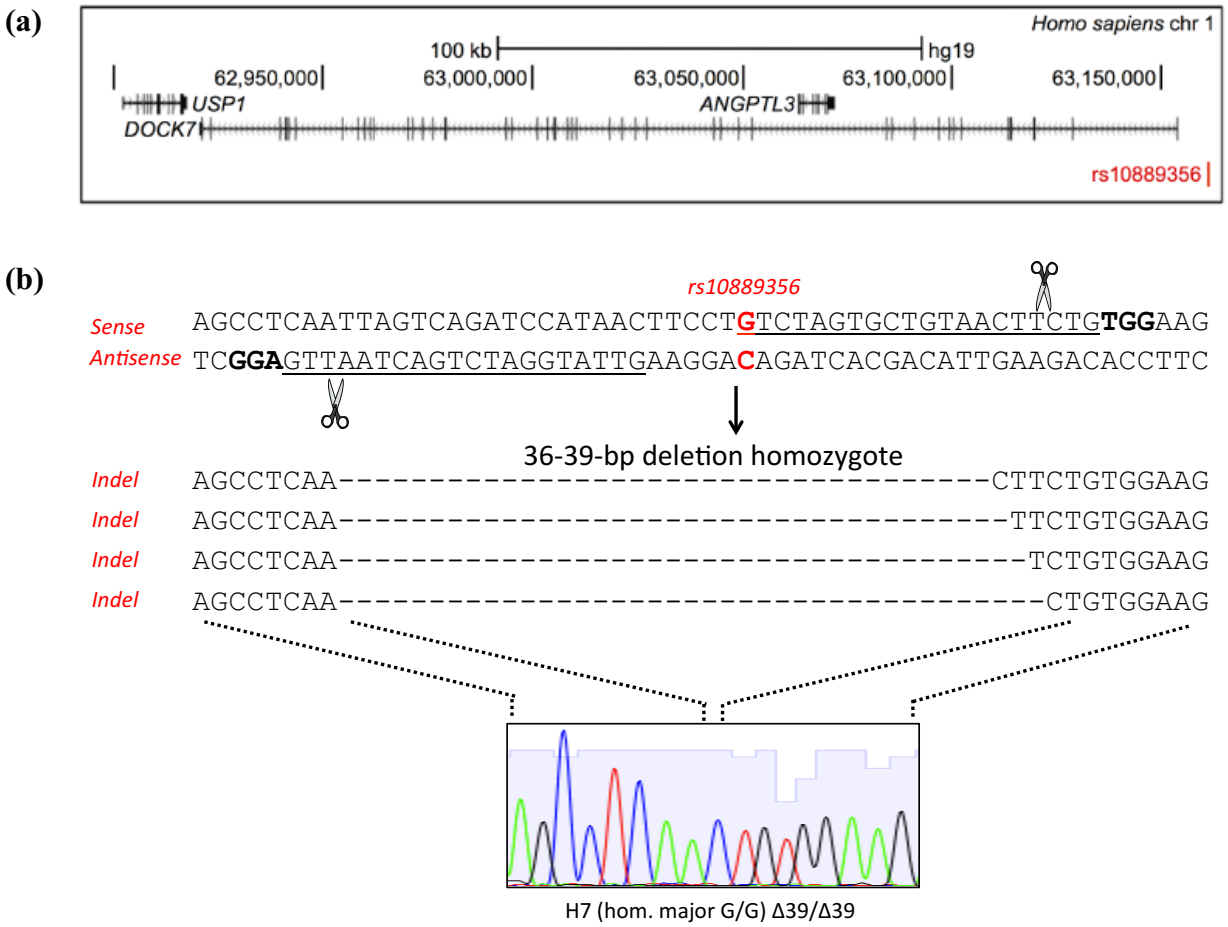


Figure 8. CRISPR/Cas genome editing at the rs10889356 locus in hPSCs. (a) Schematic of the human 1p31 locus (b) Homozygous deletions encompassing rs10889356 were generated on the H7 background (homozygous major at rs10889356) with a dual gRNA approach. Representative indels shown. Guide RNA protospacers are underlined and PAM is bolded. *Cell line generated by Avanthi Raghavan.*

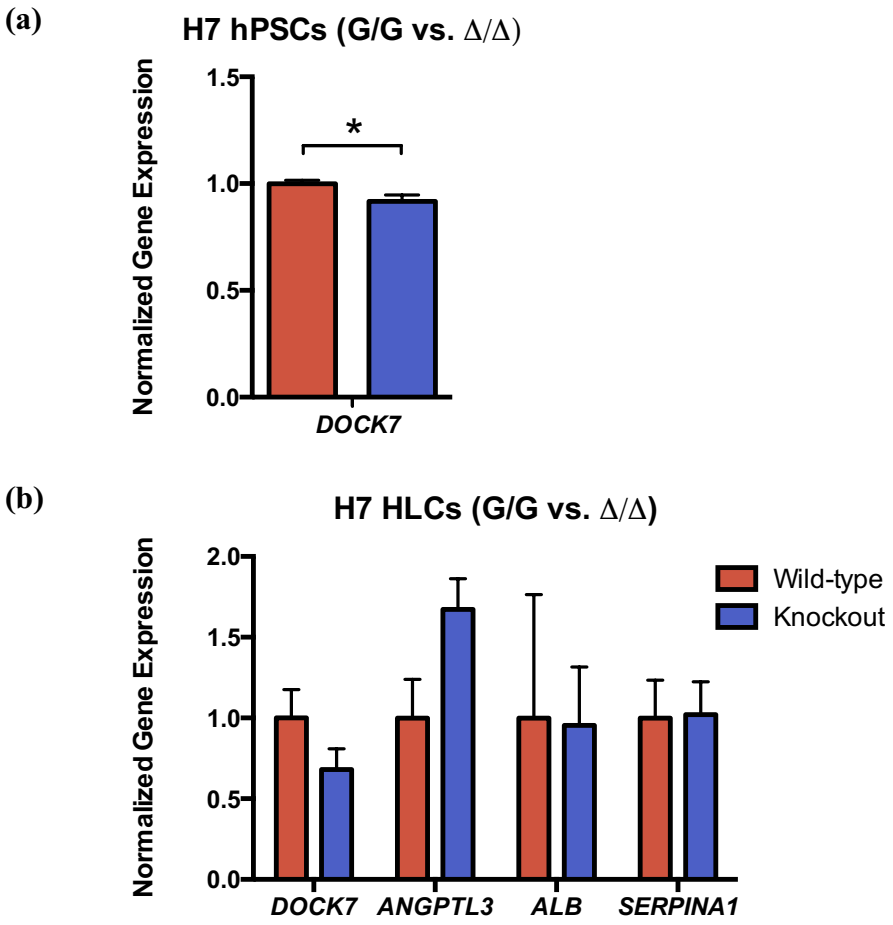
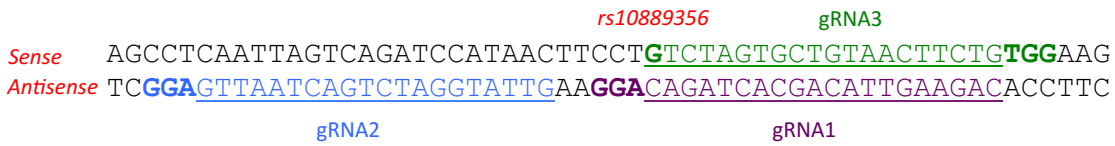


Figure 9. Gene expression analysis in rs10889356 knockout hPSCs and differentiated HLCs. Homozygous deletion mutants for rs10889356 were generated on the H7 (homozygous major) background. (a) Gene expression in undifferentiated H7 cells (n=12 wild-type and 8 knockout clones, 3 wells per clone). (b) Gene expression in differentiated H7 HLCs (n=4 wild-type and 4 knockout clones, 1 well per clone). Values are normalized to mean expression levels in wild-type clones. Statistical analysis by Mann-Whitney U test. Data represented as mean ± SEM (ns non-significant, *P<0.05, **P<0.01, ***P<0.001). *Experiments performed by Avanthi Raghavan.*

(a)



(b)

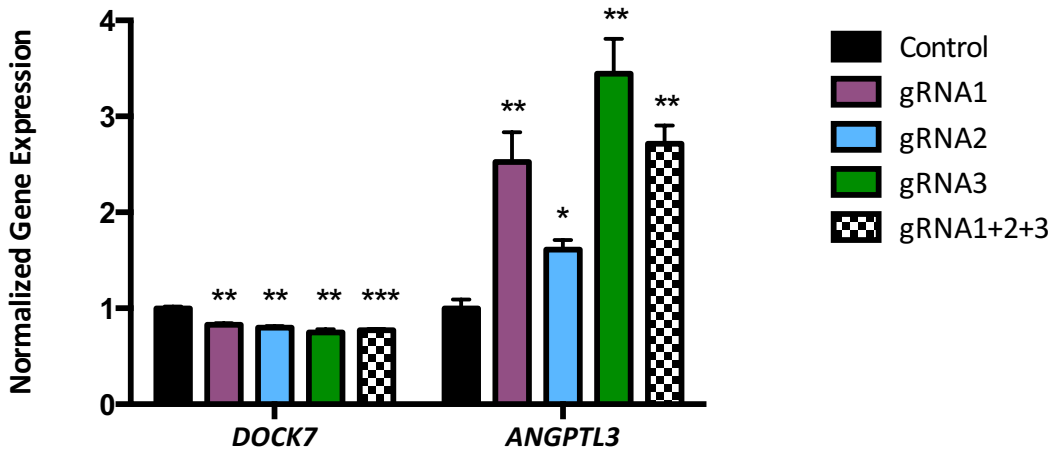


Figure 10. CRISPR interference enables modulation of gene expression from the rs10889356 locus. (a) gRNAs for the rs10889356 locus. Guide RNA protospacers are underlined and PAM is bolded. (b) Gene expression in HepG2 hepatoma cells (homozygous major at rs10889356) transfected with dCas9 and various gRNAs targeting the rs10889356 locus, either singly or in combination (n=3 wells per group). Control cells received the dCas9 construct without an accompanying gRNA. Values are normalized to mean expression levels in control cells. Statistical analysis by unpaired Student t-test. Data represented as mean ± SEM (*P<0.05, **P<0.01, ***P<0.001). *Experiment performed by Avanthi Raghavan.*

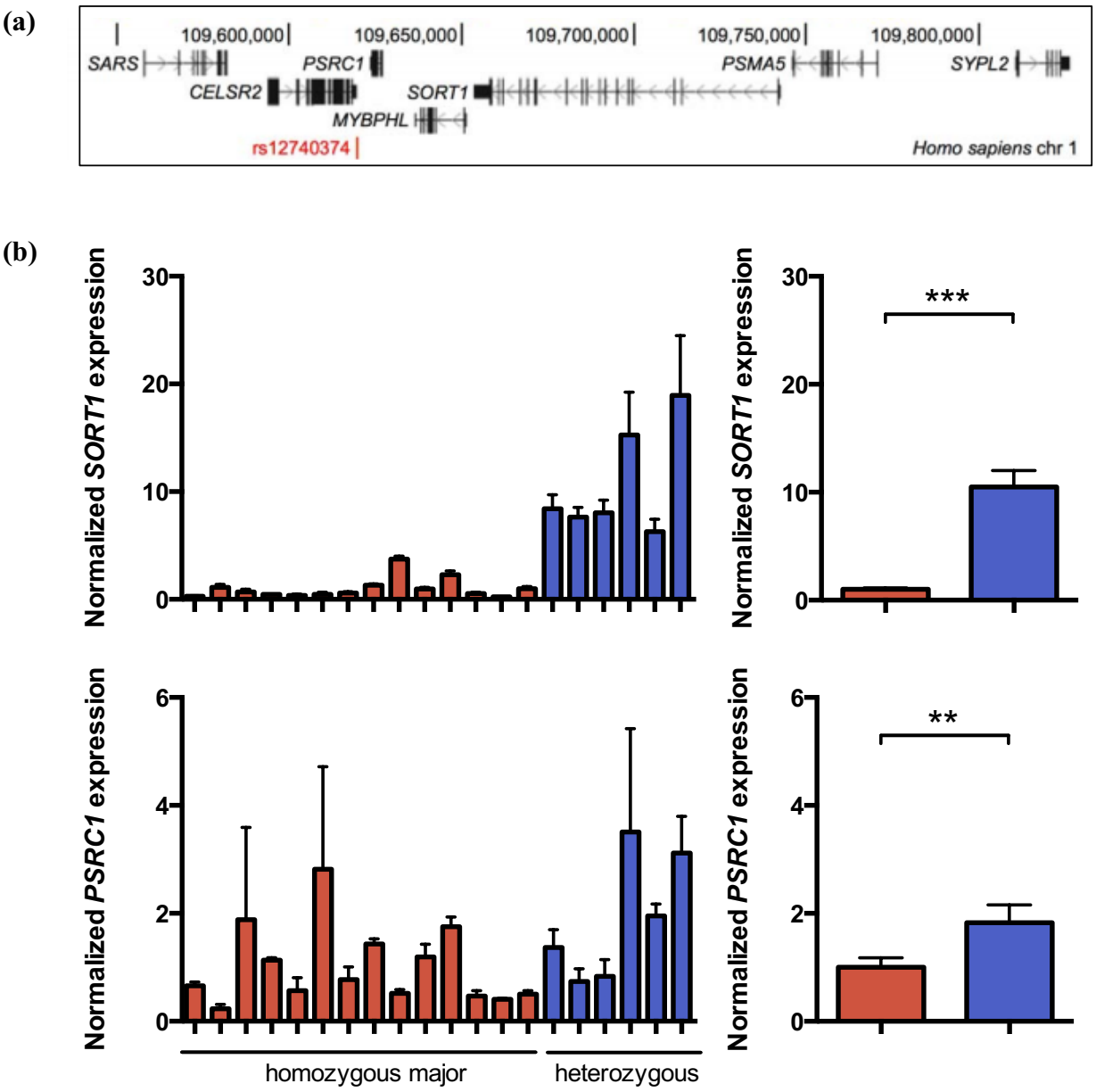


Figure 11. Gene expression analysis in primary human hepatocytes with varying genotypes at rs12740374. (a) Schematic of human 1p13 locus (b) *SORT1* and *PSRC1* expression were assessed in primary human hepatocytes that were homozygous major (n=14 independent lots, 3 wells per line) or heterozygous (n=6 independent lots, 3 wells per line) for rs12740374. Gene expression data is shown for individual lots (left) as well as aggregated by genotype (right). Values are normalized to mean expression level in all homozygous major lines. Statistical analysis by Mann-Whitney U test. Data represented as mean ± SEM (*P<0.05, **P<0.01, ***P<0.001). *Experiment performed by Avanthi Raghavan.*

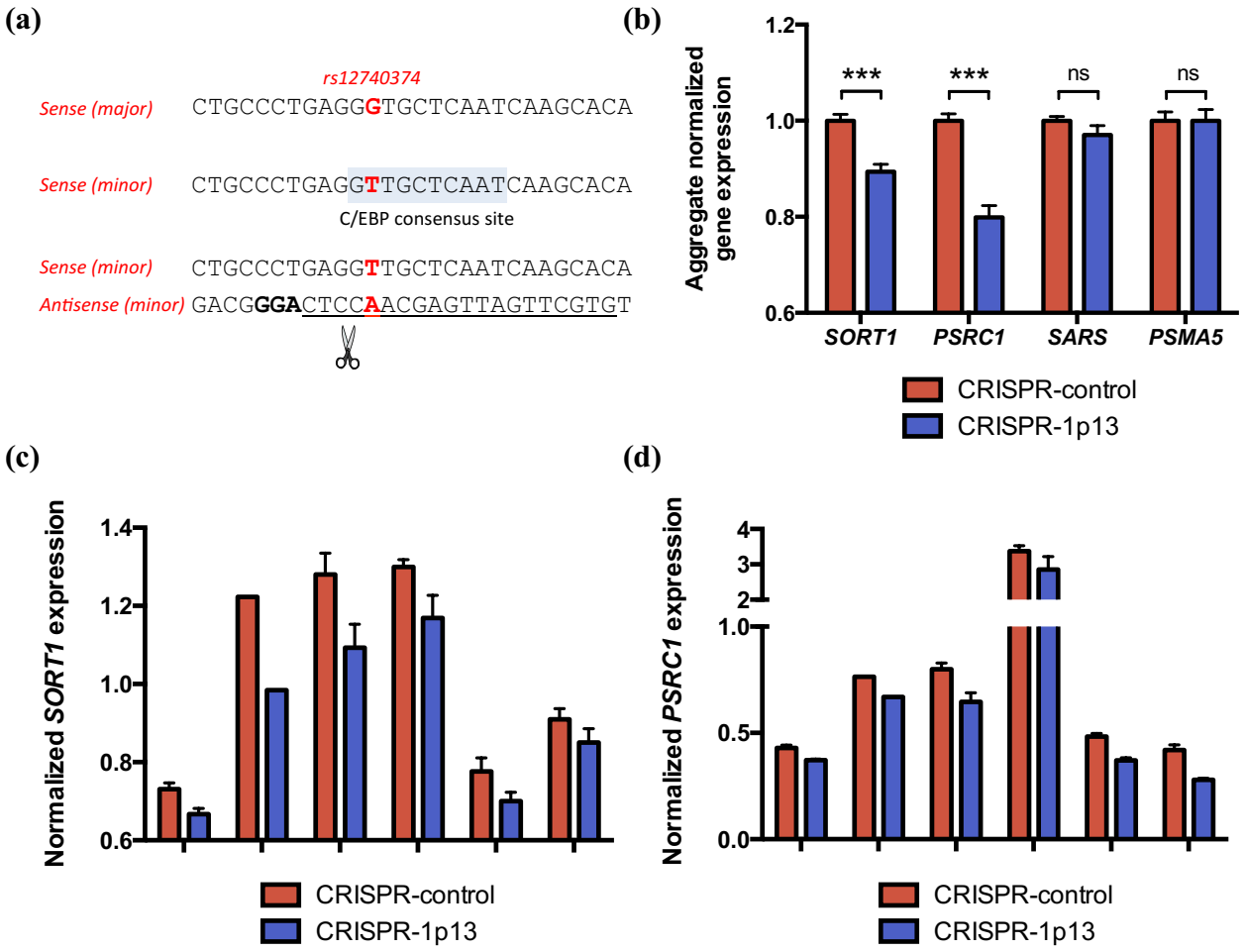


Figure 12. Genome editing at rs12740374 in primary human hepatocytes. (a) Genomic sequence at rs12740374, with minor allele-specific CRISPR gene editing strategy shown. Guide RNA protospacer is underlined and PAM is bolded. (b) Aggregate gene expression in rs12740374 heterozygous primary human hepatocytes treated with CRISPR-control or CRISPR-1p13 adenovirus (n=19 paired sets with paired wells derived from six heterozygous lots; variable numbers of paired wells were plated from each lot depending on the availability of cells in the original vial). (c) *SORT1* expression in each of six lots of heterozygous primary human hepatocytes treated with CRISPR-control or CRISPR-1p13 adenovirus (d) *PSRC1* expression in each of six lots of heterozygous primary human hepatocytes treated with CRISPR-control or CRISPR-1p13 adenovirus. Values are normalized to mean expression in CRISPR-control-treated cell lines. Note y-axis ranges for all graphs. Statistical analysis by Wilcoxon signed-rank test. Data represented as mean ± SEM (ns non-significant, *P<0.05, **P<0.01, ***P<0.001). *Experiment performed by Avanthi Raghavan.*

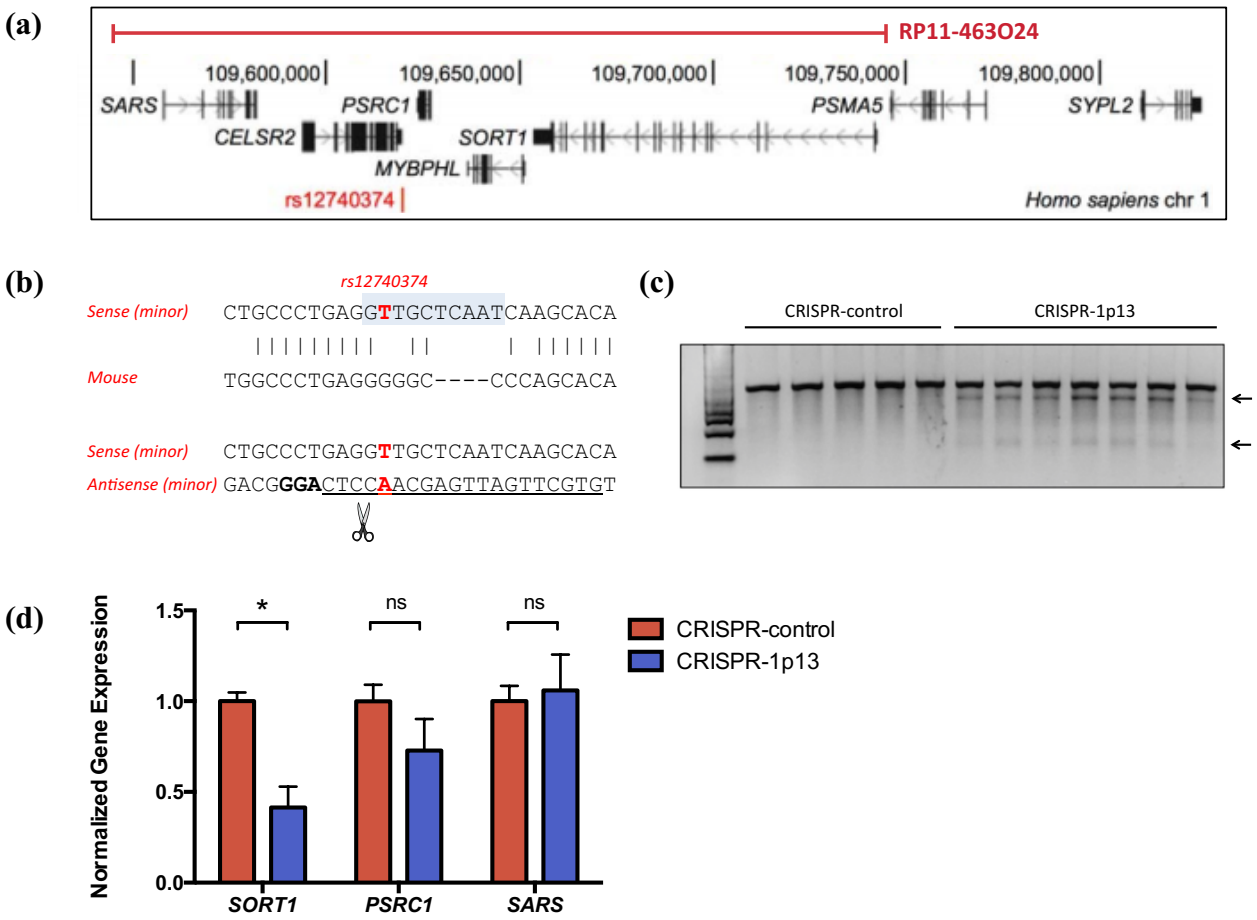


Figure 13. *In vivo* somatic genome editing at rs12740374 in locus-humanized mice. (a) Schematic of 1p13 locus (including BAC RP11-463O24). (b) Genomic sequence at rs12740374, with orthologous murine sequence and minor allele-specific CRISPR gene editing strategy shown. Guide RNA protospacer is underlined and PAM is bolded. (c) CEL I nuclease assays to assess for cleavage activity near the rs12740374 site in locus-humanized BAC transgenic mice administered CRISPR-control or CRISPR-1p13 adenovirus. Arrows denote cleavage products. (d) Gene expression levels of the human 1p13 genes were compared in CRISPR-control (n=5) and CRISPR-1p13 (n=7) mice, using mouse *Actb* as the reference gene. Values are normalized to mean expression in CRISPR-control-treated mice. Statistical analysis by Mann-Whitney U test. Data represented as mean ± SEM (ns non-significant, *P<0.05, **P<0.01, ***P<0.001). *Initial creation and genotyping of BAC transgenic mice was performed by Derek Peters. Experiment performed by Avanthi Raghavan, with assistance from Xiao Wang.*