Isogenic Human Pluripotent Stem Cell Models of Cardiovascular Disease-Associated Genetic Variation

A dissertation presented

by

Derek Tilghman Peters

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

December 2015
Isogenic Human Pluripotent Stem Cell Models of Cardiovascular Disease-Associated Genetic Variation

Abstract

A complex interplay of genetic and environmental factors underlies the development of common human diseases such as cardiovascular disease. Despite widespread use of existing medications, coronary heart disease (CHD), including myocardial infarction (MI), remains a significant cause of morbidity and mortality worldwide. The results of recent human genetic studies have provided unprecedented opportunities to elucidate the genes and molecular pathways that underlie CHD and risk factors like plasma lipid concentrations. Translating these findings into mechanistic insight promises to improve our understanding of disease pathogenesis and aid in the development of novel therapies. Precise functional characterization is required to bridge this gap; yet doing so can be challenging using traditional approaches.

The properties of human pluripotent stem cells (hPSCs) make them uniquely suited to address this challenge - they retain a normal human genome in culture, can be genetically modified, and can be differentiated into multiple cell types. The work presented in this thesis demonstrates the use of hPSCs and genome editing technology to generate human disease models in vitro. We developed a genome editing system optimized for hPSCs that can be used to efficiently generate hPSC lines with targeted
genetic modifications. Modified and isogenic control hPSCs are then differentiated into a relevant cell type for phenotypic characterization. We applied this approach to investigate the functional effects of human genetic variation underlying plasma lipid concentrations.

We used TALEN genome editing to clarify the role of the \textit{SORT1} gene as a mediator of cellular metabolic processes. We targeted the \textit{ANGPTL3} gene using TALEN and CRISPR/Cas9 genome editing to create an \textit{in vitro} model of a monogenic disorder, familial combined hypolipidemia, and investigated the putative role of \textit{ANGPTL3} in the regulation of low-density lipoprotein cholesterol metabolism. To facilitate the use of hPSC-models to study subtle phenotypes involving human liver, we developed a method for purifying hepatocyte-like cells following hPSC differentiation. Finally, we combined these approaches to model the tissue-specific effects of a common non-coding genetic variant in the chromosome 1p13 locus that is associated with risk of MI. This result demonstrates the feasibility of using hPSCs to characterize disease-associated common genetic variation.
# Table of contents

Title page ................................................................................................................. i
Copyright notice ....................................................................................................... ii
Abstract .................................................................................................................... iii
Table of contents ...................................................................................................... v
Acknowledgements ................................................................................................. xi
Statement of contribution ....................................................................................... xiii

**CHAPTER 1: FUNCTIONAL EVALUATION OF GENETIC VARIATION UNDERLYING COMPLEX HUMAN DISEASES** ........................................................................... 1

SUMMARY .................................................................................................................. 2

INTRODUCTION .......................................................................................................... 3

  Cardiovascular Disease: Progress, Challenges, and Opportunities ...................... 3
  Opportunities to Connect Genotype to Phenotype .............................................. 4

CHARACTERIZING COMMON DNA VARIANTS .................................................. 6

  Loci Associated with Blood Lipids ........................................................................ 6

  Variation within a Gene Desert—the 9p21 Locus ............................................. 8

CHARACTERIZING NOVEL CODING VARIANTS ............................................... 11

MASSIVELY PARALLEL ENHANCER CHARACTERIZATION ............................ 14

CELLULAR MODELS OF HUMAN GENETIC VARIATION ................................. 16

  Human Pluripotent Stem Cell Disease Modeling ............................................. 16

    Fulfilling “Genetics Version of Koch’s Postulates” Using Human Pluripotent Stem Cells... 19

CONCLUSIONS ........................................................................................................... 21

ACKNOWLEDGEMENTS .......................................................................................... 22

REFERENCES ........................................................................................................... 23
CHAPTER 2: A SYSTEM FOR EFFICIENT GENOME EDITING IN HUMAN PLURIPOTENT STEM CELLS USING TALENS AND CRISPR/CAS9

SUMMARY ................................................................................................................................. 28

2.1. INTRODUCTION ................................................................................................................ 29

2.2. PROTOCOL OVERVIEW .................................................................................................... 32

2.3. GENERATING CUSTOM TALENS ...................................................................................... 35
   Introduction to TALENs ........................................................................................................... 35
   2.3.1 Protocol .......................................................................................................................... 36
   2.3.2 Materials ........................................................................................................................ 45
   2.3.3 Troubleshooting ............................................................................................................ 45

2.4. GENERATING CUSTOM CRISPRS .................................................................................... 48
   Introduction to the CRISPR/Cas system ............................................................................... 48
   2.4.1 Protocol .......................................................................................................................... 49

2.5. GENOME EDITING IN HUMAN PLURIPOTENT STEM CELLS ......................................... 53
   2.5.1 Protocol .......................................................................................................................... 53
   2.5.2 Materials ........................................................................................................................ 65
   2.5.3 Troubleshooting ............................................................................................................ 66

2.6. DISCUSSION ....................................................................................................................... 70

ACKNOWLEDGEMENTS ............................................................................................................. 71

REFERENCES ............................................................................................................................. 72

CHAPTER 3: ISOGENIC HESC-BASED DISEASE MODELS CLARIFY THE ROLE OF SORT1 AS A MEDIATOR OF DISEASE-RELATED CELLULAR FUNCTIONS IN HUMAN METABOLIC CELL TYPES ................................................................................................................................. 75

SUMMARY ................................................................................................................................. 76

INTRODUCTION .......................................................................................................................... 77
CHAPTER 4: IN VITRO MODEL OF ANGPTL3 DEFICIENCY IN FAMILIAL COMBINED HYPOLIPIDEMIA USING DIFFERENTIATED HUMAN HEPATOCYTE-LIKE CELLS ........ 128

SUMMARY ................................................................................................................................. 129

INTRODUCTION .......................................................................................................................... 130

RESULTS ...................................................................................................................................... 136

TALEN genome editing to generate ANGPTL3 mutant hPSCs .................................................. 136

Decreased ANGPTL3 level in mutant hepatocyte-like cells ....................................................... 138

CRISPR genome editing to generate multiple biallelic ANGPTL3-mutant hPSC lines .......... 141

Differentiation to hepatocyte-like cells ...................................................................................... 145

ANGPTL3 deficient HLCs exhibit altered apolipoprotein secretion ........................................ 148

DISCUSSION ............................................................................................................................... 152

METHODS .................................................................................................................................. 156

Cell culture ................................................................................................................................ 156

Genome editing and confirmation of mutant cell lines ............................................................. 156

Differentiation of hESCs into HLCs ........................................................................................... 156

Quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR) ............................ 156

Enzyme-linked immunosorbent assays (ELISAs) ..................................................................... 157

Statistical analysis ...................................................................................................................... 157

REFERENCES .............................................................................................................................. 158

CHAPTER 5: PURIFYING DIFFERENTIATED HEPATOCYTE-LIKE CELLS TO FACILITATE IN VITRO GENETIC STUDIES ........................................................................................................... 161

INTRODUCTION .......................................................................................................................... 162

RESULTS ...................................................................................................................................... 167

Directed differentiation of HLCs ................................................................................................ 167

Selecting a candidate for prospective isolation of hepatocyte-like cells ................................. 170
ASGR1 marks a subset of hepatocytes arising during HLC differentiation ................. 171
Enrichment of differentiated hepatocytes based on ASGR1 surface expression .......... 174
Global transcriptional profiling of ASGR1+ cells ................................................. 181
Replating ASGR1-enriched HLCs ............................................................................ 187
DISCUSSION ........................................................................................................... 193
METHODS ............................................................................................................... 196
Cell culture .............................................................................................................. 196
Differentiation of hESCs and iPSCs into HLCs ...................................................... 196
Immunocytochemistry .............................................................................................. 198
Intracellular flow cytometry ....................................................................................... 198
ASGR1 FACS of HLCs with intracellular albumin staining ..................................... 199
ASGR1 FACS of HLCs for RNA isolation ................................................................. 201
RNA isolation and qRT-PCR ..................................................................................... 203
Microarray gene expression profiling ....................................................................... 203
Analysis of published gene expression data ........................................................... 204
ASGR1 MACS and replating of HLCs ...................................................................... 205
Cellular assays of hepatocyte functions ................................................................. 207
Statistical analysis .................................................................................................... 208
REFERENCES ........................................................................................................... 209

CHAPTER 6: MODELING THE EFFECTS OF A COMMON NONCODING GENETIC VARIANT
WITHIN THE 1P13/SORT1 LOCUS USING HPSCS ..................................................... 212

SUMMARY ............................................................................................................... 213
INTRODUCTION ....................................................................................................... 214
Translating GWAS findings into mechanistic insight using hPSCs ......................... 214
Back to the future: hPSC disease modeling of the 1p13 locus ................................. 218
# RESULTS

Targeting rs12740374 in hPSCs using CRISPR/Cas9 genome editing ........................................ 223

The rs12740374 minor allele is required for the maintenance of elevated SORT1 gene expression in human hepatocyte-like cells ................................................................. 230

# DISCUSSION

235

# METHODS

237

Genome editing and confirmation of mutant cell lines ............................................................. 237

RNA isolation and qRT-PCR ........................................................................................................ 237

Analysis of published gene expression data ............................................................................. 237

Statistical analysis ..................................................................................................................... 238

# REFERENCES

239

# CHAPTER 7: IMPLICATIONS, DISCUSSION, AND FUTURE DIRECTIONS

242

INTRODUCTION .......................................................................................................................... 243

FUTURE DIRECTIONS: THE 1p13 LOCUS AND hPSC MODELS OF GENETIC VARIATION .......... 248

The 1p13 locus: regulation at a distance .................................................................................... 248

Isogenic controls for iPSC cohort experiments ......................................................................... 252

DISCUSSION ............................................................................................................................... 255

Genome editing technology ...................................................................................................... 255

Potential off-target effects of genome editing ......................................................................... 259

Controlling confounding factors in isogenic hPSC disease modeling ..................................... 261

CONCLUSIONS AND FUTURE PROSPECTS ......................................................................... 263

REFERENCES ............................................................................................................................ 266
Acknowledgements

I feel extremely fortunate to have had the opportunity to work with outstanding faculty members, advisors, and colleagues during my graduate studies.

I would like to thank Dr. Chad Cowan for allowing me to conduct my graduate research in his laboratory and for his tremendous support as my advisor. The Cowan laboratory was an exceptional environment in which I was able to pursue my interest in stem cell research and develop as a scientist.

I would also like to thank Dr. Kiran Musunuru, who served as an instrumental advisor and mentor to me throughout my graduate studies. I am grateful to have had the opportunity to work closely with Dr. Musunuru while conducting the research described in this thesis. As an aspiring physician-scientist, working with Dr. Musunuru was truly inspiring.

I would like to acknowledge all of the members of the Cowan and Musunuru laboratories who provided invaluable help and support as I conducted my graduate research. In particular I would like to thank Jennifer Shay, Dr. Qiurong Ding, Dr. Nicolas Kuperwasser, Fang Xia, and Nancy Poole.

I would also like to thank the members of my Dissertation Advisory Committee for their thoughtful advice and suggestions: Dr. David Altshuler, Dr. Kenneth Chien, Dr. Kevin Eggan, and Dr. Richard Lee. I appreciate the time that they committed to reviewing the progress of my research; it was inspiring to have the opportunity to work with leaders in their respective scientific fields.

I would also like to thank the members of my dissertation examination committee:
Dr. Sangeeta Bhatia, Dr. Gustavo Mostoslavsky, and Dr. Qiao Zhou.

I would like to thank the following faculty members who kindly shared advice and reagents in support of the research described in this thesis: Dr. Sekar Kathiresan, Dr. George Church, Dr. Feng Zhang, and Dr. J. Keith Joung.

I would like to acknowledge the following people who provided much appreciated advice, help, and support over the course of my graduate studies: Dr. Rick Mitchell, Dr. Loren Walensky, Amy Cohen, Dr. David Cardozo, Dr. Dennis Kim, Dr. Susan Dymecki, Kate Hodgins, Patricia Cunningham, Maria Bollinger, and Daniel Gonzalez.

I would like to thank my earlier scientific mentors, Dr. Hartmut Land and Dr. Patrick Hu for their support and for inspiring me to pursue my interest in scientific research.

Finally, I would like to thank my family for their incredible love and support! Thank you to my Mom, Dad, and sister Kimberly. Thank you to my wife Hannah and children Jonathan, and Isabelle.
Statement of Contribution

The research presented in this thesis was conducted in the laboratories of Dr. Chad Cowan and Dr. Kiran Musunuru. Except where noted below, I designed, executed, and analyzed all experiments. All work was done under the supervision of Dr. Cowan and Dr. Musunuru. With the exception of Chapter 3, I am the primary author of all text and figures within this thesis.

Sections of Chapter 1 were previously published in the journal Human Molecular Genetics and were written by Dr. Musunuru and myself. Chapter 2 describes optimized protocols for efficient genome editing in human pluripotent stem cells (hPSCs). I developed these protocols in collaboration with Dr. Qiurong Ding based in part on published methods referenced specifically within the text. I produced all text and figures. Sections of this chapter were previously published in StemBook, an online journal of peer-reviewed chapters related to stem cell biology.

Chapter 3 was previously published in the journal Cell Stem Cell. I am the second author of the publication. The co-first authors are Drs. Qiurong Ding, Youn-Kyoung Lee, and Esperance A.K. Schaefer. My primary contribution to the work was the development of efficient genome editing methods for hPSCs (as described in Chapter 2), which was a central aspect of the studies reported in the publication. I also contributed to editing the manuscript. Sections of the original publication (e.g. studies pertaining to the genes AKT2 and PLIN1) are not included in Chapter 3 of this thesis, as they are not directly related to the focus of the thesis.

Work presented in Chapter 5 has been submitted for publication. The studies
described in this chapter were designed and conducted in collaboration with Christopher A. Henderson, a colleague in the laboratory. Dr. Curtis R. Warren assisted in preparing the manuscript for submission. I conducted all FACS and flow cytometry experiments, analyzed microarray gene expression data, and produced all text and figures.
Chapter 1: Functional Evaluation of Genetic Variation Underlying Complex Human Diseases

Sections of this chapter were previously published in:

SUMMARY

Genome-wide association studies and, more recently, next-generation sequencing studies have accelerated the investigation of complex human traits by providing a wealth of association data linking genetic variants to diseases and other phenotypic traits. These data promise to transform our understanding of the molecular pathways underlying complex human traits, but only if functional evaluation of the novel genetic variants is undertaken. Here we review recent examples in which such functional evaluation has been attempted, with varying degrees of success, and highlight new technological advances that should greatly enhance our ability to identify and dissect causal genotype-phenotype relationships. We focus in particular on a versatile human pluripotent stem cell (hPSC)-based approach, which can be used to perform precise functional characterization of disease-associated genetic variation.
INTRODUCTION

Cardiovascular Disease: Progress, Challenges, and Opportunities

Cardiovascular disease (CVD) is the leading cause of death worldwide (1). Coronary heart disease (CHD), including myocardial infarction (MI), is responsible for significant morbidity and mortality as well as healthcare expenditure in the United States (2). Approximately one in two men and one in three women over the age of 40 will suffer a coronary event during their lifetimes (3). The prevalence of CHD reached epidemic proportions by the middle of the twentieth century; deaths due to heart disease peaked in 1968 at 374 out of 100,000 individuals (4).

Since this time, significant advances in the understanding of disease risk factors and pathogenesis have resulted in the development of improved treatments for patients with existing disease (secondary prevention) and successful primary prevention strategies targeting established risk factors. As a result of these achievements, the incidence of MI has declined in recent years (5). Unfortunately, this progress is in jeopardy due to the rising prevalence of obesity and type 2 diabetes (6). This highlights the importance of continued public health efforts and the development of new therapeutic strategies.

Established risk factors for CHD include hypertension, dyslipidemia, obesity, diabetes mellitus, cigarette smoking, and physical inactivity (7). Elevated plasma cholesterol, especially low-density lipoprotein cholesterol (LDL-C), is a causal risk factor for MI (8); moreover, aberrant lipid metabolism is a key component of the metabolic syndrome. Despite the success of LDL-C lowering medications, including statins, a substantial portion of residual MI-risk remains (clinical studies have generally
demonstrated an overall risk reduction of 30 – 40%). Human lipid levels exhibit significant heritability; as much as one-half of inter-individual variability in LDL-C concentration may be due to inherited genetic factors (9). The identification of these factors, along with experimental characterization, promises to improve our understanding of disease pathogenesis and facilitate the development of future strategies to predict, prevent, and treat CHD.

**Opportunities to Connect Genotype to Phenotype**

Genetic variation gives rise to the heritability of complex human traits, including predisposition to common diseases. Many disorders ranging from autoimmune conditions to neuropsychiatric and cardiovascular diseases exhibit substantial heritability, as do traits such as fasting glucose level and plasma lipid concentrations, which are proven causal risk factors for disease. Recent discoveries in the field of human genetics have provided unprecedented opportunities to elucidate the genes and molecular pathways that underlie complex traits. Genome-wide association studies (GWASs) have identified numerous novel genetic loci associated with diseases and risk factors, while exome sequencing and candidate gene sequencing studies have uncovered a number of putative causal mutations in familiar as well as novel genes. Translating these findings into new mechanistic insight promises to improve our understanding of disease pathogenesis and aid in the development of novel therapies (Figure 1-1). However, precise characterization is required to bridge the gap from genotype to phenotype.
Figure 1-1. Connecting genotype to phenotype
Translating disease-related human genetic findings into new insight entails “bridging a gap” between genotype and phenotype, where the “phenotype” is predisposition to a disease or disease risk factor. As depicted in the diagram above, this entails elucidating details of the causal mechanisms connecting disease-associated genetic variation (“Locus/DNA variants”) to disease pathogenesis (“Disease/Phenotype”).
Functionally evaluating human genetic variation that underlies complex traits requires the application of an array of experimental approaches. As this area of investigation continues to mature, pioneering functional studies have utilized traditional *in vitro* and *in vivo* techniques and model systems as well as recently developed sequencing, epigenetic, and pluripotent stem cell technologies. Despite the wealth of genetic association data that have been generated in the past few years by increasing large GWASs as well as next-generation sequencing studies, to date there are relatively few examples of functional evaluation that has sought to clarify the connection between genotype and phenotype, and these examples demonstrate the advantages as well as pitfalls of the specific approaches that can be employed.

**CHARACTERIZING COMMON DNA VARIANTS**

GWASs identify loci in the human genome that contain common single nucleotide polymorphisms (SNPs) that are associated with a complex trait of interest. Functional evaluation of GWAS loci begins with the search either for a causal gene within the implicated locus or for a causal DNA variant, which in turn may lead to the identification of a causal gene. We present two illustrative examples here.

**Loci Associated with Blood Lipids**

GWASs of blood lipid concentrations, complex traits that are risk factors for cardiovascular disease, have identified 95 associated genomic loci, including loci containing known lipid genes, loci with established targets of lipid-lowering medications,
and novel loci with no known connection to lipids (10). The last group of loci likely harbor genes that play previously unappreciated roles in the regulation of lipoprotein metabolism and have been the focus of a variety of functional studies.

One such locus, located on chromosome 1p13, contains genetic variants that are highly associated with plasma low-density lipoprotein cholesterol (LDL-C) concentration as well as myocardial infarction (MI) risk, but at the time of discovery were not in proximity to any obvious candidate causal genes. Genetic mapping narrowed the region of LDL-C-associated genetic variation within the locus to a haplotype in an intergenic interval between the \textit{CELSR2} and \textit{PSRC1} genes; the haplotype was found to have a liver-specific association with the expression levels of \textit{CELSR2}, \textit{PSRC1}, and a nearby gene, \textit{SORT1} (11). This observation suggested that the haplotype might affect an enhancer. Extensive \textit{in vitro} characterization in cultured hepatoma cells using luciferase reporter constructs showed that a single SNP, located within a predicted C/EBP transcription-factor-binding site (TFBS), is responsible for the enhancer activity of the haplotype and is the likely causal variant. Additional \textit{in vitro} experiments supported a genetic mechanism by which the minor allele of the causal SNP creates a C/EBP TFBS (or, conversely, the major allele disrupts the TFBS), thereby increasing \textit{CELSR2}/\textit{PSRC1}/\textit{SORT1} expression in hepatocytes. This mechanism is further studied in Chapter 6 of this thesis.

Once a putative causal gene is identified for a GWAS locus, \textit{in vivo} overexpression and knockdown studies of the orthologous gene in mouse can be a useful approach to establishing that variation in the gene’s function likely underlies the

\footnote{Subsequent GWAS of blood lipid concentrations (41), published in 2013, identified a substantial number of additional associated genomic loci, raising the total number of lipid-associated loci to over 150.}
GWAS trait. Motivated by the association between the suspected causal SNP and 
SORT1 expression in human liver, experiments were undertaken in which Sort1 was 
overexpressed in mouse liver via adeno-associated virus or knocked down by siRNA. 
These perturbations resulted in significant and opposite effects on blood LDL-C levels in 
the mice, and together with in vitro studies in primary mouse hepatocytes, they strongly 
suggested that SORT1 is the causal gene within the locus (11). In contrast, studies with 
PSRC1 and CELSR2 in mice argued against their involvement in lipoprotein metabolism 
(11). In Chapter 3 of this thesis, the function of SORT1 is further characterized in 
several human cell types in vitro.

Overexpression and knockdown of the mouse orthologs of suspected causal 
genes at other lipid-associated loci were used to show that GALNT2, PPP1R3B, and 
TTC39B are likely to be regulators of high-density lipoprotein cholesterol (HDL-C) levels 
in humans (10). In the case of GALNT2, this gene was the only one within the locus in 
question, making it the obvious candidate for functional testing. In contrast, PPP1R3B 
and TTC39B were prioritized for functional characterization above several other genes 
in their respective loci because their expression levels in human liver were found to be 
associated with the genotypes of the GWAS SNPs, whereas the expression levels of 
the other genes were not.

Variation within a Gene Desert—the 9p21 Locus

Characterization of the LDL-C-associated 1p13 locus illustrates how in vitro 
reporter experiments and in vivo mouse experiments can be used to identify a causal 
non-coding genetic variant and provide the foundation for the elucidation of a pathway
connecting genotype to phenotype. In other instances where disease-associated variants fall within large non-coding regions of the genome, the underlying biology has proven more difficult to dissect.

Chromosome 9p21 contains common DNA variants that have been identified by various GWASs to be associated with MI risk, type 2 diabetes mellitus, melanoma, or breast cancer, making 9p21 one of the most tantalizing loci identified by GWASs. MI-associated SNPs within the 9p21 locus are not associated with traditional risk factors for MI such as lipids, blood pressure, and diabetes (12-14), but they are associated with abdominal aortic aneurysm and intracranial aneurysm (15), suggesting the existence of a novel pathway in vasculature that contributes to the pathophysiology of each of these diseases. These SNPs are located in a large (58 kb) gene-poor genomic locus, which includes part of a long non-coding RNA termed CDKN2BAS (also known as ANRIL), and lie more than 100 kb away from the closest protein-coding genes, CDKN2A and CDKN2B—cell cycle-related genes that encode cyclin-dependent kinase inhibitors—making the locus extremely challenging to characterize.

To investigate the role of this region, one group chose an approach based on a unique application of one of the most commonly used and powerful techniques in experimental biology: modification of the mouse genome. The investigators deleted a 70 kb non-coding interval on mouse chromosome 4 that is orthologous to the MI-associated region within the 9p21 locus, which resulted in decreased Cdkn2a and Cdkn2b expression in the mouse aorta and heart (16). Concordant with decreased expression of the cell cycle regulators Cdkn2a and Cdkn2b, knockout of the 70 kb interval resulted in increased proliferation and disrupted senescence of primary aortic
smooth muscle cells in culture. However, weakening the relevance of these findings to MI risk in humans is the fact that only about 50% sequence homology exists between the deleted interval and the orthologous region in humans. Moreover, knockout of this interval had no effect on atherosclerosis progression; this could reflect either that the causal DNA variant detected by GWASs in humans may not exist in or may not be relevant in mouse, or that human atherosclerotic disease is poorly phenocopied in mice—in either case, highlighting the potential limitations of studying human complex traits in non-human models.

Another group took a different approach to evaluate the 9p21 locus and its association with MI, performing an integrated analysis incorporating bioinformatics datasets and examining enhancer elements and long-range interactions involving the locus (17). First the authors analyzed DNase hypersensitivity and TFBSs as well as chromatin modification profiles—including marks associated with promoters of active genes, insulators, and enhancers—in multiple human cell types to discover potential regulatory elements in the 9p21 locus. Several of these elements displayed in vitro enhancer activity in luciferase reporter assays. After cataloging additional DNA variants by sequencing the region in 50 individuals, they narrowed the list of candidate regulatory variants by focusing on those that fell within enhancer elements, were consistently associated with increased MI risk, and disrupted consensus TFBSs. Two neighboring SNPs met these criteria, located within a predicted STAT1 binding site/enhancer preserved by the non-risk haplotype (and disrupted by the risk haplotype). Experiments in human vascular endothelial cells (HUVECs) and lymphoblastoid cell lines suggested a pathway through which interferon-gamma (IFN-gamma) signals via
STAT1 through the TFBS modified by the two SNPs, thereby resulting in the differential expression of CDKN2BAS and CDKN2B depending on which of the SNP alleles are present.

Extending their analysis of the STAT1 enhancer, the authors performed chromatin conformation capture (3C) in HUVECs combined with DNA selection and ligation (3C-DSL), showing that the enhancer participates in long-range interactions (up to a megabase) with the CDKN2A, CDKN2B, MTAP, and IFNA21 genes, especially upon IFN-gamma treatment of the cells, suggesting a dependence on STAT1 activity. This study demonstrates the power of an integrated approach for the identification of transcriptional regulatory elements in a complex non-coding region of the genome, and it provides insight into how non-coding genetic variants can alter gene regulatory pathways, potentially at great distances. However, although the study suggests that there is interplay between inflammation, variants in the 9p21 locus, and cell cycle regulation, it does not provide evidence that the implicated SNPs examined are the causal variants underlying the association of the 9p21 locus with MI, or that IFN-gamma signaling is the causal mechanistic link connecting the locus to the pathogenesis of MI. Indeed, no experimental approaches are undertaken to connect the STAT1 TFBS to atherosclerotic disease or some other phenotype that is directly relevant to MI.

CHARACTERIZING NOVEL CODING VARIANTS

Gene resequencing studies ask whether a gene of interest contributes to the pathogenesis of a complex human trait. Sequencing the coding region of a gene in a population may reveal a preponderance of variants present in a subset of individuals
who exhibit a specific trait, suggesting an association between the gene’s function and the observed trait. Whole-exome or whole-genome sequencing studies, which are unbiased, large-scale alternatives to the one-gene-at-a-time resequencing approach, aims to uncover novel causal mutations underlying a specific trait by performing next-generation DNA sequencing across the entire genome in a few affected individuals. Both approaches can provide strong evidence for an association between a specific gene and a trait; however, definitively establishing that a suspected gene contributes to the trait of interest requires functional validation. Moreover, in this context it would be valuable to assess the functional impact of each of the actual coding variants that are discovered—some might prove to be highly debilitating to gene function, whereas others may be of no functional consequence.

In one study, the coding region of the ANGPTL3 gene was resequenced in a population from the Dallas Heart Study to address whether ANGPTL3 modulates blood triglyceride (TG) levels in humans (18), in light of prior evidence that the mouse ortholog plays this role. While several nonsense, frameshift, and splice-site mutations were found in individuals within the lowest quartile of blood TG levels, suggesting that loss-of-function alleles are associated with low TG levels, a number of missense mutations of unclear functional consequence were also identified in individuals with low TG levels. In vitro experiments were performed in HEK293A cells to investigate the effects of each of the discovered mutations on ANGPTL3 expression, secretion, and activity. Notably, the results of these experiments were compared to predictions made by two computational annotation programs (PolyPhen and SIFT), revealing an approximately 35% false positive and 40% false negative rate for the prediction of deleterious mutations—
highlighting that, at the present time, computational methods cannot substitute for functional experiments in characterizing the effects of novel coding variants. Chapter 4 of this thesis describes the development of an hPSC-derived cellular model of ANGPTL3 deficiency.

Functional analysis of novel variants has been successfully conducted in model organisms, where the effects of coding variants on more complex phenotypes can be assessed with moderate throughput. One group followed up on resequencing studies with functional analysis in zebrafish as well as mammalian cells to gain insight into the genetic architecture of ciliopathy disorders (19). The investigators characterized mutant versions of the human TTC21B gene containing novel variants identified in ciliopathy patients or in healthy controls by assaying the genes’ ability to rescue a zebrafish ciliary phenotype (produced by knockdown of endogenous Ttc21b expression); they found that about one-quarter of the variants led to complete loss of function, about one-half resulted in partial loss of function, and the remainder were benign (i.e., non-functional). Of note, the investigators found that there was no significant difference in the burden of novel TTC21B variants in ciliopathy cases versus controls. When they restricted the analysis to novel variants shown by their experiments to affect protein function, they found a striking enrichment of these functional variants in the cases.

These two studies highlight the utility of experimental data in understanding the relationships between genotypes and phenotypes in humans. They also reinforce the point that an appropriate experimental system for functional analysis should be chosen based on the nature of the trait being studied as well as the function of the gene of interest, when known.
MASSIVELY PARALLEL ENHANCER CHARACTERIZATION

The regulation of gene expression is integral to most biological processes. As exemplified by the 1p13 locus, genetic variation within regulatory elements can have significant phenotypic consequences. Advanced DNA synthesis and sequencing technologies have made possible the development of creative solutions to address the challenge of dissecting genetic regulatory sequences, which can be used to interrogate the non-coding regions identified by complex-trait GWASs.

The “massively parallel reporter assay” (MPRA) approach is used to systematically interrogate each nucleotide position within a transcriptional regulatory element to identify functional TFBSs (20). Transcriptional activity using MPRA is assessed in cultured cells *in vitro* so it can be performed in an appropriate available cell type, accounting for cell type-specific differential regulation. Briefly, in MPRA, enhancer variants (containing one or more desired nucleotide substitutions) for an element of interest are generated and coupled to distinguishing tags (“barcodes”) using microarray-based DNA synthesis. Next the variants are cloned into a plasmid backbone, an invariant promoter-open reading frame segment is inserted, and the reporter plasmid pool is transfected into cells. Lastly, high-throughput RNA sequencing and counting of the distinguishing tags determines the relative quantities of RNA transcripts expressed from the plasmid pool and, by extension, the regulatory activities of the enhancer variants, defining the location of important regulatory positions.

“Massively parallel functional dissection” (MPFD) is a technology related to MPRA; however, it facilitates characterization of enhancer elements *in vivo* and involves the generation of a library of enhancer haplotypes with a programmable level of
degeneracy without requiring individual synthesis of each enhancer variant (21). These features make MPFD preferable for the efficient examination of longer regulatory elements in an in vivo context. Notably, initial experiments using MPFD have shown that not all functional motifs are associated with predicted TFBSs and that evolutionary constraint (i.e., phylogenetic conservation) poorly predicts the magnitude of functional impact for specific nucleotide positions—two points that should be considered when analyzing regulatory regions in the human genome by computational methods alone.

The MPRA and MPFD technologies could readily be applied to identify causal DNA variants within regions of non-coding DNA identified by GWASs, assuming that the causal DNA variants act by modulating the expression of nearby genes. Identification of a potential causal regulatory variant could then help to nominate candidate causal genes that could be functionally tested, as was the case with the 1p13 locus. Advances in genome editing technology (introduced in detail in Chapter 2 of this thesis)– in particular CRISPR/Cas genome editing – have enabled the development of additional experimental approaches similar to the MPRA/MPFD techniques. A particular method, termed “saturation editing”, entails multiplexed mutagenesis of a target genomic region in cells, enabling the functional assessment of DNA sequences in the endogenous chromosomal context (42).
CELLULAR MODELS OF HUMAN GENETIC VARIATION

Human Pluripotent Stem Cell Disease Modeling

Despite the insights that were gained using the approaches described above, each system ultimately falls short of establishing a direct causal connection linking genetic variation to its associated complex trait. These studies are limited by several factors that have traditionally made conducting appropriate functional studies challenging. First, complex human diseases and phenotypes may be poorly replicated in animal model systems, and regulatory variants may not be evolutionarily conserved. Second, human clinical samples are not amenable for experimental characterization due to the limited availability of most tissue types. Finally, immortalized human cell lines may in some cases be inappropriate for the evaluation of human genetic discoveries because the impact of natural genetic variation may be obscured by potential genomic instability, karyotypic abnormalities, and variability resulting from transformation and extensive passaging in cell culture.

Ongoing research provides clues as to the most appropriate way in which to study the genetic variation underlying complex traits. As observed with the 1p13 locus, genetic contributors to the development of complex traits may act in a tissue-specific manner. Indeed, a majority of common DNA variants that regulate gene expression appear to act in a cell type-dependent fashion (22), and disease-associated DNA variants are often located in predicted cell type-specific enhancer elements (23).

The challenges and insights described above suggest that human pluripotent stem cell (hPSC) technology could be a powerful tool for the evaluation of human genetic variation. hPSCs are defined by their ability to self-replicate and to differentiated
into all cells of the three primary germ layers. In 1998, Thomson et al first demonstrated the isolation and maintenance of hPSCs in culture (24). These cells, human embryonic stem cells (hESCs), are derived from the inner cell mass of donated blastocyst-stage embryos. Unlike transformed or immortalized human cell lines, hESCs can potentially be expanded indefinitely in culture while maintaining a normal karyotype. Soon after their initial isolation, it was shown that hPSCs could be directed to differentiate in vitro (25). These properties make hPSCs extremely unique in terms of their potential utility for research and as a source of cells for future regenerative medicine or cell-replacement therapies.

The promise of hPSCs was expanded further by the remarkable discovery and development of human induced pluripotent stem cells (iPSCs; 26-27). iPSCs possess the central properties of hESs, but are generated by direct reprogramming of somatic cells, rather than by isolation from embryos. Since the initial isolation of hPSCs, additional hESC lines have been derived (28) and methods for hPSC culture and iPSC reprogramming have progressed significantly (29-30). A large number of established hESC and hiPSC lines are now available for biomedical research. As of January 2014, more than 234 hPSC lines were listed on a National Institutes of Health (NIH) registry (29) and the number of available hiPSC lines is rapidly increasing.

A significant advantage of hPSCs relative to most primary human cells is that they can be genetically modified in vitro. Zinc finger nucleases (ZFNs) or transcription activator-like effector nucleases (TALENs) - chimeric proteins that contain customized DNA binding domains – can be used to introduce precise modifications into the genomic DNA of hPSCs (31-32). This targeted genome engineering (described in detail in
Chapter 2 of this thesis) is referred to as “genome editing†.” Furthermore, directed differentiation of hPSCs makes it possible to investigate cellular phenotypes in cell types which are difficult to obtain from humans (e.g., neurons) but are the most disease-relevant cell type. While hPSCs have been used to establish cellular models of monogenic disorders (33), little has been published regarding the use of hPSCs to study the genetic nature of complex traits. Nonetheless, one can envision the use of hPSCs to conduct genetic characterization of potential causal DNA variants, to interrogate putative causal genes, and to validate genotype-phenotype relationships in relevant human cell types.

† Genome editing can also be performed using the CRISPR/Cas9 system (described in Chapter 2 of this thesis). Unlike genome editing with ZFNs and TALENs, which involves engineered proteins, this versatile technology is based on an RNA-guided nuclease.
Fulfilling the “Genetics Version of Koch’s Postulates” Using Human Pluripotent Stem Cells

A number of studies have examined the impact of disease-associated genetic variation using human induced pluripotent stem cells (iPSCs) derived from affected patients and control individuals (34-39). However, this approach has two important limitations: (1) reprogramming may introduce variability that contributes to phenotypic differences when individual iPSC lines are differentiated, particularly when relatively subtle phenotypes are being examined; (2) the genetic backgrounds of disease-specific and control iPSC lines are not matched, making it difficult to attribute an observed phenotype to the influence of a specific genetic variant (Figure 1-2).

These limitations can be addressed using human iPSCs or human embryonic stem cells together with genome editing to fulfill the genetics equivalent of Koch’s postulates: a causal variant should cause the associated phenotype when introduced into a wild-type cell. One can start with a well-characterized hPSC line; perform genome editing to either introduce or “repair” a genetic variant, thereby creating an otherwise isogenic cell line that differs from the parental hPSC only with respect to the variant; differentiate the parental (control) and modified hPSC lines in parallel into the cell type of interest; and phenotype the differentiated cell lines (Figure 1-2). With this experimental design (“isogenic disease modeling”), any phenotypic differences seen between the cell lines can be directly attributed to the genetic variant. Efficiency and flexible methods for the genetic modification of hPSCs are essential to this approach. This has inspired the development of hPSC genome editing methods, which are the focus of Chapter 2 of this thesis.
Figure 1-2. Human pluripotent stem cell disease modeling

(Left) Human induced pluripotent stem cells (iPSCs) can be used to investigate the effects of disease-associated genetic variation in relevant cell types in vitro. However, if unmatched cell lines are used, observed phenotypic differences cannot necessarily be attributed to specific DNA variants due to many other genetic differences in present in the distinct genetic backgrounds. Moreover, subtle phenotypic effects of a specific DNA variant may be obscured by overall cell-line differences.

(Right) To avoid potential confounding effects resulting from unmatched genetic backgrounds, genome editing is used to generate hPSCs lines with targeted genetic modifications, which are otherwise genetically matched (i.e. presumed to be otherwise isogenic). Using different genome editing techniques, suspected causal mutations can be introduced into wild-type hPSCs or repaired in mutant iPSCs, depending on the experimental design.
CONCLUSIONS

It is clear that to truly establish causal connections between genetic variation and complex human traits, specific DNA variants must be directly interrogated in an appropriate biological context. While this type of precise functional genetic characterization is the definitive step in establishing genotype-phenotype relationships, evaluating novel genetic associations often requires several stages of investigation, where a variety of tools and approaches are used to uncover causal DNA variants, genes, and molecular mechanisms that contribute to the development of a complex trait. Creative applications of established techniques as well as a host of emerging technological developments promise to make such investigation sufficiently straightforward and efficient that it may well be feasible to cope with the wealth of data that is now emerging from GWASs and next-generation sequencing studies.
ACKNOWLEDGEMENTS

The authors thank Dr Sekar Kathiresan for useful comments on the manuscript (40). Conflict of Interest statement: none declared.
REFERENCES


Chapter 2: A System for Efficient Genome Editing in Human Pluripotent Stem Cells Using TALENS and CRISPR/Cas9

Sections of this chapter were previously published in:

SUMMARY

Genome editing is used to make targeted modifications to the genome of eukaryotic cells. There are many potential applications of genome editing in human pluripotent stem cells (hPSCs) including the generation of knockout and reporter cell lines. This protocol describes a system for efficient genome editing in hPSCs using engineered transcription activator-like effector nucleases (TALENs) or clustered regularly interspaced short palindromic repeat (CRISPR) technology.
2.1. INTRODUCTION

Gene targeting by homologous recombination has been used extensively to study gene function in model organisms – in particular in mouse embryonic stem (mES) cells for generating knockout and knock-in mice. In recognition of the importance of this technique, the Nobel Prize in Physiology or Medicine for 2007 was awarded to Mario R. Capecchi, Martin J. Evans and Oliver Smithies (Nobel Media, 2007). However, efforts to adapt this approach to other mammalian cell types, including human pluripotent stem cells (hPSCs), were hampered by prohibitively low HR efficiencies (Zwaka and Thomson, 2003) and technical challenges related to in vitro culture of hPSCs.

Although gene targeting in hPSCs was achieved, the overall rate of successful targeting was exceedingly low. This was due to (1) the low rate of HR (approximately 1 correct targeting event observed out of 1000 clones screened), (2) poor hPSC viability during single-cell cloning, and (3) inefficient delivery of DNA to hPSCs (Yates and Daley, 2006; Fenno et al., 2008). The latter two challenges were addressed by the development of improved hPSC culture and DNA delivery methods (Watanabe et al., 2007; Hohenstein et al., 2008; Schinzel et al., 2011). However, there remained a strong demand for efficient genome modification techniques. Against this background, genome editing emerged as a means to enable targeted genome engineering with increased efficiency (review by Urnov et al., 2010). Although initially demonstrated in model organisms (Bibikova et al., 2003), genome editing technology has proven to be particularly transformative when applied to hPSC research.

Genome editing is the process of generating genetically modified cells using engineered sequence-specific nucleases. Custom-engineered nucleases are introduced
into cells, in which they create targeted double-strand breaks (DSB) at a genomic site of interest for the purpose of creating insertion/deletion mutations (indels) or facilitating homologous recombination (Urnov et al., 2005). In the later approach, a short single-strand oligonucleotide (ssODN) or double-strand plasmid is used to introduce a desired DNA sequence into the genome. Transcription activator-like effector nucleases (TALENs) have been used successfully for genome editing in human pluripotent stem cells (hPSCs) (Hockemeyer et al., 2011), and custom TALEN pairs can be readily designed and constructed. Similarly, clustered regularly interspaced short palindromic repeat (CRISPR)/Cas systems have been demonstrated to achieve genome editing in hPSCs (Mali et al., 2013). A summary of the advantages and disadvantages of genome editing technologies is presented in Chapter 7 of this thesis.

There are many potential applications for genome editing in hPSCs including the following: (1) gene disruption – knocking out a gene of interest by inducing an indel mutation (Ding et al., 2013a and Chapter 3 of this thesis); (2) gene correction – correcting a causal mutation in a patient-specific induced pluripotent stem cell (iPSC) line or knocking in a disease-associated DNA variant (Soldner et al., 2011; Kiskinis et al., 2014); (3) inserting reporter genes into a genomic "safe-harbor" locus or other location to generate reporter cell lines (Hockemeyer et al., 2009; Liu et al., 2014).

Our laboratory has developed a system for rapid and efficient TALEN-mediated genome editing in hPSCs and used this system to create several cellular disease models (Ding et al., 2013a). We have also adapted this system for highly efficient CRISPR-mediated genome editing (Ding et al., 2013b). In this thesis chapter, we describe the hPSC genome editing system in detail, including the following elements:
(1) criteria for choosing a target site and designing DNA constructs for HDR; (2) optimized plasmids for the construction and expression of genome editing components in hPSCs; (3) methods for hPSC culturing, DNA delivery, and high-throughput screening of hPSC clones.

Our approach involves a cell-sorting step, which minimizes the impact of inefficient DNA delivery. Using this collection of methods, modified hPSC lines are efficiently generated without antibiotic selection or multiple hPSC cloning steps. Therefore, the entire process can be completed in less than one month. This is significantly faster than approaches that involve antibiotic selection and multiple hPSC cloning steps, which may take 16 or more weeks to complete (Soldner et al., 2011).

Overall, the genome editing system described in this chapter can be adapted for a variety of experimental applications and used to generate modified hPSC lines with very high efficiency (2% - 79%) compared with alternative methods (Mali et al., 2013).
2.2. PROTOCOL OVERVIEW

Figures 2-1 and 2-2 summarize the steps necessary to accomplish genome editing in hPSCs using our system, beginning with the design and construction of custom TALENs (Figure 2-1) or CRISPRs (Figure 2-2) and ending with the identification and expansion of successfully modified hPSC lines. TALENs are introduced in Section 2.3, which describes the process of designing and constructing custom TALENs using our publicly available vector system (http://www.addgene.org/TALEN_genome_editing_collection). Section 2.4 introduces the CRISPR/Cas system and describes the process of designing and constructing custom CRISPRs. Section 2.5 provides a detailed protocol for performing TALEN or CRISPR/Cas9 genome editing in hPSCs, including a screening step for identifying modified clones. Our system does not require the use of antibiotic selection to identify modified hPSC lines; if gene targeting with antibiotic selection is desired, the same methods can be used, but the steps described in Section 2.5.1-3 can be omitted.
Figure 2-1. Overview of methods for TALEN genome editing in hPSCs

(A) Customized TALENs are cloned using a library of preassembled domains into pTALEN expression plasmids. (B) TALEN plasmids are introduced into hPSCs by electroporation with or without exogenous HDR repair template DNA. (C) 48 hours after electroporation, GFP positive TALEN expressing hPSCs are purified by FACS. (D) Isolated hPSCs are plated at a clonal density on 10cm plates and clones are grown for 10-14 days. (E) hPSC clones are manually picked into 96-well plates – 1 clone per well. (F-H) Each 96-well plate is later passaged to two plates, one for genomic DNA preparation and PCR screening and one to create a frozen stock of each clonal cell line. The entire process takes approximately 1 to 2 months depending on how quickly step A is completed.
Figure 2-2. Overview of methods for CRISPR/Cas9 genome editing in hPSCs

(A) A customized CRISPR is generated by cloning a custom guide-RNA (gRNA) into the gRNA expression plasmid. (B) The gRNA expression plasmid and an invariant Cas9:GFP expression plasmid are introduced into hPSCs by electroporation with or without exogenous HDR repair template DNA. (C-H) Same as described in Figure 2-1C-H
2.3. GENERATING CUSTOM TALENS

Introduction to TALENs

Site-specific genome editing was originally achieved using zinc finger nucleases (ZFNs); the general design and function of ZFNs illustrate basic principles of genome editing, which are common to other genome editing reagents developed in recent years. ZFNs are engineered fusion proteins consisting of customized zinc-finger DNA binding domains linked to the nuclease domain of the bacterial FokI restriction enzyme. These proteins are designed to operate in a paired fashion: customized ZNFs bind to DNA sequences on either side of a genomic site of interest; next, the FokI domains of the two ZFNs dimerize and create a DSB at the target site. This technology has been used in a wide range of genome editing applications and specific ZFNs are being tested for therapeutic uses in clinical trials.

A disadvantage of ZFN technology is that custom ZFNs are relatively challenging to design and construct, despite the availability of open-access design protocols and reagent libraries (Urnov et al., 2010). Following the demonstration of ZFN genome editing, a potentially more adaptable class of engineered proteins subsequently emerged. These “TALENs” employ the unique DNA recognition and binding scheme of “transcription activator-like effector” (TALE) proteins found in plant pathogens; like ZFNs, TALENs include FokI domains and dimerize to create DSBs (Zhang et al., 2011; Miller et al., 2011; Hockemeyer et al., 2011). Due to the modular nature of the TALE DNA binding domain, customized TALENs that target a sequence of interest can be readily designed and constructed through a process that is simpler than that used to engineer custom ZFNs (Joung and Sander, 2013).
2.3.1 Protocol

2.3.1-1 TALEN binding site criteria and TALEN design

TALEN binding sites are chosen such that the forward and reverse TALENs will flank the desired genome-editing site. TALEN-induced DSBs, which are the substrate for the genome-editing process, occur in the region between the TALEN binding sites. Choose TALEN binding sites that satisfy the following criteria (Figure 2-3):

- Each TALEN binding site is 15 bp long
- The spacing between the forward and reverse TALEN binding sites is between 13 and 13 bp; 16 bp is optimal
- The forward TALEN binds in the 5' to 3' direction on the forward DNA strand
- The reverse TALEN binds in the 5' to 3' direction on the reverse DNA strand
- The position 0 nucleotide (the nucleotide just 5' to the first position of the TALEN binding site) should ideally be a T for both TALENs
Figure 2-3. Example genome editing target site showing TALEN design criteria
The forward and reverse TALENs recognize the underlined sequences in the 5’ to 3’ direction on the top and bottom strands respectively.
2.3.1-2 TALEN subcloning overview

Once TALEN binding sites are chosen, custom forward and reverse TALENs are constructed that are specific to the 15-bp binding sites, each in the 5’ to 3’ direction. Each TALEN consists of a sequence-specific 15mer DNA-binding domain as well as a nuclear localization signal (NLS), invariant N-term and C-term domains, and a FokI nuclease domain. For the purpose of subcloning, the 15mer DNA-binding domain is broken into four “positions”: three tetramer-recognition pieces (positions I, II, and III) and one trimer-recognition piece (position IV).

Our TALEN construction library comprises a collection of 832 plasmids containing all of the position I, II, III, and IV pieces necessary to generate a TALEN for any 15-bp target sequence (as well as two backbone vectors; see below). Using this library, a custom TALEN is generated through a five-piece subcloning ligation involving three sequence-specific tetramer-recognition pieces, one trimer-recognition piece, and an expression vector backbone (pTAL). There are two distinct pTAL vectors (pTAL_GFP and pTAL_RFP); the forward and reverse TALENs must be subcloned into different pTAL vectors so that the two FokI domains of the TALENs will correctly heterodimerize once they are bound to their target sites in the genome (the FokI domains are designed not to homodimerize). Begin by determining which position I, II, III, and IV plasmids you will use to generate sequence-specific TALENs for your chosen binding sites.

2.3.1-3 Digest tetramer and trimer pieces for TALEN subcloning

Digest and gel-purify the appropriate position I, II, III, and IV plasmids (e.g., for
the binding site “GCCGGAGCTGTCGAC” digest the “I-GCCG,” “II-GAGC,” “III-TGTC,” and “IV-GAC” plasmids). Proceed as follows:

1. If starting with a glycerol stock or agar stab culture, inoculate into LB with ampicillin and/or kanamycin (use both antibiotics for the most stringent selection), grow, and miniprep

2. Setup digest: 5–15 uL plasmid DNA, 2 uL Tango buffer, 1 uL 20 mM DTT, 1 uL BsmBI enzyme, dH$_2$O to 20 uL reaction volume; optional: can add 1 uL NcoI to the reaction (adjust dH$_2$O so that final reaction volume is 20 uL) to digest the vector backbone and reduce carryover of uncut vector to subsequent steps

3. Incubate several hours at 37°C

4. Run digests on agarose gel; expected insert sizes are about 425 bp for positions I, II, and III and about 280 bp for position IV

5. Gel-purify each insert using the QIAEX II kit (highly recommended) and elute into 20 uL elution buffer (EB)

2.3.1-4 Digest pTAL backbone vectors for TALEN subcloning

Each of the pTAL plasmids contains the CAG promoter, all parts of the TALEN protein except the DNA-binding domain, and either green fluorescent protein (GFP; specifically EGFP) or red fluorescent protein (RFP; specifically mCherry). Upon ligation of the tetramer/trimer pieces, the plasmid will co-express (via a viral 2A motif) the TALEN and the fluorescent protein. Each pTAL plasmid should be digested and gel-purified (can be done in parallel with the tetramer/trimer plasmids, or can be done beforehand to make stocks for use in numerous TALEN ligations). Proceed as follows:
1. If starting with a glycerol stock or agar stab culture, inoculate into LB with ampicillin, grow, and miniprep

2. Setup digest: 5 uL plasmid DNA, 2 uL Tango buffer, 1 uL 20 mM DTT, 2 uL BsmBI enzyme, 10 uL dH₂O to 20 uL reaction volume

3. Incubate several hours at 37°C

4. Heat inactivate at 75°C for 10 minutes, then cool to room temperature

5. Add 27 uL dH₂O, 3 uL New England Biolabs (NEB) buffer III, and 1.5 uL CIP to make a 50 uL reaction volume

6. Incubate at least 30 min at 37°C

7. Add 1 uL 0.5 M EDTA

8. Heat inactivate at 75 degrees for 10 minutes, then cool to room temperature

9. Add 50 uL dH₂O and mix; final reaction volume should be 100 uL

10. Add 100 uL phenol/chloroform/isoamyl alcohol and vortex vigorously (be careful not to let contents leak out of tube)

11. Spin at maximum speed in microcentrifuge for 5 minutes

12. Carefully remove top (aqueous) phase 20 uL at a time, avoiding the white interface, and transfer to another tube; do this 4 times for total of 80 uL

13. Run 80 uL on an agarose gel (use wide, deep wells); expected size is about 6.5 kb

14. Gel-purify the vector (there should be no discernible insert) using the QIAEX II kit, elute into 20 uL elution buffer (EB)
2.3.1-5 TALEN subcloning

Assemble gel-purified tetramer and trimer inserts and pTAL vector via a five-piece ligation:

1. Set up ligation: 0.875 uL position I insert DNA, 0.875 uL position II insert DNA, 0.875 uL position III insert DNA, 0.875 uL position IV insert DNA, 0.5 uL pTAL vector DNA, 0.5 uL 10× T4 ligase buffer, 0.5 uL T4 ligase to make a final 5 uL reaction volume
2. Mix and incubate for several hours at room temperature
3. Transform into competent bacteria (e.g., DH10B)
4. Plate onto LB agar plate with ampicillin and grow colonies at 37°C

Screen colonies (putative TALEN constructs) for appropriate inserts:

1. Pick colonies into 5 mL of LB-ampicillin (typically 5-10 colonies are picked from each plate, i.e., for each TALEN)
2. Miniprep cultures, elute in 35-50 uL elution buffer
3. Digest 2 uL miniprep DNA with XbaI and BamHI for at least 30 min (the digestion may be somewhat inefficient and not go to completion in that time frame)
4. Run digests on agarose gel
5. The correct insert size is about 2.1 kb; may see evidence of incomplete plasmid digest

Sequence plasmids with correctly sized inserts using the primers TALseq+ (5’-
GGCCAGTGTGCTGAAGATCG-3’) and TALseq– (5’-CGCTACAAGATGATCATTAGTG-3’), which will read from either end of the insert and should cover the full DNA-binding domain if using a high-quality commercial DNA sequencing service (need at least 850 bp of reliable sequence from each read). Merge the DNA sequences via the overlap at the ends of the sequence reads, “translate” into amino acid sequences, and check to insure that the “code” of the DNA-binding domain is correct (Figure 2-4).
Figure 2-4. Generic TALEN amino acid sequences (to recognize 15 bp DNA sequences)

Repeat domain sequences shown in blue font with repeat variable di-residues shown in bold;
TAL effector N-term and C-term shown in orange font.
2.3.1-6 Maxipreps of TALEN plasmids

Once the forward and reverse TALENs are sequence-confirmed, produce maxiprep quantities of endotoxin-free TALEN plasmid DNA in preparation for delivery of the TALENs to hPSCs via electroporation. Grow maxiprep cultures in LB-ampicillin. Elute or resuspend (depending on maxiprep protocol) plasmid DNA in dH$_2$O to concentrations between 1 ug/uL and 3 ug/uL. 15 to 25 ug of each TALEN plasmid will be needed for each electroporation (see Section 2.5 for details).

2.3.1-7 Designing a knock-in ssODN

If an ssODN will be used to introduce a desired sequence into the genome, design a 60-70 bp oligonucleotide (matching the antisense strand if targeting a coding region; otherwise, the oligonucleotide can match either strand) with the non-homologous nucleotide(s) located near the center of the oligonucleotide sequence. To aid in screening for hPSC clones that correctly incorporated the desired sequence, it is helpful if the ssODN is designed to insert or remove a restriction site. To insert a restriction site into a coding sequence, consider incorporating a synonymous nucleotide substitution that creates a new restriction site without changing the protein-coding sequence into the ssODN.

Obtain a salt-free oligonucleotide produced at an appropriate scale (e.g., 50 nmol); 30 ug of the oligonucleotide will be needed for each electroporation (see Section 2.5 for details). PAGE purification is not necessary for oligonucleotides less than 70 bp long. Resuspend the oligonucleotide in dH$_2$O to a concentration of 2 ug/uL.
2.3.2 Materials

- TALEN construction library (available from Addgene)
- LB, LB agar, ampicillin, kanamycin (any source)
- BsmBI enzyme (Thermo Scientific/Fermentas)
- Tango buffer (included with BsmBI enzyme)
- 20 mM DTT (any source)
- dH₂O (any source)
- Ncol, XbaI, and BamHI enzymes (New England Biolabs or other source)
- QIAEX II kit (Qiagen)
- NEB buffer III (New England Biolabs)
- CIP (New England Biolabs or other source)
- 0.5 M EDTA (any source)
- Phenol/chloroform/isoamyl alcohol (any source)
- T4 ligase (New England Biolabs or other source)
- T4 ligase buffer (included with T4 ligase)
- Competent bacteria cells (e.g., DH10B), either homegrown or commercial
- TALseq+/TALseq– primers and ssODN (any source)

2.3.3 Troubleshooting

2.3.3-1 TALEN assembly ligations

If the TALEN assembly ligation repeatedly fails to produce any colonies or if none of the screened colonies have an insert that is the correct size, consider repeating the tetramer/trimer vector and/or backbone vector digests and gel purifications before
repeating the assembly ligation.

If the TALEN ligation products have inserts that are ~300-400 bp in size, these likely represent uncut or partially cut tetramer/trimer vector DNA that has been propagated through the steps of the protocol and ended up in the final ligation reaction. Adding Ncol to the original tetramer/trimer vector digestions can minimize this “background” vector.

If the TALEN ligation products have inserts that are the correct size but sequencing shows that the final 15mer DNA-binding domain is incorrect, consider sequence-confirming the tetramer/trimer plasmids (sequence with the universal T7 primer; expected sequences are available on the Addgene page) that were used in Section 2-3.1-3.

2.3.3-2 Sequence-confirming assembled TALENs

Sequencing the final TALEN construct may be difficult due to the reparative TAL repeat domain sequences; high-quality, concentrated plasmid DNA (grown from >5 ml culture and eluted in a small volume, e.g., 35 uL) is best for sequencing.

2.3.3-3 Maxipreps of TALEN plasmids

TALEN plasmids are relatively unstable; when doing maxipreps, use LB broth with fresh ampicillin and do not grow the maxiprep culture for more than 18 hours. Consider inoculating the maxiprep culture from a medium sized LB-ampicillin culture (<50 mL) grown for a few hours. In addition to using a spectrophotometer to quantitate the maxiprep DNA, digest a small amount of DNA and run it on an agarose gel to
confirm that the actual amount of digested DNA matches the expected amount.
2.4. GENERATING CUSTOM CRISPRs

Introduction to the CRISPR/Cas system

In 2013, a novel genome editing technology – derived from components of a bacterial adaptive immunity system – was first demonstrated. The clustered regularly interspaced short palindromic repeat (CRISPR)/Cas system consists of an RNA-guided nuclease (Cas protein) and a customizable RNA (“guide RNA”), which directs the CRISPR/Cas complex to a specified target site (Jinek et al., 2012; Mali et al., 2013; Cong et al., 2013). Like ZFNs or TALENs, CRISPRs generate DSBs at their target sites in genomic DNA, and can be used for genome editing in hPSCs (Mali et al., 2013). Notably, there are several key advantages of the CRISPR/Cas technology compared with previous systems. Many of these advantages stem from the fact that the CRISPR/Cas system is RNA guided, in contrast to ZFNs and TALENs, which are engineered DNA binding proteins (advantages and disadvantages of each technology are discussed in more detail in Chapter 7 of this thesis).

Since the target specificity of the CRISPR/Cas system is determined simply by a short sequence within the guide RNA, it is more substantially more straightforward to generate custom CRISPRs than it is to construct custom TALENs or ZFNs. Moreover, the CRISPR/Cas system is especially versatile compared with other technologies; in particular, it is more amenable to multiplexed or high-throughput genome editing applications than TALENs (Dr. George Church, Personal communication, 2010). To assess the efficiency of CRISPR genome editing in hPSCs, we adapted the CRISPR/Cas9 technology to our genome editing system and compared the activity of TALENs and CRISPRs targeting 8 loci within 7 genes in hPSCs. Strikingly, CRISPRs
were significantly more active than TALENs, producing at least one mutant allele with efficiencies ranging from 51% to 79% (Ding et al., 2013b; Chapter 4 of this thesis).

2.4.1 Protocol

2.4.1-1 CRISPR/Cas9 binding site criteria and CRISPR design

CRISPR/Cas9 binding sites are chosen to conform to the sequence G(N)_{19}NGG; the CRISPR RNA component hybridizes with the first 20 DNA nucleotides, and the Cas9 protein recognizes the NGG motif (Figure 2-5). CRISPR-induced DSBs occur three nucleotides proximal to the NGG motif. The CRISPR/Cas9 can recognize and bind to the motif on either DNA strand. Choose CRISPR/Cas9 binding sites that satisfy the following criteria:

- Match the G(N)_{19}NGG sequence in proximity to the desired target site; usually, but not always, there will be a matching sequence within 16 nucleotides of the target site
- Ideally, the binding site should not exactly match or mismatch by one nucleotide any other sequence in the genome
**Figure 2-5.**

**WT binding site:** CACAGCTGAGGTGCTGGGGGAGGTGCTACGGGACAGCG

**“Q121X” ssODN:** CACAGCTGAGGTGCTGGGGGAGGTGCTACGGGACAGCG

**Figure 2-5. CRISPR protospacer design with example ssODN for generating knock-in alleles**

In this example of CRISPR design, the codon of interest for targeting (number 121) is bolded and underlined; the desired mutation in this example (Q121X) is indicated in red. Highlighted in gray and yellow respectively are the recognition site (the “protospacer”) and protospace-adjacent motif (PAM) of the chosen CRISPR. It is ideal for the protospacer sequence to begin with a “G” (shown by green highlighting), because this base is part of the promoter sequence required for proper gRNA expression. The position of this recognition site relative to the desired knock-in mutation (the red “T”), should minimize mutagenesis by CRISPR re-cutting upon incorporation of the ssODN into the genome, because a mismatch in the 3’ portion of the protospacer sequence is predicted to decrease recognition of this site by the CRISPR.
2.4.1-2 CRISPR/Cas9 overview

The Cas9 component is invariant. Cas9 is expressed from the pCas9_GFP plasmid, which is identical to the pTAL_GFP plasmid except that the TALEN has been replaced with a human-codon-optimized Cas9 gene. This plasmid co-expresses Cas9 and GFP.

The CRISPR RNA component is expressed as a small guide RNA (gRNA) from the human U6 promoter. The 5’ end of the gRNA contains the 20-bp sequence [G(N)_{19}] that hybridizes to the DNA target site; accordingly, changing this 20-bp sequence is all that is necessary to direct the complex to a different site in the genome. Any of a variety of methods can be used to specify the 20-bp sequence. The gRNA Empty Vector from George Church’s laboratory (Mali et al., 2013) can be obtained from Addgene (https://www.addgene.org/CRISPR/Church/); protocols for inserting the 20-bp sequence into this vector is available there (https://www.addgene.org/static/cms/files/hCRISPR_gRNA_Synthesis.pdf).

2.4.1-3 Maxipreps of CRISPR/Cas9 plasmids

Once the gRNA plasmid is sequence-confirmed, produce maxiprep quantities of endotoxin-free Cas9 and gRNA plasmid DNA in preparation for delivery of the CRISPR/Cas9 components to hPSCs via electroporation. Grow maxiprep culture of pCas9_GFP in LB-ampicillin; if using a gRNA plasmid based on the Church laboratory’s vector, grow maxiprep culture in LB-ampicillin/kanamycin. Elute or resuspend (depending on maxiprep protocol) plasmid DNA in dH_{2}O to concentrations between 1 ug/μL and 3 ug/μL. 15 to 25 ug of each plasmid will be needed for each electroporation
(see Section 2.5 for details).

2.4.1-4 Designing a knock-in ssODN

Same as for TALENs (see Section 2.3.1-7).
2.5. GENOME EDITING IN HUMAN PLURIPOTENT STEM CELLS

2.5.1 Protocol

2.5.1-1 Expansion of hPSCs and preparation of conditioned media

1. Expand an hPSC line (e.g., HUES 9, Cowan et al., 2004) to 15-cm dishes. One
15-cm dish is used per electroporation; typically 1 to 4 electroporations are
performed for a particular genome-editing experiment. Culture the cells using
feeder-free adherent culture conditions in chemically defined mTeSR1 media
supplemented with penicillin/streptomycin on dishes pre-coated with Geltrex
matrix (250 uL Geltrex per 15-cm dish) (Schinzel et al., 2011). Passage cells with
Accutase [diluted 1:3 with calcium- and magnesium-free PBS (PBS–/–)]
dissociation.

2. Replace mTeSR1 with fresh media (15-20 mL) every 24 hours; collect
conditioned mTeSR1 (except during first media change, 24 hours after plating)
until the cells are approximately 80% confluent, replace with fresh mTeSR1.

3. Make “post-FACS recovery media” using conditioned mTeSR1: 1 volume
conditioned mTeSR1, 1 volume fresh mTeSR1, 35 uL 10,000× bFGF per 500 mL
total media volume; filter-sterilize using tissue-culture media filter bottle (note:
gentamicin can be added to the media for a few days following FACS to prevent
FACS-related bacterial contamination).

2.5.1-2 TALEN or CRISPR/Cas9 delivery by electroporation

Electroporate cells once they have grown to approximately 80% confluence:

1. Pre-treat cells with 10 uM ROCK inhibitor (Y-27632) for 3–4 hours prior to
electroporation; add ROCK inhibitor directly to media.

2. Dissociate cells with Accutase; assure single cell dissociation.

3. Count cells and resuspend 10-12×10^6 cells (roughly the number of cells on a 15-cm dish) in 800 uL ice-cold PBS−/− containing 50 ug of the TALEN pair (25 ug of each plasmid) or a mix of 30 ug of the TALEN pair (15 ug of each plasmid) plus 30 ug ssODN or double-strand plasmid DNA for each electroporation; alternatively, setup one electroporation per 15-cm dish of cells without counting the cells. Perform 1-4 electroporations simultaneously. If using the CRISPR/Cas9 system, resuspend the cells in 800 uL ice-cold PBS−/− containing 25 ug pCas9_GFP plasmid and 25 ug gRNA expression plasmid or 15 ug pCas9_GFP plasmid, 15 ug gRNA plasmid, and 30 ug ssODN or double-strand plasmid DNA for each electroporation. Note: it may be useful to also setup a “mock electroporation”; see Section 2.5.1-3, step 6. The mock electroporation can be smaller in scale: use one-third the number of cells.

4. Transfer the 800 uL cell/DNA mixture into a 0.4 cm electroporation cuvette and incubate on ice for 5 minutes. If also performing a mock electroporation, transfer the 800 uL cell/PBS mixture (no DNA added) into a separate electroporation cuvette.

5. Electroporate using the following settings: 250 V and 500 uF (time constant should be between 10 and 14 seconds).

6. Quickly add 0.5 mL of mTeSR1 to each cuvette and transfer the contents of the cuvettes into 15-mL centrifuge tubes containing 2 mL of mTeSR1 per electroporation.
7. Gently pellet cells (200 g, 5 min, 25 °C).

8. Gently resuspend the pellet in 5-10 mL mTeSR1.

9. Plate cells to one Geltrex-coated 15-cm dish per cuvette for a total volume of 20 mL mTeSR1 per dish with 10 uM ROCK inhibitor. If also performing a mock electroporation, plate cells to one Geltrex-coated 10-cm dish (100 uL Geltrex per 10-cm dish) for a total volume of 10 mL mTeSR1 with 10 uM ROCK inhibitor. Note: if performing genome editing with antibiotic selection, the post-electroporation plating density should be optimized so that isolated antibiotic resistant hPSC clones are generated over the course of antibiotic selection.

10. The day after electroporation, replace media with fresh mTeSR1 without ROCK inhibitor. Note: if using antibiotic selection, begin selection 24 to 48 hours after electroporation and continue selection until suitably sized antibiotic-resistant colonies appear (see Section 2.5.1-4).

### 2.5.1-3 Fluorescence-activated cell sorting (FACS) enrichment and clonal isolation

The following steps are used for efficient genome editing without antibiotic selection; proceed to Section 2.5.1-4 if using selection.

1. After ~48 hours following electroporation, confirm successful delivery of TALEN plasmids (or CRISPR/Cas9 plasmids) by visualizing some green and/or red fluorescent marker expression using fluorescence microscopy (optional); proceed with FACS enrichment of fluorescence-expressing cells.

2. Dissociate cells with Accutase; assure dissociation to single cells. Combine cells from multiple replicate electroporations into a single tube.
3. Carefully resuspend cells in about 500 μL PBS–/– with 10 uM ROCK inhibitor, pass through a 35-um cell strainer cap into a FACS tube, and dilute with additional PBS–/– as needed. Higher cell concentrations will allow for more rapid FACS and possibly greater cell viability (see Section 2.5.3-3) but may increase the risk of cell clumping during FACS.

4. Prepare 1.5 mL Eppendorf tube (for collection of cells) containing 700 μL mTeSR1 with 10 uM ROCK inhibitor.

5. Keep cells and collection tube on ice.

6. Using FACS, collect single cells that express green and/or red fluorescence into Eppendorf tube containing mTeSR1 (Figures 2-6 and 2-7). Use unelectroporated or mock-electroporated (PBS only) cells to set gating for fluorescence.

7. Plate post-FACS cells on Geltrex-coated 10-cm dishes at a density of 15,000 cells per 10-cm dish in post-FACS recovery media with 10 uM ROCK inhibitor. Plating density may be increased up to 50,000 cells per 10-cm dish if too few colonies are obtained (see Section 2.5.3-3).

8. Change media with fresh post-FACS recovery media (without ROCK inhibitor) every 24 hours. Once multi-cellular colonies are clearly visible (typically 2 or 3 days after FACS), mTeSR1 can be used instead of recovery media.
Figure 2-6. Representative FACS plots illustrating isolation of TALEN expressing hESCs
(A) Mock electroporated hESCs (cells electroporated without TALEN plasmid DNA). (B) TALEN-electroporated hESCs. Percentage of P2-gated cells expressing GFP and/or RFP based on gates shown: GFP only 12%, GFP + RFP 0.7%, RFP only < 0.1%. For the electroporated sample, cells from the GFP, RFP, and GFP + RFP gates are sorted into a single 1.5 mL collection tube.
Figure 2-7. Representative FACS plots illustrating isolation of Cas9 expressing hESCs

(A) Gating strategy: gating for live cells and single cells, left and right plots respectively. (B) GFP gating of mock electroporated hESCs (cells electroporated without Cas9 plasmid DNA, left plot) and CRISPR/Cas9-electroporated hESCs (right plot). Percentage of single cells expressing GFP based on gates shown: 14.9% (CRISPR/Cas9 electroporation), 0.12% (mock electroporation). For the CRISPR/Cas9-electroporated sample, cells from the GFP+ gate are sorted into a 1.5 mL collection tube.
2.5.1-4 Clone expansion and freezing

Begin picking colonies (typically 12 to 14 days after electroporation) when they are about the size of a nickel under the microscope with a 10× objective (Figure 2-8A); do not allow colonies to grow too large and merge.

1. Coat a 96-well plate with Geltrex, using the same amount of Geltrex as for a 10-cm plate (100 uL Geltrex per plate).

2. Wash the colony plate with calcium- and magnesium-free PBS (PBS−/−).

3. Add 12 mL fresh PBS−/− to the plate and return the plate to the incubator for about 5 minutes to allow the cells to begin to detach (this greatly facilitates re-plating into the 96-well plate).

4. Prepare the 96-well plate with 120 uL mTeSR1 per well plus 10 uM ROCK inhibitor (not including volume of PBS/cells that you will be transferring into the wells).
Figure 2-8. Representative images showing the expansion and screening of hESC lines during the genome editing process

(A) A hESC colony 12 days after clonal plating following FACS. (B) hESCs in a 96-well 1 day after manual picking. (C) hESCs in a 96-well 4 days after picking when they are ready to be passaged. (D) PCR screening the targeted genomic locus using genomic DNA from individual hESC clones. Clones that likely harbor TALEN-generated mutations are indicated (arrows).
5. Using a microscope setup in a tissue-culture hood or colony-picking hood pick colonies using a P200 pipette (set to about 45 uL) with filter tips by scraping the colony into small clumps of cells, collecting the cells with the pipette, and gently pipetting cells into a well of the 96-well plate. Use a new pipette tip for each colony. Try to collect all of the dissociated cells from a scraped colony; excess dissociated cells will float away in the PBS and may be accidently collected when subsequent colonies are picked. Pick colonies until a suitable number of colonies are isolated or the 10-cm plate contains a significant number of unwanted floating cells that cannot be avoided (typically after about 1 hour of picking). Note: the efficiency of TALEN-mediated genome editing using this system typically results in the generation of modified hPSC lines at a rate of 5% to 20% among the total number of colonies picked; CRISPR/Cas9-mediated genome editing appears to be even more efficient.

6. Change the medium (120 uL per well) in the 96-well plate every 24 hours (optional: add gentamicin to the media for a few days to prevent picking-related bacterial contamination) until the wells are about 80% confluent (typically 4 days after picking; Figures 2-8B and 2-8C), when they can be split and frozen. Refer to Section 2.5.3-5 if the cell density varies significantly from well to well on an individual 96-well plate.

7. Coat 1 or 2 Geltrex-coated 96-well plates per 96-well plate being passaged; 1 plate is for making DNA and 1 plate is for passaging and serves as an optional backup plate.

8. Prepare the Geltrex-coated 96-well plate(s) with 133 uL of mTeSR1 plus 10
uM ROCK inhibitor per well (the amount of ROCK inhibitor added should be based on the final volume per well, which is 200 uL; see step number 13).

9. Prepare an uncoated 96-well plate, which will be used for freezing.

10. Aspirate media from the original 96-well plate; wash once with 150 uL PBS—/— per well and aspirate.

11. Add 35 uL Accutase (1:3 diluted in PBS—/—) per well; treat in incubator for 5-10 min until most of the cells are detached from the plate (lightly tapping the side of the plate near the end of the treatment can help detach the cells).

12. Add 165 uL mTeSR1 per well directly to the plate.

13. Using a multichannel pipette, distribute 66 uL of cells (one third of the total volume of cells per well) to each of the 2 or 3 96-well plates—DNA plate, backup plate (optional), and freezing plate—and pipette up and down gently 1.5 times before disbursing the cells to ensure that the cells are equally distributed. Return the cells to the incubator.

14. Add 70 uL mFreSR per well to the freezing plate; wrap plate with Parafilm and move to a Styrofoam container at −80°C for freezing and storage.

15. On the next day, exchange the media on the new 96-well plates with 120 uL fresh mTeSR1 per well.

16. Change media every 24 or 48 hours.

**2.5.1-5 Clone screening**

Isolate genomic DNA (gDNA) from the cells in the 96-well plate designated for making DNA when the cells in each well are about 90% confluent:
1. Wash twice with 150 uL PBS per well.

2. Add 50 ul 96-well lysis buffer per well: 10 mM Tris pH 7.5 (or 8.0), 10 mM EDTA, 10 mM NaCl, 0.5% sarcosyl, 40 ug/mL proteinase K (added fresh before using the buffer).

3. Cover the plate with PCR plate film, replace the plate top, wrap in Parafilm, and wrap in plastic wrap or place in sealed plastic box with damp paper towel.

4. Incubate at 56-60°C overnight or for ≥6 hours.

5. Precipitate DNA by adding 100 uL cold 95% ethanol (not 100%) plus 75 mM NaCl (150 uL 5 M NaCl per 10 mL ethanol) per well; leave at –20°C for 2 hours.

6. Spin plate for 5 minutes at 1000 rpm.

7. Carefully decant liquid out of plate; the gDNA should stick to the bottom or side of each well.

8. Wash three times with 150 uL 70% ethanol per well; carefully decant after each wash.

9. Let residual ethanol evaporate at room temperature or 60°C; be careful not to let the gDNA overdry.

10. Resuspend gDNA in 30 uL dH₂O plus 0.1 mg/mL RNase A per well, cover the plate with PCR plate film, and place on shaker at moderate speed (to avoid shearing of DNA) for about 2 hours. Leave at room temperature overnight and store at –20°C.

11. Screen hPSC clones by PCR amplification of the targeted genomic region: use primers that amplify a short product (150-200 bp) around the genomic
target site; set up a PCR reaction for each hPSC clone using 96-well PCR plates and a multichannel pipette to pipette PCR master mix and gDNA; use 2.5 uL gDNA per 25 uL PCR reaction (consider increasing amount of template DNA if no PCR products are obtained); include a negative control PCR reaction (H₂O only); run PCR products on a high percentage (2.5%) agarose gel (load gel using a multichannel pipette) for 2-3 hours to separate wild type from indel-containing products (Figure 2-8D). If a restriction site was inserted into the genomic target site to facilitate screening (see Section 2.3.1-7), digest 3 uL of the PCR reaction for each hPSC clone in a 20 uL restriction digest; PCR purification prior to restriction digestion is not necessary if the volume of the PCR reaction digested is kept relatively low; run the restriction digest on a high percentage (2.5%) agarose gel to separate undigested and digested PCR products.

12. Confirm suspected mutant clones by direct sequencing of the PCR products and sequencing of TOPO-cloned PCR products.

2.5.1-6 Clone recovery

1. Prepare Geltrex-coated 24-well plate (1 well each for mutant and control clones of interest).

2. Add 500-750 uL mTeSR1 per well plus 10 uM ROCK inhibitor (taking into account the final volume per well).

3. Remove frozen 96-well plate from –80°C freezer, clean outside of plate with 70% ethanol, remove Parafilm, and thaw quickly by placing the plate in a
37°C tissue-culture incubator (wells on the borders of the plate will thaw more quickly).

4. When wells of interest have only a small piece of ice remaining, move plate into a tissue-culture hood and add 100 uL room temperature mTeSR1 directly to each well.

5. Quickly transfer the thawed cells using a P200 pipette (using a new filter tip for each well) into the wells of the 24-well plate.

6. Replace the media with fresh mTeSR1 (without ROCK inhibitor) every 24 hours.

7. Once large colonies with hPSC morphology appear, expand to one well of a 6-well plate, and then expand to a 10-cm plate (some colonies will not thaw successfully and will be lost due to poor viability or differentiation, so it is recommended that several suspected mutant and control clones be thawed).

2.5.2 Materials

- 15-cm, 10-cm, 6-well, 24-well, and 96-well tissue culture plates (any source)
- mTeSR1 media (STEMCELL Technologies)
- Penicillin/streptomycin (Gibco or other source)
- Geltrex matrix (Life Technologies)
- Accutase (STEMCELL Technologies)
- PBS, calcium- and magnesium-free (Gibco or other source)
- bFGF (Aldevron)
- Media filter bottles (VWR or other source)
• Gentamicin reagent solution (Gibco or other source)
• ROCK inhibitor, Y-27632 (Santa Cruz Biotechnologies, sc-281642A)
• Electroporator (Bio-Rad or other source)
• Electroporation cuvettes, 0.4 cm (Bio-Rad or other source)
• FACS tube (BD Biosciences or other source)
• mFreSR (STEMCELL Technologies)
• Proteinase K (Roche or other source)
• RNase A (Invitrogen or other source)
• PCR plate film (VWR or other source)
• TOPO TA Cloning Kit (Invitrogen)

2.5.3 Troubleshooting

2.5.3-1 Low percentage of GFP- and/or RFP-positive cells after electroporation

Unknown factors lead to variable percentages of GFP- and/or RFP-positive cells from one electroporation attempt to another, and the percentage of GFP- and/or RFP-positive cells may be low (< 5% GFP-positive and nearly 0% RFP-positive, see Section 2.5.3-2). However, we have not noticed a correlation between the percentage of fluorescent cells and the efficiency of genome editing using this protocol. Therefore, the observed fluorescence percentages are not important if a suitable number of cells can be collected for clonal plating. Although the fluorescent cell percentage may not be important, note the following: 15,000 to 50,000 cells are needed for post-FACS clonal plating per 10-cm plate (see Section 2.5.1-3); it is important that the cells being collected are truly fluorescent (see gating strategy illustrated in Figures 2-6 and 2-7);
and FACS should be completed relatively quickly (less than 1.5 hours of total FACS time, see Section 2.5.3-3) to avoid reduced cell viability upon clonal plating.

The percentage of GFP- and/or RFP-positive cells may be influenced by the rate of cell division after electroporation, which may be greater for some hPSC lines; cell division after electroporation appears to result in a lower GFP/RFP-positive percentage during FACS, likely due to dilution of plasmid DNA and GFP/RFP as the cells divide.

2.5.3-2 Very low percentage of RFP-positive cells after electroporation

For unclear reasons, the percentage of RFP-positive and GFP/RFP-double-positive cells during FACS is typically an order of magnitude less than the percentage of GFP-positive cells (see Figures 2-6 and 2-7). This discrepancy does not affect the efficiency of genome editing. GFP-positive, RFP-positive, and GFP/RFP-double-positive cells can be collected into a single collection tube (see Section 2.5.1-3).

2.5.3-3 Few or no surviving colonies on post-FACS 10-cm plate

As described in Section 2.5.1-3, different hPSC lines exhibit different survival rates during clonal plating to 10-cm plates after FACS, necessitating some optimization of plating density for a given cell line (generally ranging from 15,000 to 50,000 cells per 10-cm plate). Therefore, if no colonies survive clonal plating, consider plating at a higher density.

Conditioned media (see Section 2.5.1-1) increases the survival rate during clonal plating after FACS and should be used after FACS until small colonies appear (typically 2 to 4 days after FACS), at which point normal mTeSR1 media is used.
The post-FACS survival rate may decrease as the overall FACS time increases.

It is best to complete the entire FACS-enrichment process (see Section 2.5.1-3)—beginning with collecting the cells and ending with plating to 10-cm plates—within 1.5 hours. If the FACS is proceeding slowing (due to a low percentage of fluorescent cells), consider plating the appropriate number of cells to a 10-cm plate once that amount of cells has been collected, while the FACS is still proceeding.

2.5.3-4 Bacterial contamination occurs on post-FACS 10-cm plates or 96-well plates

If bacterial contamination is observed on any of the post-FACS 10-cm plates or 96-well plates (likely due to contamination from the flow cytometer or the multichannel aspirator respectively), throw away the contaminated plates and add 50 ug/mL gentamicin to the media of any remaining plates.

2.5.3-5 Cell density in 96-well plates varies from well to well

After picking hPSC colonies to 96-well plates, the cell density may be appreciably different from well to well within an individual plate. If this is the case, some wells may be ready to be passaged (i.e., reach 80% confluence) before other wells. To avoid this problem, when picking colonies, choose colonies that are similar in size. When considering when to passage a 96-well plate where the cell density varies between wells, wait to passage the plate until the majority of the wells have reached a suitable level of confluence.
Screening PCR reactions fail for a significant number of hPSC clones

The PCR reaction should be optimized with control genomic DNA before setting up large-scale screening PCR reactions to ensure that PCR products will be obtained for most of the hPSC clones during the screening PCR. Failure of a significant number of PCR reactions during the large-scale screening PCR may occur if the PCR reaction is not well-optimized; if this is the case, you may observe small bright primer dimer bands on the gel. Setting up the screening PCR reactions on ice as quickly as possible will increase the chances of success.

Poor-quality or low-concentration template genomic DNA from most or all of the hPSC clones may cause the screening PCR reactions to fail. Ensure that the DNA is fully dissolved before setting up the screening PCR reactions; consider increasing the volume of template genomic DNA used.

Lastly, be careful not to lose the precipitated genomic DNA pellets from the wells of the 96-well plate when preparing genomic DNA (Section 2.5.1-5). Precipitating the genomic DNA for several hours at –20°C and spinning the plates down after precipitating the DNA helps diminish this outcome.
2.6. DISCUSSION

The protocols described here facilitate efficient genome editing in hPSCs using TALENs or CRISPR/Cas9. Notably, our approach employs cell-sorting following electroporation to isolate hPSCs that express TALENs or CRISPR/Cas9 (the mediators of genome editing), overcoming a significant obstacle to facile gene targeting in hPSCs: inefficient delivery of DNA. This overall approach is straightforward, takes only a few weeks in total, does not require antibiotic selection, and has a wide array of potential applications.

It should be noted that there are a number of potential caveats related to the genome editing approach, in particular the possibility of off-target mutagenesis caused by engineered nucleases. To begin to investigate this possibility, in Chapter 3 of this thesis, exome sequence analysis is performed on TALEN genome-edited hPSC clones. In addition, specific strategies to circumvent potentially confounding effects related to the genome editing process are demonstrated. These issues are discussed at greater length in Chapter 7 of this thesis, taking into account the findings of a number of pertinent recently published studies.

As with any research approach, the advantages and possible pitfalls of genome editing in hPSCs should be carefully considered when designing an experimental strategy. A growing number of published examples suggest ways in which genome editing techniques can be used to empower hPSC research, in particular for in vitro disease modeling studies. The efficient genome editing system detailed in this thesis chapter greatly expands what is possible within the field of hPSC research.
ACKNOWLEDGEMENTS

This work was supported in part by the Sternlicht Director’s Fund Award for Graduate Students from the Harvard Stem Cell Institute (D.T.P.); grants R00-HL098364 (K.M.), U01-HL107440 (C.A.C.), and R01-DK097768 (K.M., C.A.C.) from the United States National Institutes of Health (NIH); the Broad Institute’s Lawrence H. Summers Fellowship and the Carlos Slim Foundation (K.M.); the Harvard Stem Cell Institute (K.M., C.A.C.), and Harvard University (K.M., C.A.C.). The authors report no relevant conflicts of interest. We thank J. Keith Joung, Feng Zhang, George Church, Prashant Mali, Qiurong Ding, Nicolas Kuperwasser, Jennifer Shay, and the staff of the HSCRB-HSCI Flow Cytometry Core for assistance and suggestions.
REFERENCES


Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome


Watanabe K, Ueno M, Kamiya D, Nishiyama A, Matsumura M, Wataya T, Takahashi


Chapter 3: Isogenic hESC-Based Disease Models Clarify the Role of SORT1 as a Mediator of Disease-Related Cellular Functions in Human Metabolic Cell Types

The majority of this chapter was previously published in:

SUMMARY

Transcription activator-like effector nucleases (TALENs) are a new class of engineered nucleases that have proven to be easier to design to bind and cleave at desired sites in the genome than previous types of nucleases. Here we report the use of TALENs to rapidly and efficiently generate knockout or mutant alleles of 13 genes in either cultured somatic cells or human pluripotent stem cells, the latter of which we differentiated both the targeted lines and isogenic control lines into various cell types relevant to metabolic disorders. We demonstrate cell-autonomous phenotypes that are directly linked to disease—dyslipidemia, insulin resistance, hypoglycemia, lipodystrophy, motor neuron death, and hepatitis C infection—and yield new biological insights. Given the speed and ease with which we were able to derive these cell lines, especially in comparison to transgenic animals, and the ability to interrogate for clinically relevant phenotypes in multiple cell types, we anticipate that TALEN-mediated genome editing of human cells will become a mainstay for the investigation of human biology and disease. For this thesis chapter, we will focus in particular on the studies pertaining to APOB and SORT1, which are the most relevant to cardiovascular disease and illustrative of key aspects of in vitro disease modeling.
INTRODUCTION

The study of human disease has been facilitated by the ability to identify responsible gene mutations; at the same time, it has been hampered by the lack of an inexhaustible supply of easily accessible tissues from patients bearing those mutations. Another limitation is that many gene mutations that would be informative for disease biology if they could be studied in isolated cells are incompatible with human life (i.e., embryonic lethal). Classical gene targeting technology uses homologous recombination to target an investigator-specified gene for disruption or modification. It has proven to be an invaluable tool of experimental biology through its use in mouse embryonic stem cells to generate germline knockout and knock-in mice; however, its use in mammalian systems has been limited primarily to studies in mice. In many cases, mice do not faithfully phenocopy human physiology and disease, e.g., cholesterol metabolism, coronary artery disease, and human hepatitis C virus (HCV) infection. The emergence of genome editing with engineered nucleases, as well as human pluripotent stem cell (hPSC) technology and differentiation protocols to obtain a variety of cell and tissue types in vitro, now make it possible to rapidly interrogate the effects of genetic modification in otherwise isogenic human model systems.

Transcription activator-like effector nucleases (TALENs) are a new class of engineered nucleases that due to their modular domain structure have proven more straightforward to design and construct to perform genome editing than previous types of nucleases (Bogdanove and Voytas, 2011). TALENs are typically designed as a pair that binds to genomic sequences flanking a target site and generates a double-strand break (DSB), which is repaired by the cell using either homology-directed repair (HDR)
or the error-prone process of non-homologous end-joining (NHEJ) (Christian et al., 2010; Li et al., 2011; Miller et al., 2011; Hockemeyer et al., 2011). NHEJ can be exploited to introduce small insertions or deletions (indels) resulting in frameshift mutations that effectively knock out a protein-coding gene. An exogenously introduced double-stranded DNA or single-stranded DNA oligonucleotide (ssODN) can serve as a repair template for HDR to incorporate an alteration into the genome (Radecke et al., 2010; Chen et al., 2011; Soldner et al., 2011). In principle, TALEN pairs can be generated de novo in a matter of days (Cermak et al., 2011; Sanjana et al., 2012; Reyon et al., 2012).

To demonstrate the utility, efficiency, and rapidity of TALEN technology in generating human cellular models with which to derive new biological insights, we created mutations in 13 genes and performed detailed phenotypic analysis of four genes for which novel roles in disease biology have emerged in recent years—APOB, SORT1, AKT2, and PLIN1. For this thesis chapter, we will focus in particular on the studies pertaining to APOB and SORT1, which are the most relevant to cardiovascular disease and illustrative of key aspects of in vitro disease modeling.
RESULTS

Modular Assembly and Use of TALENs for Efficient and Rapid Genome Editing

The DNA-binding domain of a TALEN comprises arrays of 33- to 35-amino-acid monomers that are “coded” to recognize and bind specific DNA basepairs in a 1:1 fashion (Moscou and Bogdanove, 2009; Boch et al., 2009). Current methods for TALEN design and construction require either several intermediate steps to assemble full TAL repeat arrays from individual monomers, which can take days or weeks to accomplish and verify, or specialized magnetic and liquid-handling equipment to perform the assembly in an iterative fashion (Li et al., 2011; Cermak et al., 2011; Sanjana et al., 2012; Reyon et al., 2012). We built upon previously described modular Golden Gate methodologies to allow assembly of multiple DNA fragments in an ordered fashion in a single reaction to eliminate intermediate steps, such that a single ligation of pre-assembled tetramers/trimers generates TALENs that recognize any 15-bp recognition site in the genome (Figure 3-1). This assembly method requires only 1-2 days for completion and is not prone to errors that complicate methods that rely on polymerase chain reaction (PCR) amplification of monomers. Further, we have developed a set of optimized vectors and methods for the delivery of TALENs into mammalian cells and, in particular, hPSCs. Briefly, we transflect or electroporate TALEN pairs into cells and then subject them to fluorescence activated cell sorting (FACS) 48 hours post-transfection based on green and/or red fluorescent marker expression (Figure 3-1). We replate the sorted cells at either limiting dilution or low density and allow them to recover and grow for 1 week, resulting in the formation of distinct single colonies. Colonies are expanded, genomic DNA purified, and mutations analyzed by PCR and Sanger sequencing.
(Figure 3-1). The entire process from start to finish can be completed in less than one month.
Figure 3-1.

Digest desired multimers from 832-plasmid library with BsmBI (type II* enzyme)

Ligate into a plasmid encoding TALEN + fluorescent protein

Transfect/electroporate into cells

Fluorescence activated cell sorting

Replate, grow, pick colonies, expand

Isolate DNA, PCR screen for mutations

wild-type (control) clones  mutant clones

Figure 3-1. Schematic of a system for efficient and rapid genome editing with TALEN
Utilizing these methods we generated TALEN pairs to target 14 distinct sites in 13 genes in human somatic cell lines, human embryonic stem cell (hESC) lines, or human induced pluripotent stem cell (iPSC) lines; the alterations included a variety of knockout mutations as well as a knock-in missense mutation and a functional frameshift mutation (Table 3-1). We observed that the efficiency of mutation varied by genomic location as well as among different cell lines, with indels from NHEJ occurring in roughly 2% to 30% of clones screened and the efficiency of knock-in by HDR occurring at a frequency of 1.6%. We then proceeded to perform detailed phenotypic analyses of cells harboring mutations in four human disease-related genes—APOB, SORT1, AKT2, and PLIN1.
Table 3-1. Targeting Efficiency of TALENs at 14 Loci in 13 Genes in Various Cell Types

* HUES 1 and HUES 9 are human embryonic stem cell lines (Cowan et al., 2004); BJ-RiPS and CF-RiPS are induced pluripotent stem cell lines (Warren et al., 2010); HuH-7 cells are cultured human hepatocellular carcinoma cells

† Successfully inserted E17K knock-in mutation using single-stranded DNA oligonucleotide

<table>
<thead>
<tr>
<th>Gene</th>
<th>Target Sequence (Flanked by TALEN Binding Sites)</th>
<th>Cell Line*</th>
<th>Number of Clones Screened</th>
<th>Number of Mutants</th>
<th>% Efficiency of Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT2</td>
<td>TTCAGGTGAATACAT</td>
<td>HUES 9</td>
<td>192</td>
<td>17</td>
<td>8.9</td>
</tr>
<tr>
<td>AKT2 E17K</td>
<td>TTCAGGTGAATACAT</td>
<td>HUES 9</td>
<td>192</td>
<td>3†</td>
<td>1.6</td>
</tr>
<tr>
<td>ANGPTL3</td>
<td>TCCTAGAAGAAAAAAA</td>
<td>HUES 9</td>
<td>424</td>
<td>84</td>
<td>19.8</td>
</tr>
<tr>
<td>APOB</td>
<td>CCTATCTTATGTGTA</td>
<td>HuH-7</td>
<td>126</td>
<td>9</td>
<td>7.1</td>
</tr>
<tr>
<td>ATGL</td>
<td>CCGCCCAGCCCCACG</td>
<td>HUES 9</td>
<td>192</td>
<td>28</td>
<td>14.6</td>
</tr>
<tr>
<td>CIITA</td>
<td>CCCCTGTGCCTCTAC</td>
<td>BJ-RiPS</td>
<td>292</td>
<td>37</td>
<td>12.7</td>
</tr>
<tr>
<td>CELSR2</td>
<td>CCCTGAGGGTGTGTCAAT</td>
<td>HUES 1</td>
<td>506</td>
<td>18</td>
<td>3.5</td>
</tr>
<tr>
<td>CFTR</td>
<td>GGAGGCTCAACGAG</td>
<td>CF-RiPS</td>
<td>140</td>
<td>3</td>
<td>2.1</td>
</tr>
<tr>
<td>LINC00116</td>
<td>AGTTGTCCCGGTGTAGT</td>
<td>HUES 9</td>
<td>88</td>
<td>26</td>
<td>29.5</td>
</tr>
<tr>
<td>NLRC5</td>
<td>GGCTGCTCAACAAAG</td>
<td>BJ-RiPS</td>
<td>83</td>
<td>6</td>
<td>7.2</td>
</tr>
<tr>
<td>PLIN1</td>
<td>ACAGTTGGTGCAATTAC</td>
<td>HUES 9</td>
<td>293</td>
<td>70</td>
<td>23.8</td>
</tr>
<tr>
<td>PLIN1</td>
<td>ACAGTTGGTGCAATTAC</td>
<td>BJ-RiPS</td>
<td>439</td>
<td>29</td>
<td>6.6</td>
</tr>
<tr>
<td>SORT1 (exon 2)</td>
<td>TCAGTATCCTTGTC</td>
<td>HUES 1</td>
<td>576</td>
<td>128</td>
<td>22.2</td>
</tr>
<tr>
<td>SORT1 (exon 3)</td>
<td>TGGACAGTCCAAAGCT</td>
<td>HUES 9</td>
<td>192</td>
<td>21</td>
<td>10.9</td>
</tr>
<tr>
<td>TRIB1</td>
<td>CTCCGGCGTCGCCACC</td>
<td>HUES 9</td>
<td>169</td>
<td>41</td>
<td>24.2</td>
</tr>
<tr>
<td>TTN</td>
<td>ATGACAGAAGAGAA</td>
<td>BJ-RiPS</td>
<td>250</td>
<td>17</td>
<td>6.8</td>
</tr>
</tbody>
</table>
**APOB is Required for HCV Replication**

*APOB*, which encodes apolipoprotein B, the core protein of very-low-density lipoprotein (VLDL) and low-density lipoprotein (LDL) particles that transport cholesterol and triglycerides from the liver to other tissues via the bloodstream, has been suggested to play a critical role in hepatitis C virus (HCV) infection. In HCV models using cultured human HuH-7 hepatoma cells, RNA interference resulting in partial knockdown of *APOB* expression has been reported to reduce HCV secretion, albeit not HCV replication (Huang et al., 2007; Nahmias et al., 2008); the precise points of interaction of *APOB* with the HCV lifecycle remain to be defined. We sought to address this question by generating *APOB* knockout HuH-7 cells.

The human *APOB* gene encodes a 512 kDa protein termed apoB-100. We designed a TALEN pair targeting a site in exon 13 (*Figure 3-2A*); frameshift mutations at the site would generate truncated proteins about 12.5% of the size of apoB-100 (apoB-12.5). We transfected a clonal line of HuH-7 with high expression of CD81 (a co-receptor for HCV entry; HuH-7/CD81<sup>high</sup>) with the *APOB* TALEN pair. Following FACS with a co-translated fluorescent marker, replating of sorted cells at limiting dilution, and expansion of single clones, we found that of 126 screened clones, four clones had exon 13 frameshift mutations in both alleles (*Figure 3-2A*). Compared to wild-type controls from the same set of screened clones, *APOB* knockout cells had no detectable intracellular apoB protein, no secreted apoB mass in the media, and $<$3% *APOB* mRNA expression, consistent with nonsense-mediated mRNA decay (*Figures 3-2B, 3-2C, 3-2D*).

We infected *APOB*/− and wild-type cells with the tissue-culture-infectious HCV
strain JFH-1. The APOB−/− cells had significantly lower HCV RNA levels (74% reduction, \( P = 0.006 \)), with minimal detectable HCV core protein (Figures 3-2B and 3-2E). Reintroduction of apoB-100 protein into the APOB−/− cells by adding LDL particles to the media, allowing for cellular LDL uptake, resulted in partial restoration of HCV core protein levels, arguing that the HCV replication defect was the result of loss of APOB function rather than an off-target effect of the TALENs (i.e., mutagenesis at other sites in the genome) (Figure 3-2B). Together, these data suggest that apoB-100 is integral to the HCV viral lifecycle and that APOB-targeting therapeutics (e.g., mipomersen) may have efficacy in treating HCV-infected patients.
**Figure 3-2. APOB is Important for HCV Replication**

(A) Generation of APOB knockout HuH-7 clones with TALENs targeting exon 13. Boxes indicate the TALEN binding sites. Deletions, insertions, and duplications in the two alleles of each clone are indicated. The 26-bp insertion and 8-bp duplication (asterisk) in clone A are 5’-GAGTCGCTTCTCCGGGAGATAAGTCA-3’ and 5’-GACTGGCT-3’, respectively.

(B) Left panel, Western blot using whole cell lysates from two wild-type and four knockout HuH-7 cell lines (clones A–D). Right panel, Western blot from a wild-type clone and a knockout clone (clone A) infected with or without JFH-1 virus and incubated with or without LDL particles. The same wild-type clone and knockout clone (clone A) were used for all subsequent experiments.

(C) Left panel, apoB ELISAs performed on conditioned media from cells; values are normalized to level from wild-type clone. Right panel, APOB mRNA expression by qRT-PCR from whole cell lysates; expression is indicated as fold change of $2^{-\Delta\Delta C_t}$ with reference to 18S rRNA, normalized to level in wild-type clone.

(D) Immunocytochemistry for apoB.

(E) HCV RNA levels by qRT-PCR from clones infected with JFH-1 virus; expression is indicated as fold change of $2^{-\Delta\Delta C_t}$ with reference to GAPDH, normalized to level in wild-type clone.
Figure 3-2 (continued)

A

![Diagram showing targeting exon 13 in APOB with wild-type and clone A, B, C, D sequences.]

B

<table>
<thead>
<tr>
<th>HuH-7 cells</th>
<th>+/LDL</th>
<th>- JFH-1</th>
<th>+ JFH-1 virus</th>
</tr>
</thead>
<tbody>
<tr>
<td>wild-type APOB knockout</td>
<td>α-apoB</td>
<td>α-β-actin</td>
<td>α-HCV core</td>
</tr>
<tr>
<td>WT KO</td>
<td>WT KO</td>
<td>WT KO</td>
<td>WT KO</td>
</tr>
</tbody>
</table>

C

![Bar graph showing normalized level of apoB and APOB expression.]

D

<table>
<thead>
<tr>
<th>wild-type APOB</th>
<th>APOB knockout</th>
</tr>
</thead>
<tbody>
<tr>
<td>apoB</td>
<td></td>
</tr>
<tr>
<td>Hoechst</td>
<td></td>
</tr>
</tbody>
</table>

E

![Image showing HuH-7 cells with wild-type and APOB knockout under JFH-1 virus conditions.]

![Graph showing normalized level of HCV RNA.]

P<0.006
Isogenic Disease Models in hPSCs

We found that the karyotype of the HuH-7 cells was severely abnormal (Figure 3-3)—fortuitously, it harbored two APOB alleles, in contrast to SORT1, with at least five alleles—highlighting the disadvantages of cultured tumor cell lines for rigorous genetic studies. hPSCs offer several advantages: they can maintain stable genomes with normal karyotypes while propagated in culture (Figure 3-3), preserving correct gene dosage; they can be differentiated into a variety of cell types, extending studies beyond a single cell type; and they can yield human cell types that are not available as cultured cell lines, e.g., adipocytes and motor neurons.
Figure 3-3. Representative karyotypes of HuH-7/CD81\textsuperscript{high} cells, a wild-type HUES 9 clone, and a SORT1−/− HUES 9 clone (clone D from Fig. 2a).

The two HUES 9 clones have identical, normal karyotypes; the HuH-7 cells have severely abnormal karyotypes.
Figure 3-3 (continued)
These advantages are mitigated by the significant variability in differentiation capacity and phenotypic characteristics among different hPSC lines, particularly among iPSC lines. This variability is attributed to differences in genetic background, in epigenetic state, and in derivation of the cell lines and adaptation to culture, among other factors. In this variability lies the potential for confounding of any phenotypic differences observed among differentiated cell lines generated to serve as disease models or controls—a significant weakness of studies in which a few iPSC lines from patients are compared with a few iPSC lines from healthy individuals, as has been the case with most published studies to date, since any observed differences cannot be reliably attributed to the effects of disease mutations. We demonstrated this cell line-to-cell line variability by differentiating two hESC lines, HUES 1 and HUES 9, into hepatocyte-like cells (HLCs) using an adapted protocol (Si-Tayeb et al., 2010) (Figure 3-4). We found that there were significant differences in the amounts of apoB and albumin secreted by the two cell lines and retained in the media (Figure 3-5); when apoB mass was normalized to albumin mass, there was a two-fold difference between the two hESC lines ($P = 0.0001$).
Figure 3-4.
(A) Differentiation protocol of human pluripotent stem cells (hPSCs) into hepatocyte-like cells (HLCs).
(B) Differentiation with marker staining at various stages, with cultured human HepG2 cells for comparison.
(C) Relative expression of hepatocyte-specific genes in HLCs vs. hPSCs measured by qRT-PCR; expression is indicated as fold change of $2^{-\Delta\Delta Ct}$ with reference to HPRT, normalized to levels in hPSCs.
(D) Secreted apoB and apoA-I mass in media from HLCs vs. hPSCs by ELISA and/or Western blot analysis.
Figure 3-4 (continued)

A

Actvin A/PI3K inhibitor  BMP4/FGF2  HGF  OSM/HGF/Dex/HBM

Human pluripotent stem cell (hPSC)  Definitive endoderm (DE)  Hepatic endoderm (HE)  Immature hepatocyte (IMH)  Hepatocyte-like cell (HLC)

Step 1  Step 2  Step 3  Step 4

d0  d1  d5  d10  d15  d20-25

B

HepG2  hPSC  DE  HE  IMH  HLC

Oct4  Sox17  FoxA2  HNF4A  ALB

C

SERPINA1  ALB  APOB

relative expression/mRNA

HLC  hPSC  HLC  hPSC  HLC  hPSC

D

AFC  apoB ELISA

apoB ELISA

t-  apoA-I
Figure 3-5. Related to Figure 3-6

(A) Albumin and apoB mass measured by ELISA in media collected from wild-type HLCs differentiated from HUES 1 or HUES 9 cells, normalized to mean levels in HUES 1 HLCs. N = 6.

(B) ApoA-I and apoB mass measured by ELISA in media collected from wild-type and knockout HLCs (two clones each; A and B for HUES 1, D and E for HUES 9 from Figure 2), normalized to mean levels in wild-type HLCs. N = 3 for HUES 1, N = 6 for HUES 9.

(C) ApoA-I and apoB mass in media measured by ELISA and SORT1 mRNA expression measured by qRT-PCR from wild-type and knockout HUES 9 HLCs (two clones each; clones D and E from Figure 2) infected with SORT1- or GFP-expressing lentivirus. ELISA measurements are normalized to mean levels in wild-type HLCs. mRNA expression is indicated as fold change of $2^{-\Delta\Delta Ct}$ with reference to 18S rRNA, normalized to mean level in undifferentiated HUES 9 hESCs (not shown). N = 6.

(D) ANGPTL4, ANGPTL6, HGF, and FGF-19 mass in media measured by ELISA from clones in (C), normalized to mean levels in wild-type HLCs. N = 5.

(E) Lipid-related gene expression measured by qRT-PCR from clones in (C); expression is indicated as fold change of $2^{-\Delta\Delta Ct}$ with reference to 18S rRNA, normalized to mean level in undifferentiated HUES 9 hESCs (not shown). N = 6.

Error bars show s.e.m. from experiments with biological replicates. P values calculated with unpaired t test.
Figure 3-5 (continued)
Using genome editing to generate isogenic cell lines that differ only with respect to a single mutation of interest provides a superior study design, since the cell lines would have the same origin and would thus be matched in genetic background, epigenetic state, derivation and adaptation to culture, etc. This would minimize confounding of the experiment and allow for more confidence in concluding that any phenotypic differences between the cell lines are secondary to the mutation. For these reasons, our subsequent studies were all performed in genome-edited hESCs.

**SORT1 Mediates Diverse Cellular Functions in Hepatocytes, Adipocytes, and Neurons**

*SORT1* (encoding sortilin) was recently discovered by genome-wide association studies to regulate human blood LDL cholesterol levels and risk for coronary artery disease, via the modulation of the hepatic secretion of apoB-100-containing particles into the bloodstream; however, conflicting studies in humans and mice disagree about the direction of the effect of sortilin on apoB secretion (Musunuru et al., 2010; Kjolby et al., 2010). Human genetic studies have found that single nucleotide polymorphisms (SNPs) associated with increased hepatic *SORT1* expression are also associated with decreased blood LDL cholesterol levels (Kathiresan et al., 2008; Musunuru et al., 2010). Knockdown and overexpression of *Sort1* in mouse liver suggested that sortilin functions to decrease hepatocyte apoB secretion (Musunuru et al., 2010). In contrast, a study of *Sort1* knockout mice suggested that sortilin increases hepatocyte apoB secretion (Kjolby et al., 2010).

We targeted exon 2 in the hESC line HUES 1 and, in a single round of TALEN
targeting, generated three clones that were compound heterozygous for frameshift mutations (out of 576 clones screened) and confirmed that they lacked sortilin protein (Figures 3-6A and 3-6B). In parallel, we targeted exon 3 in the hESC line HUES 9 and obtained two knockout clones (out of 192 clones screened). We differentiated two SORT1–/– and two wild-type HUES 1 clones or two SORT1–/– and two wild-type HUES 9 clones into hepatocyte-like cells (HLCs) using an adapted protocol (Si-Tayeb et al., 2010) (Figure 3-4). Measuring the levels of apoB as well as albumin and apoA-I (reference controls) secreted from the HLCs and retained in the media, we found that knockout cells had significantly increased apoB mass (HUES 1: 117% increase in apoB/albumin ratio, \( P = 0.04 \); HUES 9: 65% increase in apoB/albumin ratio, \( P = 0.05 \)) (Figure 3-6C and Figure 3-5B).
Figure 3-6. SORT1 Reduces Hepatocyte Secreted ApoB Mass, is Important for Insulin-Responsive Glucose Transport in Adipocytes, and Mediates proBDNF-Induced Motor Neuron Death

(A) Generation of SORT1 knockout hPSC clones with TALENs targeting exon 2 or exon 3. Boxes indicate the TALEN binding sites. Deletions and insertions in the two alleles of each clone are indicated. The 17-bp insertion (asterisk) in clone B is 5'-TGCTATCTCAACCAGG-3'.

(B) Western blot for sortilin and qRT-PCR for SORT1 mRNA in wild-type and knockout HUES 1 clones (clones A–C); mRNA expression is indicated as fold change of $2^{-\Delta\Delta Ct}$ with reference to 18S rRNA, normalized to mean level in wild-type clones.

(C) Albumin and apoB mass measured by ELISA in media collected from wild-type and knockout HLCs (two clones each; A and B for HUES 1, D and E for HUES 9), normalized to mean levels in wild-type HLCs. N = 3 for HUES 1, N = 6 for HUES 9.

(D) Western blots of lysates and ELISAs in media from wild-type and knockout HLCs (one clone each, A for HUES 1; two clones each, D and E for HUES 9) infected with SORT1- or GFP-expressing lentivirus, normalized to mean levels in wild-type HLCs. N = 2 for HUES 1, N = 6 for HUES 9.

(E) Ratios of glucose uptake to total protein content in wild-type and knockout HUES 1 adipocytes (one clone each; clone A) infected with SORT1-expressing or control lentivirus and treated with or without insulin, all normalized to mean ratio in wild-type adipocytes without insulin. N = 6.

(F) Immunocytochemistry for TUJ1 and ISL-1 in wild-type and knockout HUES 9 motor neurons. Arrows indicate representative double-positive cells.

(G) Counts of wild-type and knockout HUES 9 (2 clones each, D and E) motor neurons (TUJ1/ISL-1 double-positive cells) treated with BDNF vs. proBDNF. N = 12. Error bars show s.e.m. from experiments with biological replicates. $P$ values calculated with unpaired $t$ test.
Figure 3-6 (continued)

A

<table>
<thead>
<tr>
<th>Clone</th>
<th>Targeting Exon 2</th>
<th>Targeting Exon 3</th>
<th>Wild-type HUES 9</th>
<th>Wild-type HUES 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone A</td>
<td>TCAAGAAGGCTGACAGTCCTCCAGGAGG</td>
<td>AGCTGACAGGAAGG</td>
<td>1bp del</td>
<td>3bp del</td>
</tr>
<tr>
<td>Clone B</td>
<td>TCTGACCTGGAGGAGGATGAG</td>
<td>AGCTGACAGGAAGG</td>
<td>17bp del</td>
<td>15bp del</td>
</tr>
<tr>
<td>Clone C</td>
<td>TCTGACCTGGAGGAGGATGAG</td>
<td>AGCTGACAGGAAGG</td>
<td>13bp del</td>
<td>15bp del</td>
</tr>
<tr>
<td>Clone D</td>
<td>TCTGACCTGGAGGAGGATGAG</td>
<td>AGCTGACAGGAAGG</td>
<td>13bp del</td>
<td>15bp del</td>
</tr>
<tr>
<td>Clone E</td>
<td>TCTGACCTGGAGGAGGATGAG</td>
<td>AGCTGACAGGAAGG</td>
<td>13bp del</td>
<td>15bp del</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Wild-type</th>
<th>SORT1-/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUES 1 hPSCs</td>
<td></td>
</tr>
<tr>
<td>α-sortilin</td>
<td></td>
</tr>
<tr>
<td>α-actinin</td>
<td></td>
</tr>
</tbody>
</table>

C

<table>
<thead>
<tr>
<th>HUES 1 HLCs</th>
<th>HUES 9 HLCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized level</td>
<td>Normalized level</td>
</tr>
<tr>
<td>Albumin mass</td>
<td>Apo8/albumin</td>
</tr>
<tr>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>1.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

D

<table>
<thead>
<tr>
<th>Wild-type</th>
<th>SORT1-/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUES 1 HLCs</td>
<td>HUES 9 HLCs</td>
</tr>
<tr>
<td>Normalized level</td>
<td>Normalized level</td>
</tr>
<tr>
<td>Apo8/albumin</td>
<td>Apo8/albumin</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

E

<table>
<thead>
<tr>
<th>HUES 1 white adipocytes</th>
<th>HUES 9 white adipocytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized level</td>
<td>Normalized level</td>
</tr>
<tr>
<td>Glucose uptake/protein</td>
<td>Glucose uptake/protein</td>
</tr>
<tr>
<td>0</td>
<td>1.5</td>
</tr>
<tr>
<td>1.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

F

<table>
<thead>
<tr>
<th>Wild-type</th>
<th>SORT1-/-</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProBDNF</td>
<td>ProBDNF</td>
</tr>
<tr>
<td>100μm</td>
<td>100μm</td>
</tr>
</tbody>
</table>

G

<table>
<thead>
<tr>
<th>HUES 9 neurons</th>
<th>Motor neuron %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized level</td>
<td>Normalized level</td>
</tr>
<tr>
<td>0.2</td>
<td>1.2</td>
</tr>
<tr>
<td>1.2</td>
<td>2.2</td>
</tr>
</tbody>
</table>
We infected knockout HUES 1 HLCs or HUES 9 HLCs with a lentivirus expressing the SORT1 cDNA or a control lentivirus and found that reconstitution of SORT1 to the levels observed in wild-type HUES 1 or HUES 9 HLCs resulted in normalization of the apoB mass (Figure 3-6 and Figure 3-5C), confirming that the observed differences in apoB mass are specific to SORT1 function and not the result of off-target effects. We found that secreted levels of additional hepatic proteins—ANGPTL4, ANGPTL6, HGF, and FGF-19—did not differ among the various experimental conditions (Figure 3-5D), nor did mRNA levels of APOB and other lipid-related genes such as HMGCR, LDLR, and SREBP1 (Figure 3-5E). Our data suggest that, in humans, sortilin acts in hepatocytes to reduce apoB-containing particle levels in the blood, resulting in lower cholesterol levels and reduced risk of coronary artery disease—consistent with human genetic studies (Kathiresan et al., 2008; Musunuru et al., 2010) and, notably, contradicting the results reported from Sort1 knockout mice (Kjolby et al., 2010).

SORT1 has also been suggested to play an important role in regulating blood glucose levels by modulating insulin-dependent translocation of the fat- and muscle-specific glucose transporter, Glut4, to the plasma membrane via the formation and transport of Glut4 storage vesicles, based on studies in cultured mouse 3T3-L1 cells (Shi and Kandror, 2005). We differentiated two SORT1−/− and two wild-type HUES 1 clones into white adipocytes using a recently published protocol (Ahfeldt et al., 2012), and we observed a substantial increase in glucose uptake in wild-type adipocytes upon treatment with insulin (63% increase, P = 0.009) but not in SORT1−/− adipocytes (Figure 3-7A). We infected the knockout adipocytes with a SORT1 or control lentivirus
and found that reconstitution of SORT1 restored insulin-responsive glucose uptake (60% increase, $P = 0.002$), confirming that the loss of insulin response in the knockout adipocytes was specific to SORT1 function and not the result of off-target effects (Figure 3-6E and Figure 3-7B). Thus, SORT1 appears to be critical for insulin-responsive glucose uptake in human adipocytes and may play a role in insulin sensitivity in humans.
Figure 3-7. Related to Figure 3-6

(A) Ratios of glucose uptake to total protein content in wild-type and knockout HUES 1 adipocytes (two clones each; clones A and B from Figure 2) treated with or without insulin, all normalized to mean ratio in wild-type adipocytes without insulin. N = 6.

(B) SORT1 and adipose-related gene expression measured by qRT-PCR from wild-type and knockout HUES 1 adipocytes (one clone each; clone A from Figure 2) infected with SORT1-expressing or control lentivirus; expression is indicated as fold change of $2^{-\Delta\Delta Ct}$ with reference to 18S rRNA, normalized to mean level in wild-type adipocytes. N = 6.

(C) Total cell numbers (determined as number of cells marked with DAPI), motor neuron numbers (determined as number of cells doubly positive for TUJ1 and ISL-1), and ratios of motor neuron number to total cell number upon differentiation of wild-type and knockout HUES 9 cells into neurons (two clones each; clones D and E from Figure 2). N = 12.

Error bars show s.e.m. from experiments with biological replicates. $P$ values calculated with unpaired t test.
Figure 3-7 (continued)

**A**
HUES 1 white adipocytes
- wild-type, – insulin
- wild-type, + insulin
- Sort1<−/−, – insulin
- Sort1<−/−, + insulin

**B**
- wild-type, control virus
- Sort1<−/−, control virus
- Sort1<−/−, Sort1 virus

**C**
HUES 9 neurons
- wild-type
- Sort1<−/−
Finally, *SORT1* has also been implicated in the viability and function of neurons (Nykjaer and Willnow, 2012). In motor neurons, sortilin has been found to regulate neuronal survival during a temporally and spatially specific period of programmed cell death. Specifically, induction of motor neuron cell death by the pro-form of brain-derived neurotrophic factor (proBDNF) has been reported to be dependent on the presence of sortilin (Teng et al., 2005; Taylor et al., 2011). We differentiated two *SORT1*−/− and two wild-type HUES 9 clones into TUJ1+/ISL-1+ motor neurons using an adapted protocol (Di Giorgio et al., 2008; Chambers et al., 2009) and observed that while both *SORT1*−/− and wild-type hPSCs generated similar numbers of motor neurons (Figure 3–7C), *SORT1*−/− motor neurons were refractory to proBDNF-mediated programmed cell death, whereas wild-type motor neurons exhibited a substantial reduction after three days of proBDNF treatment (23% reduction, *P* = 0.004) (Figures 3–6F and 3–6G). These data confirm that *SORT1* is required for proBDNF-induced programmed cell death in human motor neurons.

**TALENs Exhibit Minimal Off-Target Effects**

As the extent of off-target effects of TALENs (i.e., mutagenesis at other sites in the genome) in hPSCs remains to be defined, we performed whole-genome sequencing of six cell lines: the parental HUES 1 cell line (clone X); the three *SORT1* knockout HUES 1 clones (A–C in Figure 3–6); a control HUES 1 clone that had been grown in parallel with the *SORT1* knockout clones (i.e., had been exposed to the *SORT1* exon 2 TALEN pair but retained two wild-type alleles; clone W); and a clone that had been successfully targeted in the CELSR2 gene with a different TALEN pair (clone Y).
Because the whole-genome sequencing was performed at low coverage (6-12 fold coverage on average), it was not possible to perform de novo genome assembly and ascertain all sequence variation among the genomes. Instead, we used the sequencing data to interrogate the sites in the genome at which one or the other TALEN of the SORT1 exon 2 TALEN pair would be most likely to bind and, thus, be most likely to induce an off-target sequence change. Because TALENs virtually always induce indels by NHEJ, rather than single nucleotide variants (SNVs), we focused our search on off-target indels.

Almost 40,000 genomic sites were identified to vary at most by two inserted bases, deleted bases, or substituted bases from one or the other SORT1 exon 2 TALEN binding site (just 21 genomic sites were 100% identical); we screened all of these sites for evidence of nearby indels. Nine sites with potential indels (Table 3-2) passed our criteria (see Experimental Procedures). The site with the strongest evidence for targeting by the TALENs was the on-target site in SORT1 exon 2, with indels detected in clones A, B, and C, as expected. In each of the other eight sites—all of which represent degenerate TALEN binding sites with two differences from the SORT1 exon 2 sites—the candidate indel appeared in only one clone, and the indel was in proximity to just one potential TALEN binding site; given that TALENs can only induce double-strand breaks as a dimer, it is unlikely that any of these indels resulted from TALEN activity. At six of the eight degenerate TALEN binding sites, the indel lay more than 60 bp away, making it even less likely that the TALENs induced the indel (Table 3-2). Thus, although we are not able to rule out TALEN off-target effects, we conclude that indels appear rarely at the most likely degenerate TALEN binding sites in clones in which on-target
indels have successfully been introduced.
Table 3-2. On-Target and Potential Off-Target Effects of TALENs Detected by Whole-Genome Sequencing

<table>
<thead>
<tr>
<th>position</th>
<th>indel length</th>
<th>reads with indels (out of total reads) in clones</th>
<th>predicted TALEN binding site</th>
<th>predicted TALEN binding site</th>
<th>distance from binding site (bp)</th>
<th>sequences surrounding indels</th>
</tr>
</thead>
<tbody>
<tr>
<td>on-target</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr1:109912174</td>
<td>–8</td>
<td>8 (14 reads)</td>
<td>chr1:109912157–109912171</td>
<td>chr1:109912187–109912201</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>chr1:109912175</td>
<td>–13</td>
<td>2 (4 reads)</td>
<td>chr1:109912157–109912171</td>
<td>chr1:109912187–109912201</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>chr1:109912177</td>
<td>–5</td>
<td>2 (9 reads)</td>
<td>chr1:109912157–109912171</td>
<td>chr1:109912187–109912201</td>
<td>6</td>
<td>refer to Figure 2</td>
</tr>
<tr>
<td>chr1:109912178</td>
<td>–13</td>
<td>6 (9 reads)</td>
<td>chr1:109912157–109912171</td>
<td>chr1:109912187–109912201</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>chr1:109912178</td>
<td>–1</td>
<td>4 (14 reads)</td>
<td>chr1:109912157–109912171</td>
<td>chr1:109912187–109912201</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>off-target</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2:38501598</td>
<td>–2</td>
<td>2 (12 reads)</td>
<td>chr2:38501866–38501898</td>
<td>N/A</td>
<td>88</td>
<td>GGACACAGTATCTTTAGGAGACCA</td>
</tr>
<tr>
<td>chr1:14107489</td>
<td>–1</td>
<td>2 (15 reads)</td>
<td>chr1:14107517–14107585</td>
<td>N/A</td>
<td>83</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
<tr>
<td>chr6:1521927</td>
<td>–1</td>
<td>2 (18 reads)</td>
<td>chr6:1521902–1521916</td>
<td>N/A</td>
<td>11</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
<tr>
<td>chr9:35550072</td>
<td>–2</td>
<td>2 (13 reads)</td>
<td>chr9:35550133–35550147</td>
<td>N/A</td>
<td>61</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
<tr>
<td>chr13:47901074</td>
<td>–1</td>
<td>2 (13 reads)</td>
<td>chr13:47901075–47901080</td>
<td>N/A</td>
<td>68</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
<tr>
<td>chr15:89922403</td>
<td>–1</td>
<td>2 (9 reads)</td>
<td>chr15:89922404–89922505</td>
<td>N/A</td>
<td>within</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
<tr>
<td>chr17:29982550</td>
<td>–1</td>
<td>2 (15 reads)</td>
<td>chr17:29982445–29982460</td>
<td>N/A</td>
<td>89</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
<tr>
<td>chr21:11062546</td>
<td>–4</td>
<td>3 (19 reads)</td>
<td>chr21:11062641–11062653</td>
<td>N/A</td>
<td>95</td>
<td>GGACACAGTACCTGGGGAAACCA</td>
</tr>
</tbody>
</table>
DISCUSSION

With our studies, we have used human model systems to generate strong evidence that apoB-100 is critical for HCV replication in human hepatocytes; that sortilin reduces apoB secretion by human hepatocytes, facilitates insulin-mediated glucose uptake by human adipocytes, and mediates motor neuron survival; that AKT2 E17K is a gain-of-function mutation that leads to reduced glucose production in human hepatocytes and increased triglyceride content in human adipocytes; and that PLIN1 frameshift mutations increase basal lipolysis in human adipocytes. More generally, these findings highlight the various types of studies to which genome editing in human cells may be applied to obtain novel biological insights.

With the emerging use of hPSCs to generate cell- and tissue-based models of disease, the ability to rapidly and efficiently modify genes in human cells using TALENs should become an indispensible tool for biological studies. For both SORT1 and AKT2, once we had obtained mutant clones we were able to successfully and simultaneously interrogate phenotypes in multiple somatic cell types, which previously has only been possible using primary tissues from transgenic animal models. While hPSC-based studies to date have largely been limited to “cells in a dish,” thereby restricting analyses to cell-autonomous phenotypes, we anticipate that incorporation of hPSC-derived cells into chimeric animal models will allow for interrogation of the effects of human genetic variation in whole-animal models.

Our studies suggest that TALENs incur a low burden of off-target effects. Nonetheless, the ease and rapidity of TALEN-mediated genome editing allows for rigorous study designs that can alleviate any concerns about off-target effects or other
potential confounding factors. As we have demonstrated with SORT1, it is straightforward to (1) generate multiple distinct mutant cell lines with each TALEN pair, (2) use distinct TALEN pairs to target different sites in a gene, (3) generate mutant clones in different cell lines with different genetic backgrounds, and (4) perform reconstitution experiments in knockout clones. Having used all of these approaches, we are able to conclude with great confidence that the observed hepatocyte and adipocyte phenotypes are indeed related to SORT1 function. We suggest that using at least one of these approaches will become de rigeur for future genetic studies.

Finally, we note that with the ability to use TALENs to readily insert specific gene variants into cells, the current enthusiasm for the generation of iPSC lines from patients with genetic disorders may shift instead to the use of genome editing to engineer isogenic cell lines with and without disease mutations. The time it takes to recruit a patient for the donation of tissue from which to make iPSCs (assuming such a patient is readily accessible, which may not be the case for rare disorders), to perform reprogramming to derive iPSC clones, to perform quality control to identify clones that are pluripotent and that will readily differentiate into the desired cell type, by use of a “scorecard” (Bock et al., 2011) or other methods, and then to undertake differentiation and phenotypic studies—in the absence of isogenic control cell lines—is a minimum of six months and usually longer. Within a shorter timeframe, we have found it to be quite feasible to use TALENs to edit a well-characterized and pre-validated (with respect to differentiation capacity) hPSC line and yield both mutant cell lines and isogenic control cell lines—allowing for a more rigorous study design—and to undertake differentiation and phenotypic studies, without any need for patient contact. Furthermore, genome
editing potentially allows for the interrogation of a large number of mutations, such as those now emerging from next-generation sequencing studies of human populations, on a single genetic background. Such studies will represent a significant advance in our ability to dissect genotype-phenotype relationships and thereby better elucidate human biology and disease.
METHODS

TALEN construction

TALEN genomic binding sites were chosen to be 15 bp in length or, in a few cases, 13 bp in length such that the target sequence between the two binding sites was between 14 and 18 bp in length; each binding site was anchored by a preceding T base in position “0” as has been shown to be optimal for naturally occurring TAL proteins (Moscou and Bogdanove, 2009; Boch et al., 2009). A library of 832 tetramer or trimer TAL repeats were constructed using methods based on the PCR-based protocol of Zhang et al. (2011); these multimers were designed to have complementary sticky ends when digested out of library plasmids with the type IIIs restriction enzyme BsmBI. As outlined in Figure 3-1, multimers were assembled into an array and subcloned into a full-length TALEN harboring, in order: a N-terminal FLAG tag, a nuclear localization signal, the N-terminal portion of the TALE PthXo1 from the rice pathogen X. oryzae pv. Oryzae (a kind gift of Dr. Daniel Voytas, University of Minnesota) lacking the first 176 amino acids (after Miller et al., 2011), the engineered TAL repeat array, the following 63 amino acids from the corresponding C-terminal portion of PthXo1 (after Miller et al., 2011), and one of two enhanced FokI domains. The FokI domains used were obligate heterodimers with both the Sharkey (Guo et al., 2010) and ELD:KKR (Doyon et al., 2011) mutations to enhance cleavage activity, engineered by PCR. Each TALEN was in a plasmid with the CAG promoter for optimal expression in hPSCs, with the TALEN being coexpressed with a fluorescent marker [enhanced green fluorescent protein (EGFP) or turbo red fluorescent protein (tRFP; Evrogen)] via an intervening viral 2A sequence. Full plasmid sequences are available upon request.
Cell culture, transfection/electroporation, and sorting

**HuH-7 cells.** HuH-7/CD81\textsuperscript{high} cells were grown in adherent culture in DMEM High Glucose containing glutamine and pyruvate (Invitrogen) and supplemented with 10% FBS and penicillin/streptomycin. Transfection of the plasmids expressing the APOB TALEN pair into HuH-7 cells was performed using Fugene 6 (Roche) in 10-cm tissue culture plates according to manufacturer instructions.

**hESC lines.** HUES 1 and HUES 9 cells (Cowan et al., 2004) were grown in feeder-free adherent culture in chemically defined mTeSR1 (STEMCELL Technologies) supplemented with penicillin/streptomycin on plates pre-coated with Geltrex matrix (Invitrogen). The cells were disassociated into single cells with Accutase (Invitrogen), and 10 million cells were electroporated with 50 µg of the TALEN pair (25 µg of each plasmid), or with a mix of 30 µg of the TALEN pair (15 µg of each plasmid) and 30 µg of the ssODN (5’-CAGGA AGTAC CGTGG CCTCC AGGTC TTGAT GTACT TACCT GAAAT GAGGC AGGAA GGGAG GGAGA GA-3’), in a single cuvette and replated as previously described (Schinzel et al., 2011). The cells were collected from the culture plates 48 hours post-transfection or post-electroporation by trypsin or Accutase treatment, respectively, and resuspended in PBS. Cells expressing green and/or red fluorescent markers were collected by FACS (FACSaria II; BD Biosciences) and replated on 10-cm tissue culture plates at 15,000 cells/plate to allow for recovery in growth media.

Isolation of targeted clonal cell populations

Post-FACS, the cells were allowed to recover for 7-10 days, after which single
colonies were manually picked and dispersed and replated individually to wells of 96-well plates. Colonies were allowed to grow to near confluence over the next 7 days, at which point they were split using trypsin (for HuH-7 cells) or Accutase (for hESCs) and replica-plated to create a working stock and a frozen stock. The working stock was grown to confluence. Genomic DNA was extracted in 96-well format from working stocks in lysis buffer (10 mM Tris pH 7.5, 10 mM EDTA, 10 mM NaCl, 0.5% Sarcosyl) containing proteinase K at 56°C overnight in a humidified chamber. Genomic DNA was precipitated by addition of 95% ethanol containing 75 mM NaCl for 1 hr at room temperature. The DNA was then washed 2 times in 70% ethanol, allowed to dry at room temperature, and then resuspended in nuclease-free water.

Genotyping at the TALEN target site was then performed for each sample by PCR amplification (94°C 30 sec; 56°C 30 sec; 68°C 30 sec) using FastStart Taq (Roche) and a primer pair designed to yield small amplicons (~150-200 bp) around the target site. Amplicons were subjected to electrophoresis on 2.5% agarose gels to discriminate clones with indels; for potentially positive clones, PCR amplicons were subcloned using the TOPO TA Cloning Kit (Invitrogen) and subjected to numerous sequence reads to confirm the presence of mutant alleles. Clones with confirmed compound heterozygous mutant alleles were retrieved from the frozen stocks and expanded for further experiments. When no compound heterozygous clones were identified, a heterozygous clone with one mutant allele was expanded and subjected to a second round of TALEN targeting.
Differentiation of hPSCs into HLCs

Following the protocol of Si-Tayeb et al. (2010), we used feeder cell-free, virus-free differentiation conditions entailing the addition of a variety of growth factors and chemicals to the growth media (Figure 3-4). We incubated cells in (1) RPMI-B27 (RPMI-1640 from Sigma; B27 supplement from Invitrogen) medium supplemented with recombinant activin A (100 ng/mL; PeproTech) and LY-294002 (5 µM; Promega), a phosphatidylinositol 3-kinase (PI3K) chemical inhibitor, for 4 days to obtain definitive endoderm; (2) RPMI-B27 supplemented with the growth factors BMP4 (20 ng/mL; PeproTech) and FGF2 (5 ng/mL; Millipore) and 0.5% DMSO for 5 days, yielding hepatoblasts; (3) RPMI-B27 supplemented with the growth factor HGF (20 ng/mL; PeproTech) and 0.5% DMSO for 5 days, yielding immature hepatocytes; and (4) HCM Hepatocyte Culture Medium (Lonza) supplemented with HGF (20 ng/mL), Oncostatin M (20 ng/mL; PeproTech), dexamethasone (100 nM; Sigma), and 0.5% DMSO for 10 days, yielding mature HLCs. Infection with lentiviruses expressing SORT1 or GFP was performed by standard methods at the end of the 3rd stage (day 15) for 2 hours, followed by the 4th stage of differentiation.

Differentiation of hPSCs into white adipocytes

Differentiation of white adipocytes from HUES 1 and HUES 9 was performed as previously described (Ahfeldt et al., 2012). In brief, cells were grown as embryoid bodies, replated, serially passaged to obtain mesenchymal progenitor cells (MPCs), and programmed with inducible PPARG2 expression to obtain white adipocytes. For SORT1 reconstitution experiments, SORT1−/− MPCs were either co-infected with lentiviruses
expressing *PPARG2*, rtTA, and *SORT1* or, for control conditions, with lentiviruses expressing *PPARG2* and rtTA only, followed by adipocyte differentiation.

**Differentiation of hPSCs into motor neurons**

Differentiation of motor neurons from HUES 9 was performed as previously described (Di Giorgio et al., 2008; Chambers et al., 2009). In brief, cells were grown as embryoid bodies in DMEM/F12 containing 2% B27 and 1% N2 (Life Technologies). The embryoid bodies were treated with Dorsomorphine and SB431542 (Stemgent) to induce a neural lineage and, additionally, with retinoic acid (Sigma) and Smoothened Agonist (Millipore). After 10 days of embryoid body formation, colonies were plated on polyornithine and laminin (PO/LAM)-coated plates to differentiate neural progenitors to motor neurons. After 30 days of differentiation, the neural culture containing motor neurons was dissociated to single cells with Papain solution (Worthington) and plated on PO/LAM-coated 384-well plates in Neurobasal-containing 2% B27 and 1% N2 (Life Technologies). After 5 days of pre-incubation, cells were treated with 10 ng/mL of GDNF and BDNF (R&D Systems) or proBDNF (ProSpecBio). After 3 days of treatment, cells were fixed with 4% PFA (Electron Microscopy Sciences) and subjected to immunocytochemistry.

**Enzyme-linked immunosorbent assays (ELISAs)**

For HuH-7 clones, cells were seeded at 50,000 cells/well in 24-well cell culture plates and grown to 80% confluency. Growth medium was added fresh at time = 0 hr and collected at time = 8 hr. The secreted apoB mass was quantitated using an apoB
ELISA kit (Mabtech) according to manufacturer instructions. The measurements were expressed as absorbance at 450 nm, adjusted for the background absorbance level.

For HLCs, after 25 days of differentiation, medium was added fresh at time = 0 hr and collected at time = 16 hr, and ELISAs were performed according to manufacturer instructions for apoB and apoA-I (Mabtech); albumin (Bethyl Laboratories); and ANGPTL4, ANGPTL6, HGF, and FGF-19 (Millipore). For adipocytes, after 21 days of differentiation, medium was added fresh at time = 0 hr and collected at time = 12 hr, and ELISAs were performed according to manufacturer instructions for adiponectin, MCP-1, IL-8, and PAI-1 (Millipore).

**Glucose production and uptake**

Glucose production was measured using a protocol adapted from Hagiwara et al. (2012). Briefly, after 25 days of differentiation, HLCs were serum starved overnight, followed by culturing in low-glucose DMEM containing dexamethasone (100 nM; Sigma) in the presence or absence of insulin (100 nM; Sigma) for 16 hrs. Cells were then incubated with glucose production solution containing sodium pyruvate (2 mM; Sigma), lactate (20 mM; Sigma), dexamethasone (100 nM), and forskolin (10 µM; Sigma) with or without insulin (100 nM). After 12 hr of incubation, the glucose concentration in the media was measured using the Amplex Red Glucose/Glucose Oxidase Assay Kit (Sigma) according to manufacturer instructions. Albumin mass in the media was measured by ELISA (Bethyl Fisher).

Glucose uptake was measured as previously described (Ahfeldt et al., 2012). Briefly, white adipocytes after 21 days of differentiation were starved overnight in DMEM
supplemented with 0.2% BSA and then incubated in KRH buffer (121 mM NaCl, 4.9 mM KCl, 1.2 mM MgSO₄, 0.33 mM CaCl₂ and 12 mM HEPES, at pH 7.4) for another 5 hours. Glucose uptake was measured by incubating cells with 0.5 µCi mL⁻¹ 2-deoxy-D-[³H]glucose (PerkinElmer) for 5 min after insulin (100 nM) stimulation for 10 min at 37 °C. After three washes with cold PBS, cells were lysed with 0.1% Triton X-100 solution and subjected to scintillation counting. CPM values were normalized to total protein content measured by the Bradford assay (Bio-Rad).

**Triglyceride measurement and lipolysis assay**

Cell triglyceride content was measured using Infinity Triglycerides solution (VWR International). White adipocytes after 16 days of differentiation were lysed using 0.1% SDS solution; 30 µL of cell lysate was mixed with 150 µL Infinity Triglycerides solution, followed by incubation at 37°C for 5 min. Absorbance at 525 nm was measured and normalized to total protein content measured by the Bradford assay (Bio-Rad).

To measure lipolysis activity, white adipocytes after 24 days of differentiation were starved in DMEM containing 1% FBS for 1 hour, followed by incubation in Hank’s Balanced Salt Solution (HBSS) with 2% fatty acid-free BSA alone or with 10 µM forskolin for 1 hour. The culture media was collected for glycerol measurement using the Free Glycerol reagent (Sigma). Absorbance at 540 nm was measured and normalized to total protein content measured by the Bradford assay (Bio-Rad) or total triglyceride content.
**Immunocytochemistry**

HuH-7 clones were plated on glass coverslips in a 24-well plate and allowed to adhere for 48 hr in standard growth medium. The cells were then fixed for 10 min in 4% formaldehyde, washed extensively with PBS, permeabilized with 0.1% Triton X-100 for 10 min, and blocked for 1 hr in 3% BSA in PBS. The cells were then incubated with a 1:1000 dilution of mouse monoclonal anti-apoB (C1.4) antibody (sc-13538; Santa Cruz) at 4°C overnight, washed with PBS, incubated for 1 hr at room temperature with Alexa 488-conjugated anti-mouse secondary antibody (1:500 dilution) (Invitrogen). Hoechst (1:5000 dilution) (Invitrogen) was used for nuclear staining. Cells were then mounted with an anti-fade mounting medium (Vector Laboratories), and immunofluorescence was visualized by confocal microscopy (LSM 710; Zeiss).

**Quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR)**

mRNA was extracted from cells using TRIZOL (Invitrogen) and purified using the RNeasy Mini Kit (QIAGEN). cDNA was generated using the Superscript III RT Kit (Invitrogen) and reverse priming with a 1:1 mixture of random hexamers and oligo-dT. Gene expression was measured in isolates using SYBR Green (Applied Biosystems) and oligonucleotides with the following sequences: APOB, 5’-ACCAGCACAGACCATTTCAG-3’ and 5’- GCGTAGAGACCCATCACATG-3’; ADIPOQ, 5’-GATGAAAGTCCTGCTTTGGAAGG-3’ and 5’-CAGCACTTAGAGATGGAGTTGG-3’; G6PC, 5’-TTCTACGTCTTTGCTCTTCTGC-3’ and 5’-AACACCGAAGACTCCACATC-3’; HMGCR, 5’-CTTTGCATGGCTCTTGAACAC-3’ and 5’-CTTTAATGGAAGCAAGTGCTCC-3’; HSL, 5’-CTCAGTGCTGCTCTCAAGTG-3’ and 5’-
CACCCAGGCGGAAGTCTC-3'; **LDLR**, 5'-CACATTTGCCACAACCAGG-3' and 5'-TCTTTGAATAAAAACAAGGGCGG-3'; **PCK1**, 5'-AGTGGAGCTCAGAGGATGG-3' and 5'-GCCTCCAAAGATAATGCCTTC-3'; **PLIN1** (N-terminus), 5'-CCCCCCTGAAAGATTGCTTCT-3' and 5'-GGAACGCTGATGCTGTTTCTG-3'; **PLIN1** (C-terminus), 5'-GCCATGTCCCTATCAGATGC-3' and 5'–GTTGTCGATGTCCCGGAATT-3'; **PPARG2**, 5'-GCAGGAGATCTACAAGGACTTG-3' and 5'-CCCTCAGAATAGTGCAACTGG-3'; **SORT1**, 5'-CAGGAGTGCTCATTGTGAAGA-3' and 5'-TTTATTAGTGTGGGAGGTCTGTG-3'; **SREBP1**, 5'-TGCTTTCTTTTGCTACCAGTG-3' and 5'-GCCTGGGCAAAGTCTCC-3'; 18S rRNA, 5'-CGGC\_TACCACATCCAAGGAA-3' and 5'-GCTGGAATTACCGCGGCT-3'. The last was used as a reference control. Biological replicates were performed, and reactions were carried out using a ViiA 7 System (Applied Biosystems) using standard real-time PCR conditions. Relative quantitation of mRNA levels was performed using the 2^-ΔΔCt^ method.

**Western blot analysis**

Western blot analyses were performed using standard methods with antibodies against sortilin (AF3154; R&D Systems), albumin (A80-129A; Bethyl Fisher), beta-actin (A5316; Sigma), AKT2 (L79B2; Cell Signaling Technology), AKT1 (C73H10; Cell Signaling Technology), apoB (sc-13538 from Santa Cruz; ab20737 from Abcam), and perilipin (for N-terminus recognition, GP29 from Progen; for C-terminus recognition, 3470 from Cell Signaling Technology), as well as anti-HCV core 6G7 monoclonal antibody (a kind gift of Drs. Harry Greenberg and Xiaosong He, Stanford University).
HCV infection of HuH-7 cells

HuH-7 cells were grown in DMEM supplemented with 10% FBS and penicillin/streptomycin. Cells were seeded at 75,000 cells/well in 12-well cell culture plates with 1 mL of media in each well and allowed to attach overnight. The growth media was then replaced with 250 µL of media containing the full-length, genotype 2b, tissue culture-infectious JFH-1 hepatitis C virus for 6 hrs; mock-infected wells were replaced with 250 µL of virus-free media. After 6 hours, the virus media was removed and replaced with fresh DMEM supplemented with 10% lipid-depleted FBS (Gemini Bio-Products) and penicillin-streptomycin. To restore intracellular apoB-100, LDL (Sigma) was added to the media to give a final LDL concentration of 9.68 mg/mL.

Protein and RNA were isolated from cell lysates 72 hrs following infection. RNA was isolated using the RNeasy Mini Kit (QIAGEN) and reverse transcribed into cDNA using the GeneAmp RNA PCR Kit (Applied Biosystems) according to manufacturer instructions. HCV RNA levels were quantitated using the DyNAmo HS SYBR Green qPCR kit (Finnzyme). The following primers were used: JFH-1 RNA, 5’-CTGTCTTCACGCAGAAAGCG-3’ and 5’-TCGCAACCCAACGCTACTCG-3’; and GAPDH, 5’-ACCTTCCCCCATGGTGCTGA-3’ and 5’-GCTCCTCCTGTTCGACAGTCA-3’. Relative quantitation of viral RNA levels was performed using the $2^{-\Delta\Delta Ct}$ method with GAPDH as the reference control.

Whole-genome sequence analysis

Genomic DNA from six hESC clones was extracted using the DNeasy Tissue Kit (QIAGEN), subjected to quality assessment, and sequenced to between 6× and 12×
haploid coverage on Illumina GA-II or HiSeq sequencers as paired-end 101-nucleotide reads as previously described (Stransky et al., 2011). Pre-processing, alignment to the human reference genome (assembly 19), and post-filtering of Illumina sequence data was performed using the Broad Institute Sequencing Platform’s “Picard” pipeline and Genome Analysis Toolkit (McKenna et al., 2010), including table recalibration (TableRecalibration), BWA alignment, and local realignment near indels (IndelRealigner). The SAMtools (Li et al., 2009) command mpileup was used to generate pileups for the regions immediately upstream and downstream of sites that are highly similar (within two insertions, deletions, or substitutions) to the predicted TALEN binding sequences in SORT1 exon 2. We identified these sites (39084 in total) in the human reference genome using Mathematica 8.0 (Wolfram) and liftOver (Hinrichs et al., 2006), and we recorded every potential indel in all the reads that were mapped within 100 bp upstream or downstream of any of the sites. Indels were further considered if they appeared in >1 independent reads in regions of good coverage (≥4×) in at least one of the clones that was exposed to the SORT1 exon 2 TALENs (A, B, C, and W), appeared in ≥10% of the total available sequence reads in the clone, and did not appear in either of the two control clones (clones X and Y). Indels meeting these criteria were manually inspected using the SAMtools Text Alignment Viewer, and any potential indel appearing at a site of highly repetitive DNA was eliminated as being likely to have resulted from sequencing artifact.
ACKNOWLEDGEMENTS

This work was supported in part by a Roche Postdoc Fellowship (Q.D.); the Sternlicht Director's Fund Award for Graduate Students from the Harvard Stem Cell Institute (D.T.P.); grants T32-HL007604 (R.M.G.), T32-DK007191 (D.L.M.), K08-DK088951 (L.F.P.), R00-HL098364 (K.M.), and U01-HL107440 (C.A.C.) from the United States National Institutes of Health (NIH); the Broad Institute’s Lawrence H. Summers Fellowship (K.M.); and Harvard University (L.L.R., K.M., C.A.C.). A.M. is a full-time employee of Roche Pharmaceuticals; the other authors report no relevant conflicts of interest. We thank David Altshuler, Noel Burtt, Guillermo del Angel, Mark DePristo, Emmanuel Figueroa, Stacey Gabriel, Namrata Gupta, J. Keith Joung, Adam Kaplan, Heng Li, Elyse Macksoud, Khalid Shakir, Alanna Strong, Kristin Thompson, Jayaraj Rajagopal, Stephanie Regan, Jennifer Shay, and Yulei Xia for assistance and suggestions.
REFERENCES


Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove,


Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li,


Tansey, J.T., Sztalryd, C., Gruia-Gray, J., Roush, D.L., Zee, J.V., Gavrilova, O.,


Chapter 4: *In Vitro* Model of ANGPTL3 Deficiency in Familial Combined Hypolipidemia Using Differentiated Human Hepatocyte-like Cells
SUMMARY

Recent human genetics studies have implicated ANGPTL3, which codes for angiopoietin-like 3 (ANGPTL3), as a regulator of plasma triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C) levels—risk factors for cardiovascular disease. While it has been shown that ANGPTL3, which is synthesized and secreted by the liver, inhibits lipoprotein lipase and endothelial lipase, increasing plasma TG and HDL-C levels in rodents, the mechanism by which ANGPTL3 modulates LDL-C levels in humans remains unclear. Guided by the observation that individuals with rare nonsense mutations in ANGPTL3 have decreased rates of very-low-density lipoprotein (VLDL) production, we sought to evaluate the effects of ANGPTL3 deficiency in human hepatocytes. We used TALEN and CRISPR/Cas genome editing to generate isogenic ANGPTL3 knockout and control human embryonic stem (hES) cell lines and differentiated these cell lines into hepatocyte-like cells for characterization. The studies presented in this chapter demonstrate the potential utility of hPSCs and genome editing for investigating the effects of human genetic variation in vitro. Yet, despite the use of this rigorous experimental approach, we were ultimately unable to gain new insight into the in vivo function of ANGPTL3, due to the fact that in vitro assays yielded ambiguous results. This outcome serves as a cautionary tale, illustrating challenges that may be commonly encountered in hPSC disease modeling studies.
INTRODUCTION

Despite the use of LDL-C lowering medications, including statins, myocardial infarction (MI) remains a leading cause of death\(^1\). Therefore, there is a critical need for additional medications to prevent MI. Plasma lipid concentrations are heritable risk factors for cardiovascular disease (CVD) in humans and aberrant lipid metabolism is a key component of the metabolic syndrome. Studies employing Mendelian randomization, a genetics approach used to assess the causality of disease-associated traits, have confirmed a causal relationship between LDL-C and MI\(^2\). As described in Chapter 1 of this thesis, human genetic studies have uncovered numerous genes and genomic loci associated with diseases and risk factors including CVD and blood lipid levels\(^11-12\). These findings represent promising opportunities to uncover molecular pathways that underlie disease pathogenesis, enabling the development of new therapeutic strategies.

Coding sequence mutations in the human \textit{ANGPTL3} gene, which encodes for angiopoietin-like 3, are associated with altered plasma lipid levels as well as plasma glucose concentration and insulin sensitivity\(^7-8\). Exome sequencing of two individuals from a family with extremely low triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), and LDL-C levels led to the identification of \textit{ANGPTL3} as the causal gene for this monogenic disorder, termed familial combined hypolipidemia (FCH; OMIM \#605019; \textbf{Figure 4-1})\(^9\). In addition, common DNA variants near \textit{ANGPTL3} are associated with TG, LDL-C, and total cholesterol levels\(^3,13\).
Figure 4-1. Nonsense mutations in ANGPTL3 (S17X, E129X) underlie a monogenic disorder of lipid metabolism characterized by extremely low TG, HDL-C, and LDL-C levels (Figure adapted from Musunuru K. NEJM 2010). Pedigree depicts “original” FCH family, showing genotypes (represented by symbol shading pattern), ages, and plasma lipid concentrations of family members. Exome sequencing, which revealed mutations in the ANGPTL3 gene, was performed on individuals II-4 and II-5. The panels in the upper right are the lipid levels in family members grouped by genotype after adjustment for age and sex.
ANGPTL3 is a member of the angiopoietin-like (ANGPTL) family of secreted proteins, which are involved in lipid trafficking and metabolism\textsuperscript{14-16}. Studies in mouse have shown that ANGPTL3 is synthesized in the liver and secreted into the bloodstream, where it inhibits the activity of intravascular lipases, in particular lipoprotein lipase (LPL) and endothelial lipase (EL, encoded by the \textit{LIPG} gene), thereby increasing plasma TG and HDL-C levels (\textbf{Figure 4-2})\textsuperscript{10}. However, the mechanism by which ANGPTL3 modulates LDL-C levels in humans remains unclear. This highlights the challenge of modeling certain human disease processes using mouse models. Fundamental differences in physiology between humans and rodents make it particularly difficult to study cardiovascular disease using mouse models; mice transport very little cholesterol in LDL particles and do not naturally develop CAD and MI.
Figure 4-2. Schematic representation of the physiologic functions of secreted ANGPTL3 protein, highlighting both its established and hypothesized roles as a regulator of human lipoprotein metabolism

Studies in mouse have shown that ANGPTL3 is synthesized in the liver and secreted into the bloodstream where it inhibits lipoprotein lipase (LPL) and endothelial lipase (EL, encoded by the LIPG gene), increasing plasma TG and HDL-C levels\(^10\). However, the mechanism by which ANGPTL3 modulates LDL-C levels in humans remains unclear.
Human genetic evidence suggests that ANGPTL3 regulates LDL-C by a mechanism distinct from its inhibition of LPL or EL. Common variants near LPL and LIPG are associated with TG and HDL-C respectively, but not LDL-C\textsuperscript{3,13}. Notably, individuals with nonsense mutations in ANGPTL3 have decreased rates of very-low-density lipoprotein (VLDL) production and increased LDL particle catabolism\textsuperscript{9}, suggesting that ANGPTL3 may play a direct role in the regulation of LDL-C metabolism in the liver. Therefore, ANGPTL3 could potentially be a therapeutic target for reducing LDL-C and decrease MI risk, although the FCH phenotype suggests that inhibiting ANGPTL3 may also reduce HDL-C – a potentially undesirable effect. However, the LDL-C and HDL-C traits in FCH families exhibit distinct modes of inheritance, supporting the notion that ANGPTL3 regulates LDL-C via a separate, unknown mechanism\textsuperscript{9}. To establish ANGPTL3 as a potential therapeutic target, additional characterization is needed to clarify its connection to human LDL-C metabolism.

We sought to evaluate the effects of ANGPTL3 deficiency in human hepatocytes \textit{in vitro} using hPSCs. Although ANGPTL3 mutant hepatocytes could be generated from FCH patient-specific iPSC cell lines, this would require the recruitment of rare FCH individuals. Moreover, phenotypic differences observed between FCH and control hepatocytes could not be specifically attributed to ANGPTL3 mutations due to genetic background differences. Therefore, we used genome editing technology to generate human embryonic stem (hES) cell lines with ANGPTL3 nonsense mutations, and differentiated these cell lines with matched wild-type controls into hepatocyte-like cells (HLCs) for characterization (Figure 4-3A).
Figure 4-3. TALEN genome editing to create an isogenic hPSC model of ANGPTL3 deficiency

(A) Experimental design.
(B) Targeting ANGPTL3 mutation sites using TALENs. Schematic showing the sites of the ANGPTL3 S17X and E129X mutations (underlined) and the chosen TALEN binding sites (boxes), with CEL-I activity results (performed in HEK 293T cells) for each TALEN pair shown.
RESULTS

TALEN genome editing to generate ANGPTL3 mutant hPSCs

In order to model the FCH phenotype as faithfully as possible \textit{in vitro}, we designed a genome editing strategy based on the locations of actual coding sequence mutations identified in FCH individuals – in particular the S17X and E129X mutations found in compound heterozygous members of an FCH family (\textbf{Figure 4-1}). We constructed and confirmed the activity of transcription activator-like effector nuclease (TALEN) pairs designed to target the sites of the S17X and E129X mutations in \textit{ANGPTL3} (\textbf{Figure 4-3B}). We also designed single-stranded DNA oligonucleotides (ssODNs) containing the actual FCH mutations, which can be used to introduce these mutations into a wild-type hESC line by TALEN-mediated genome editing.

We performed one round of genome editing using the E129X TALEN pair and generated modified HUES9 hESC clones containing small insertion or deletion mutations (indels) near the site of the E129X mutation. We identified 2 cell lines with biallelic frameshift mutations as well as multiple heterozygous lines and wild-type controls (\textbf{Figure 4-4A}). The two biallelic mutant cell lines were homozygous for either a 7 bp or 1 bp deletion mutations, which were predicted to cause a frameshift that would result in the production of truncated protein, either 136 or 138 amino acids in length (\textbf{Figure 4-4B}; wild-type ANGPTL3 is 461 amino acids). The overall modification rate observed with this TALEN pair in the first round of genome editing was 18%.
Figure 4-4. Confirmation of ANGPTL3 mutant hESC lines generated using TALENs

(A) Multiple clones with biallelic ANGPTL3 indel mutations were identified and sequence confirmed; sequences from two representative clones harboring predicted frameshift-causing mutations.

(B) Predicted amino acid sequence of ANGPTL3 protein resulting from indel mutations. The predicted mutant protein sequence is very similar to that produced by actual patient mutations despite.

(C) ssODN design and screening results for TALEN-mediated introduction of the E129X-causing mutation. A short fragment of genomic DNA containing the codon 129 was PCR amplified from individual hESC clones; EcoRI digestion of resulting PCR products was used to identify a clone containing the desired knock-in allele.

(D) Sanger sequencing confirmation of a heterozygous knock-in clone. An additional substitution was introduced to allow screening by EcoRI digestion; this mutation results in a predicted conservative amino acid substitution from Lys to Arg.
Next we used the S17X or E129X TALEN pairs with the corresponding ssODN to generate HUES hESC clones containing the FCH mutations. Using the S17X TALEN pair we identified several cell lines with indel mutations (approximately 5% overall modification rate) but did not identify any lines in which the ssODN had been incorporated out of 300 hESC clones screened. With the E129X TALEN pair and ssODN we generated 12 additional mutant cell lines (11% modification rate) including 2 lines containing the E129X mutation (Figure 4-4C and 4-4D; 1.8% knock-in rate based on 109 clones screened). In each knock-in cell line, the second allele of ANGPTL3 remained wild type, so the cell lines were not predicted to be ANGPTL3 deficient.

**Decreased ANGPTL3 level in mutant hepatocyte-like cells**

In light of the low knock-in efficiency, we assessed the possibility of using the biallelic indel clones as an approximation of the S17X/E129X genotype. The expression pattern of ANGPTL3 in humans is highly liver-specific\(^7\). Therefore, we differentiated ANGPTL3 mutant and wild-type control hESCs to “hepatocyte-like cells” (HLCs) using an established directed differentiation protocol (Figure 4-5A\(^{25}\)). As described in Chapter 3 of this thesis, HLCs generated with this protocol express hepatocyte-specific markers and possess cellular attributes of hepatocytes including apolipoprotein secretion. We assessed the expression level of ANGPTL3 in HLCs differentiated from 2 mutant and 2 wild type HUES9 hESC clones into HLCs at the mRNA and protein level. When normalized to albumin gene expression (ALB), to account for differences in differentiation between hPSC lines, we found that ANGPTL3 mRNA expression was significantly lower in mutant HLCs compared to the wild-type controls (Figure 4-5B) –
likely the result of nonsense-mediated mRNA decay in the mutant HLCs. Similarly, secreted ANGPTL3 protein was quantifiable in the media of wild-type HLCs, but undetectable in media from mutant HLCs (Figure 4-5C). These results suggest that ANGPTL3 gene expression is significantly diminished and ANGPTL3 protein is either not produced or is unable to be secreted in ANGPTL3 mutant HLCs. Therefore, cell lines containing biallelic ANGPTL3 frame-shift mutations near the locations of FCH mutations were considered effectively ANGPTL3 deficient.
Figure 4-5. ANGPTL3 production was significantly reduced in biallelic ANGPTL3 mutant Hepatocyte-like cells (HLCs) compared with matched wild-type control HLCs

(A) Schematic of protocol used for directed HLC differentiation of ANGPTL3 mutant and wild-type hESCs. Each cell line was differentiated in at least 6 well of a 24-well plate per differentiation experiment.

(B) qRT-PCR analysis of ANGPTL3 gene expression level in mutant and wild-type control HLCs from a representative differentiation. Relative gene expression levels were normalized to wild-type; error bars represent S.E.M.; results of statistical testing not shown.

(C) Representative ELISA measurement of secreted ANGPTL3 protein in the culture medium on day 25 of differentiation. ANGPTL3 protein level in media from mutant HLCs was below the detection limit of the assay.
CRISPR genome editing to generate multiple biallelic ANGPTL3-mutant hPSC lines

Due to relatively low modification efficiency obtained with TALEN genome editing, we were unable to produce a sufficient number of biallelic ANGPTL3 mutant cell lines for the proposed studies; doing so would have required a second round of genome editing. Sequential rounds of genome editing necessitate longer culturing times and two subcloning steps, which can have significant deleterious consequences, including the following: the accumulation background mutations, the emergence of karyotypic abnormalities, and an increased chance of mycoplasma contamination. Therefore, we utilized a different genome editing technology to generate additional hESC lines in a single round of genome editing.

The clustered regularly interspaced short palindromic repeat (CRISPR)/Cas system is a relatively new genome editing technology, which has been widely adopted because of its straightforward adaptability and high efficiency. As with TALENs, the CRISPR/Cas system can be effectively used for genome editing in hPSCs. As described in Chapter 2 of this thesis, we found that CRISPRs were significantly more active than TALENs in hPSCs, generating on-target modifications at a rate of 51% to 79%.

We generated CRISPR guideRNAs (gRNAs) targeting the same sites as the ANGPTL3 TALENs, which overlapped the locations of the S17X and E129X mutations. As expected, modification of ANGPTL3 was much more efficient with CRISPRs. Using the S17X CRISPR, we generated HUES9 hESC clones with indels in one or both alleles at a 79% efficiency, with 13% of the clones analyzed having biallelic indel mutations (i.e.
homozygous or compound heterozygous mutations, Table 4-1). For differentiation experiments we focused on 4 mutant HUES9 clones containing biallelic frameshift mutations as well as 4 wild-type control clones (Table 4-2). Although we did not generate cell lines with the exact mutations found in FCH individuals, the frameshift mutations present in the chosen mutant clones are predicted to have very similar effects on the ANGPTL3 transcript and produce truncated proteins comparable to those resulting from the S17X and E129X mutations.
Table 4-1. Summary of *ANGPTL3* genome editing efficiency in hESCs using TALENs and CRISPRs

TALEN and CRISPR target sites are underlined. "|" marks predicted CRISPR cut site. PAM sequences required by CRISPR/Cas9 are shown in bold. *ANGPTL3* S17 and E129 codons are indicated by red font. Efficiency is the percentage of modified cell lines among clones screened. 2 knock-in lines were identified when the E129X TALEN pair was used with a single stranded oligonucleotide for homology directed targeting (1% efficiency).

<table>
<thead>
<tr>
<th>Target Site</th>
<th>Target Sequence(^3)</th>
<th>TALENs Efficiency</th>
<th>CRISPRs Efficiency (Homozygous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S17X</td>
<td>TATTGTTCCTCTAGTTATTTCTCCCAGAATTGATCAAGACAATTTCATCA</td>
<td>5% (15/300)</td>
<td></td>
</tr>
<tr>
<td>S17X</td>
<td>TATTGTT CCTCTAGTTATTTCTCCCAGAATTGATCAAGACAATTTCATCA</td>
<td>79% (76/96)</td>
<td>13% (12/96)</td>
</tr>
<tr>
<td>E129X</td>
<td>TCAAAACTTGAAGGCTCTAGAAGAAGAAAAATTTCT ACTTTCAACAAAA</td>
<td>11% (12/109)</td>
<td>ND</td>
</tr>
<tr>
<td>E129X</td>
<td>TCAAAACTTGAAGGCTCTAGAAGAAGAAAAATTTCT ACTTTCAACAAAA</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>
Table 4-2. *ANGPTL3* genotypes of CRISPR-generated mutant hESC clones

4 mutant cell lines and 4 *ANGPTL3* wild-type control cell lines (not shown) were generated and used for HLC differentiation experiments. “DEL#” refers to an *ANGPTL3* deletion mutant allele of length ‘#’.

<table>
<thead>
<tr>
<th>Clone designation</th>
<th><em>ANGPTL3</em> genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 1</td>
<td>DEL4/DEL1</td>
</tr>
<tr>
<td>M 2</td>
<td>DEL1/DEL1</td>
</tr>
<tr>
<td>M 3</td>
<td>DEL14/DEL29</td>
</tr>
<tr>
<td>M 4</td>
<td>DEL1/DEL1</td>
</tr>
</tbody>
</table>
Differentiation to hepatocyte-like cells

As introduced above, hPSCs can be differentiated into HLCs using an optimized directed differentiation protocol. This protocol consists of 4 sequential differentiation steps and does not require embryoid body formation, feeder cells, or undefined reagents. Using this protocol we differentiated 4 homozygous mutant and 4 control cell lines in 24-well plates (6 to 12 wells per line, per experiment). Not all cell lines differentiated successfully in each experiment; unsuccessful differentiations were not included in phenotypic studies. Unsuccessful differentiations included experiments in which pluripotent cells remained at the end of the DE stage (based on cell morphology); too few cells remained at the end of the DE stage (due to excessive cell death or inadequate initial seeding density); or there was a preponderance of fibroblast-appearing cells at the later stages of differentiation rather than HLCs, which are polygonal-shaped cells with multiple lipid droplets and granules.

To assess variability in the degree of differentiation efficiency between the different cell lines we determined the expression level of hepatocyte marker genes by qRT-PCR. These included alpha-1 antitrypsin (A1AT), albumin (ALB), cytochrome P450 3A4 (CYP3A4), and alpha-fetoprotein (AFP) (Figure 4-6). Although AFP is not normally expressed in adult human liver, it is typically expressed by HLCs and is an indication of successful differentiation to the “hepatoblast” fate. A1AT, ALB, and CYP3A4 are markers of functional hepatocytes and should be expressed by HLCs. This analysis revealed noticeable variation between the different cell lines, especially with respect to AFP expression, as well as some variability in gene expression between different samples from the same cell lines. Despite these differences, we proceeded to examine
apolipoprotein secretion by *ANGPTL3* mutant and wild-type HLCs.
Figure 4-6. Hepatocyte marker genes are expressed at variable levels across multiple ANGPTL3 mutant and control cell lines differentiated to HLCs

Multiple matched hESC lines – ANGPTL3 wild-type and mutant clones (designated “WT” “M” respectively) – were differentiated to HLCs in 24-well plates ($n = 6 – 12$ wells/line); cell lines M4 and WT2 failed to differentiate and were not included in this experiment. RNA was made on day 25 of differentiation. qRT-PCR analysis of hepatocyte marker gene expression levels in mutant and wild-type control HLCs. Expression levels are normalized to the lowest mean expression level. Error bars represent S.E.M.
ANGPTL3 deficient HLCs exhibit altered apolipoprotein secretion

Individuals with compound heterozygous nonsense mutations in *ANGPTL3* have reduced VLDL production, increased LDL uptake, and extremely low levels of LDL-C in the bloodstream\(^9\). Therefore, despite differences in differentiation efficiency between hESC clones, we expected to observe a noticeable overall discrepancy in apolipoprotein secretion or uptake by HLCs generated from multiple mutant and wild-type clones. To assess for differences in VLDL secretion due to *ANGPTL3* mutations, we determined the level of apoB in the media of mutant and wild-type HLCs. apoB is the key apolipoprotein of VLDL and LDL particles and serves as a marker for these particles. We also determined the level of apoA-I, which is a marker of HLD particles. Although secreted ANGPTL3 affects HDL-C levels, it does so by inhibiting endothelial lipase in multiple tissues. Therefore, we did not expect ANGPTL3 deficiency to significantly effect apoA-I secretion by hepatocyte-like cells.

We performed multiple independent differentiations in 24-well plates (6–12 wells per cell line, per experiment) and quantified the amount of apoB, apoA-I, and albumin (ALB) in the media by ELISA. ALB secretion is an important synthetic function that is specifically performed by hepatocytes. To account for differences in the number of HLCs between differentiation wells, we normalized the levels of secreted apoB and apoA-I to the concentration of secreted ALB. We analyzed the results of 3 differentiation experiments, which included a total of 7 matched HUES9 clones – 3 mutant and 4 wild-type lines (mutant cell line #4 failed to differentiated in each experiment). Contrary to our hypotheses, we observed an overall increase in apoB and apoA-I (normalized to ALB) in the media of *ANGPTL3* mutant HLCs compared with wild type controls, over 3
independent experiments (average apoB, 39.9% greater; average apoA-I, 35.9% greater; \( P < 0.05 \); Figure 4-7).

However, significant inter-experiment variability, calling into question these overall conclusions. The average level of ALB, apoB, and apoA-I protein in the media of mutant and control HLCs varied substantially across the 3 differentiation experiments. The level of secreted ALB in mutant HLCs relative to wild type was higher in experiments 1 and 2 (\( P < 0.05 \) and \( P < 0.001 \), respectively), but lower in experiment 3 (not statistically significant; \( P > 0.05 \); Figure 4-7). The normalized level of secreted apoB and apoA-I was also inconsistent across experiments. In experiment 3, secreted apoB (normalized to ALB) was more than 2 fold higher in mutant HLCs compared with wild-type (\( P < 0.001 \)), yet it was not significantly different than wild-type in experiments 1 and 2 (\( P > 0.05 \); Figure 4-7). Similarly, secreted apoA-I (normalized to ALB) was higher in mutant compared with wild-type HLCs in experiments 2 and 3 (\( P < 0.001 \) and \( P < 0.05 \), respectively), but somewhat lower in experiment 1 (\( P > 0.05 \), Figure 4-7).

Thus, we did not obtain a reproducible phenotype with respect to apoB or apoA-I secretion, making it difficult to draw a conclusion regarding the effect of ANGPTL3 mutations on apolipoprotein secretion by HLCs. This inconsistency could be attributed to unknown differences between the cell lines or variability introduced by inefficient differentiation. Although each cell line was handled similarly through the genome editing process and only exposed to a single round of genome editing, there appear to be some differences between the individual clones that were used (other than the ANGPTL3 mutations) because certain clones consistently differentiated more efficiently than others. I do not suspect that the mutations in ANGPTL3 itself are involved. General
variability in the differentiation process was probably the most significant factor contributing to the observed inter-experimental variability. These issues are discussed further in the discussion section of this chapter. Because we were unable to elicit a reproducible phenotype attributable to ANGPTL3 deficiency we did not proceed with additional characterization.
**Figure 4-7.**

Protein concentrations were measured by ELISA in media collected from *ANGPTL3* wild type (WT) and mutant (M) HLCs. 3 independent differentiation experiments were performed. Albumin, apoB/albumin, and apoA-I/albumin levels are shown (rows 1-3); levels are normalized to the mean wild-type level for each experiment (*n* = 6 to 12 wells per cell line per experiment). Error bars represent S.E.M.
DISCUSSION

Although the regulation of intravascular LPL activity is almost certainly the mechanism connecting ANGPTL3 to altered TG levels, we hypothesized that ANGPTL3 influences LDL-C by an unknown distinct mechanism. This notion is supported by the lack of genetic evidence for an association between LPL and LDL-C level as well as the finding that individuals with ANGPTL3 mutations have decreased rates of VLDL production and increased LDL clearance. Therefore, we expected to observe decreased apoB protein (a surrogate for VLDL production) in the media of ANGPTL3 mutant HLCs resulting from either decreased apoB secretion or increased uptake. In direct contrast to this expectation, we observed a statistically significant trend of greater apoB in the media of mutant HLCs (39.9% average increase over 3 independent experiments), although we encountered substantial inter-experimental variability. We also generally observed more apoA-I in the media of mutant HLCs compared with wild-type controls (35.9% average increase over 3 independent experiments). These results do not support our hypotheses and are generally inconsistent with the human FCH phenotype, which includes extremely low LDL-C and HDL-C levels. Despite this discrepancy, our overall findings were subsequently corroborated by a recent independent study, which also provided some additional clarification regarding the apparently contradictory phenotype.

Tikka et al. (2014) generated immortalized primary hepatocytes in which ANGPTL3 was silenced by stably transduced shRNA – across several stable lines, the average reduction in ANGPTL3 was comparable to our results with ANGPTL3 mutant HLCs (85% and 90-95% reduction in ANGPTL3 gene expression and protein secretion.
respectively). The authors were surprised to find that secretion of TG, phospholipid, and apoB protein was significantly increased in ANGPTL3 silenced cells – consistent with our findings. Notably, they also observed a pronounced reduction in the ratio of TG/apoB secretion upon insulin stimulation in ANGPTL3 silenced cells. The authors report that this result is consistent with prior human studies, in which individuals with ANGPTL3 mutations were found to have a substantially lower ratio of plasma TG to apoB ratio (4.3 fold reduction). Furthermore, it is known that insulin signaling shifts the balance of hepatic VLDL assembly and secretion, decreasing the production of larger TG-rich VLDL-1 particles without effecting the production of TG-poor VLDL-2 particles. In addition, there is a well-documented association between ANGPTL3 mutations and increased insulin sensitivity in humans. A connection between ANGPTL3 activity and insulin responsiveness was observed with human hepatocytes in vitro. Thus, it appears that the hypolipidemic effect of human ANGPTL3 mutations – including decreased LDL-C – is one component of an interconnected state of increased insulin sensitivity, although considerable uncertainty remains.

My broadly, the paradoxical increase in apoB secretion by ANGPTL3 deficient hepatocytes illustrates the challenge of extrapolating cell-autonomous in vitro findings to in vivo human physiology. Cellular phenotypes and experimental findings can be extremely informative, although special caution must be exercised in the interpretation of these results.

In summary, the studies presented in this chapter demonstrate advances as well as disadvantages of using hPSCs to model disease-associated human genetic variation. We chose an experimental approach that would allow us to test a specific
hypothesis related to the function of ANGPTL3, while maximizing the likelihood that the results obtained would be relevant to actual human genetic variation. We used several genome editing tools and approaches for the studies presented in this chapter and observed on-target modification efficiencies consistent with previously documented trends\textsuperscript{4,22}. In particular, at two different locations within ANGPTL3, CRISPRs were significantly more efficient than TALENs for generating indel mutations using the same delivery methods and protocols for both CRISPR and TALEN genome editing.

CRISPRs are clearly an invaluable technology for generating loss-of-function hPSC lines – the utility of this technology cannot be overstated. As discussed in the results section above, perhaps the most significant advantage of CRISPR/Cas genome editing is that biallelic mutant cell lines can be generated in a single step, avoiding the potential consequences of extended hPSC culturing and multiple subcloning steps. We also introduce an FCH ANGPTL3 coding sequence mutation into wild-type hESCs using TALENs and a ssODN, observing a knock-in efficiency of 1.8%, which is in line with published reports for this technique. Although this level of efficiency represents a significant improvement over traditional gene targeting by homologous recombination, there remains substantial room for improvement. High efficiency gene correction in a single genome editing step would be extremely useful for a number of applications (discussed further in Chapter 7 of this thesis).

Despite the successful application of genome editing, we ultimately failed to gain a large amount of insight into the effects of ANGPTL3 mutations or the role of ANGPTL3 as a regulator LDL-C. This was due to the fact that we were unable to discover a consistent cellular phenotype in mutant HLCs, despite the use of a rigorous
experimental approach. As a result, we were not able to proceed with more advanced experiments, such as ANGPTL3 reconstitution in mutant cells or additional phenotypic characterization. This outcome was most likely due to two major factors: (1) inconsistency and variability in HLC differentiation, despite the use of up-to-date methods and matched cell lines; (2) apparent discordance between the behavior of HLCs in vitro and their in vivo counterparts, which exist within a larger physiologic context. The first factor is the focus of Chapter 5 of this thesis and is further discussed in Chapter 7. The second is also discussed in greater detail in Chapter 7. These challenges are not uncommon within the field of hPSC disease modeling in general and should be addressed by the development of improved differentiation methods and culturing techniques. Nevertheless, it will likely always be prudent to consider these issues seriously during the design and interpretation of individual disease modeling experiments.
METHODS

Cell culture

Undifferentiated hESCs were cultured as described in Chapter 3 of this thesis. The HUES9 hESC line was used; HUES9 cells are part of the NIH hESC registry.

Genome editing and confirmation of mutant cell lines

TALEN and CRISPR/Cas9 genome editing were performed as described in Chapter 2 of this thesis. Screening of genome edited clones was performed by PCR of the target site, gel electrophoresis, and Sanger sequencing of PCR products. Screening for incorporation of ssODN was performed by gel electrophoresis after PCR products were digested with EcoRI. Ambiguous mutant sequences were confirmed by TA cloning of PCR products followed by Sanger sequencing of TA clones. Genome editing efficacy was calculated as number of isolated hESC clones containing at least one on-target mutant allele divided by the total number of hESC clones screened.

Differentiation of hESCs into HLCs

HLC differentiation of hESCs was performed as described in Chapter 3 of this thesis.

Quantitative reverse transcriptase-polymerase chain reaction (qRT-PCR)

qRT-PCR was performed as described in Chapter 3 of this thesis, using SYBR green (Applied Biosystems) and B2M as a reference control gene.
Enzyme-linked immunosorbent assays (ELISAs)

ASGR1 ELISA was performed using the Human Angiopoietin-like 3 Quantikine ELISA Kit (R&D Systems), following the manufacturer’s instructions. Albumin, apoA-I, and apoB ELISAs were performed as described in Chapter 3 of this thesis.

Statistical analysis

Statistical significance was assessed using standard Student’s t-test (2-tail); \( P < 0.05 \) was considered statistically significant. To combine results from multiple independent differentiation experiments, data were normalized relative to wild-type controls from each experiment before being combined for statistical testing. Tukey’s post-hoc test was not applied because only one comparison was made (mutant vs. wild-type).
REFERENCES


Chapter 5: Purifying Differentiated Hepatocyte-Like Cells to Facilitate *In Vitro* Genetic Studies
INTRODUCTION

The liver plays many important roles in human physiology. As a central metabolic organ, normal and aberrant liver functioning is of particular importance clinically. Its detoxification activity influences the metabolism of drugs and its involvement in energy metabolism is central to common metabolic disorders such as type 2 diabetes mellitus and the metabolic syndrome, which stem from abnormal energy homeostasis. In addition, chronic liver disease and cirrhosis are associated with significant morbidity and mortality – patients with end-stage liver disease often require life-saving liver transplantation. For these reasons, there is a significant need for research related to liver disease as well as the development of regenerative medicine strategies to address a shortage of donor organs for transplantation.

Given the myriad functional activities conducted by the liver, its cellular makeup is relatively simple. The hepatocyte, which is the major parenchymal cell type of the liver, mediates many of the organ’s functions by carrying out a multitude of activities at the cellular level. Therefore, recapitulating key liver functions in vitro is straightforward in one sense (only requiring one cell type), but complex on the other hand, given the complicated nature of the hepatocyte and the fact that it normally functions in a larger physiological context. Despite the regenerative capacity of the liver in vivo, in vitro culture of hepatocytes is not straightforward and there is a limited supply of human hepatocytes. Primary hepatocytes typically lose their proliferative capacity and mature gene expression profile upon culture in vitro, although advanced culture systems such as micropatterned coculture (MPCC) can significantly extend the viability and support the functional capacity of hepatocytes ex vivo (Khetani and Bhatia, 2008).
Hepatocellular carcinoma cell lines, such as HepG2 and HuH7 cells are useful for certain types of studies, but are subject to the limitations discussed in Chapter 1 of this thesis. Therefore, there is a substantial need for a renewable source of human hepatocytes for in vitro studies and the development of cell based therapies. Human pluripotent stem cells are a promising potential source of these cells (reviewed by Schwartz et al., 2014).

The development of protocols for hPSC hepatocyte differentiation has progressed steadily and moved though several phases, from initial approaches requiring serum-containing media and feeder cells, to more highly optimized protocols (Si-Tayeb et al., 2010). There are now well-established methods for step-wise directed differentiation of hPSCs using defined media components and feeder-free culture conditions (Mallanna and Duncan, 2013). These protocols, which are approximately 25 days in duration from the onset of differentiation, can be used to produce hepatocytes from all types of hPSCs, routinely generating a > 70% proportion of cells that are positive for the hepatocyte-specific marker albumin. These cells also express other hepatocyte-specific genes and possess many of the hallmark cellular function of hepatocytes, such as cytochrome activity and apolipoprotein secretion; they can also be transplanted into immunodeficient mouse models of acute liver injury, achieving partial functional engraftment in the liver. However, hPSC-derived hepatocytes are not equivalent to primary adult human hepatocytes and are appropriately considered “hepatocyte-like cells (HLCs).” Unlike adult hepatocytes, HLCs typically retain expression of the fetal hepatocyte marker alpha-fetoprotein (AFP) and fall substantially short of mature hepatocytes in terms of quantifiable functional capabilities, such as
albumin secretion and drug detoxification.

Despite these shortcomings, HLCs have been used successfully to study aspects of development in vitro (DeLaForest et al., 2011) as well as for disease modeling applications ranging from predicting drug sensitivity to investigating cellular lipoprotein metabolism (Takayama et al., 2014; Rashid et al., 2010; Cayo et al., 2012; Chapter 3 of this thesis). Nevertheless, substantial obstacles must be overcome before more advanced disease modeling studies can be reasonable attempted with HLCs. A particularly notable challenge is variability and inefficiency of differentiation, which is a problem common to hPSC differentiation to other cell types as well. Even with optimized protocols, there is substantial variability in the efficiency and completeness of HLC differentiation depending on the hPSC line being differentiated or even between differentiation wells for a single cell line. As an example, Takayama et al. (2014) performed HLC differentiation using 12 hiPSC lines generated by the same reprogramming method. They observed differentiation efficiencies ranging from approximately 37% to 88% as measured by percentage albumin positive cells. In addition to differences in efficiency, there is likely also variability in the degree of maturation of differentiated HLCs. These results raise serious concerns regarding the potential utility of HLCs for in vitro studies.

Evidence suggests that this characteristic variability stems from inherent differences in hPSC lines (Kajiwara et al., 2012; Bock et al., 2011) as well as a multitude of known and unknown experimental variables. This problem poses a significant challenge particularly for in vitro modeling of relatively subtle phenotypes, as well as phenotypes that could be confounded by incomplete differentiation or the
presence of undesired cell types. These issues are particularly problematic for hPSC disease modeling studies involving multiple cell lines – such affected and control iPSC lines or mutant and control genome-edited hPSC lines (discussed in Chapter 1 of this thesis). Therefore, there is a tremendous need to address the problem of variable HLC differentiation in order to facilitate more advanced hPSC disease modeling.

There are a few logical strategies for addressing incomplete or inefficient hPSC differentiation in general. One is to improve or optimize the differentiation protocol itself, for example by driving differentiation more forcefully using programming viruses. This tactic was used successfully to enable efficient differentiation of hPSCs into white and brown adipocytes (Ahfeldt et al., 2012). This is a reasonable approach, although it can be technically challenging and may not entirely solve the problem of variable differentiation efficiencies among different hPSC lines.

Other than by protocol improvements, inefficient hPSC differentiation of some cell types has been overcome by prospective identification of the desired cell type based on expression of a cell-type specific marker or combination of markers (reviewed by Sandoe and Eggan, 2013). This is accomplished by generating hPSC lines harboring a reporter gene that will be expressed exclusively by the cell type of interest during or at the end of differentiation. Reporter hPSC lines are generated either by viral delivery of an exogenous construct or targeting of an endogenous locus. Regardless of the method used, this approach requires at lease some additional manipulation of each cell line being differentiated and therefore could be impractical for routine use with multiple cell lines. An additional innovative strategy has been demonstrated to achieve efficient purification of hPSC-derived cardiomyocytes following differentiation (Tohyama
et al., 2013). This non-genetic method takes advantage of differences in glucose and lactate metabolism between cardiomyocytes and non-cardiomyocytes, whereby differentiated cells are cultured in media that only permits the survival of cardiomyocytes. An approach of this nature could be particularly relevant to the purification of HLCs given the highly specialized metabolic capabilities of hepatocytes. However, such a strategy has not been established for HLCs to date.

For the above reasons we chose to pursue the use of a hepatocyte-specific cell-surface marker to enable the isolation of HLCs following differentiation. This approach does not require the introduction of a reporter construct and therefore it is potentially more straightforward than the aforementioned alternatives, particularly for differentiating a large number of cell lines.

We describe here the validation of a strategy for prospective isolation of HLCs differentiated from a variety of hPSC lines based on expression of a highly liver-specific cell surface protein, asialoglycoprotein receptor 1 (ASGR1). We extensively characterized ASGR1+ cells, showing that ASGR1 marks a subset of albumin positive HLCs, which are more similar to mature hepatocytes than standard HLCs. This work demonstrates the utility of isolating ASGR1+ cells for gene expression studies. Furthermore, we show that ASGR1-enriched HLCs can be replated, while retaining hepatocyte marker expression and cellular functions.
RESULTS

Directed differentiation of HLCs

As introduced above, well-established protocols can be used to produce hepatocyte-like cells (HLCs) from hPSCs by step-wise directed differentiation using defined, feeder-free conditions. Depending on the hPSC line used and other experimental variables, differentiation generally results in a mixture of HLCs (the desired cell type) as well as a variable number of other cell types (Figure 5-1A). Evidence suggests that unsuccessfully differentiated cells (non-HLCs) may be the result of failed endoderm specification early in differentiation or later misdirection of endoderm cells toward a non-hepatocyte fate (Loh et al., 2014). However, the specific composition of mixed HLC differentiation cultures has not been investigated.
Figure 5-1. Directed differentiation of hPSCs to hepatocyte-like cells (HLCs)

(A) Overview of optimized protocol for directed differentiation from hPSCs to HLCs illustrating the range of typical differentiation outcomes.

(B) Row-centered heatmap showing gene expression level of representative markers during each stage of HLC differentiation in vitro and in normal liver tissue in vivo from published microarray expression data (DeLaForest et al., 2011; Su et al., 2004). Red denotes higher than average expression and blue denotes lower than average expression. ASGR1 is specifically expressed in adult relative to fetal liver tissue and most highly expressed in vitro after the final stage of differentiation.

(C) Confocal microscopy images of immunocytochemistry staining after the third (IMH) and fourth (HLC) differentiation stages showing expression of hepatic lineage (HNF4A) and adult hepatocyte (ALB and ASGR1) markers. Cells were differentiated from HUES9 hESCs. Scale bars, 100 µm.
Figure 5-1 (continued)

A

Hepatocyte-like cells

Ideal
Average
Below average

hPSC → DE → HE → IMH → HLC

Act A, CHIR → FGF2, BMP4 → HGF → HGF, OncM, Dex

d0 → d3 → d8 → d13 → d25

B

HLC differentiation (DeLaForest et al.)

Liver tissue

PSC  DE  HE  IMH  HLC

Fetal  Adult

OCT4
SOX17
HNF4A
AFP
ALB
ASGR1

row min  row max

C

Albumin

HNF4A

ASGR1
Our lab has developed an optimized HLC directed differentiation protocol based on established methods (Si-Tayeb et al., 2010) with modest modifications to the first and fourth step of the 4-stage protocol. The first step of our protocol (the generation of definitive endoderm cells, “DE”) is based on the first step of an optimized protocol for directed differentiation of functional human pancreatic beta cells (Pagliuca et al., 2014), which like hepatocytes are derived from the endoderm lineage. All other steps are identical to the established HLC protocol except for the addition of dexamethasone during the final differentiation step (Figure 5-1A). Using this protocol with cell lines amenable to HLC differentiation, we generally achieve differentiation efficiencies ranging from 30% to 70% as measured by percentage albumin positive cells. These efficiencies are comparable to those obtained by other groups using similar protocols.

Selecting a candidate for prospective isolation of hepatocyte-like cells

To facilitate the prospective isolation of HLCs we searched the literature for a gene with a highly hepatocyte-specific expression pattern, which codes for a cell surface protein – the latter attribute would enable the isolation of live cells by fluorescence-activated cell sorting (FACS) or related techniques without requiring that the cells be fixed and permeabilized, as for an intracellular protein. A protein meeting these criteria could be used as the basis of a FACS strategy to separate hepatocyte-like cells from undesired cells after differentiation. We found that the asialoglycoprotein receptor 1 (ASGR1) fulfills these requirements.

ASGR1, also known as the Ashwell receptor, is specifically expressed in mature liver tissue in vivo, and is located on the cell surface of hepatocytes where it mediates
the specific clearance of galactosyl-terminal glycoproteins from the blood stream (OMIM, 2015). Of note, ASGR1 is relevant for the targeting of therapeutics to the liver.

We analyzed published gene expression data from human tissues as well as HLC differentiation of hESCs and found that ASGR1 is expressed in adult liver tissue, is not expressed or expressed at an extremely low level in fetal liver, and is expressed most highly during HLC differentiation after the final differentiation stage, the “HLC” stage (Figure 5-1B). We confirmed this expression pattern during HLC differentiation by immunocytochemistry (Figure 5-1C) during the last two differentiation stages. Next we more carefully assessed the expression of ASGR1 during HLC differentiation by flow cytometry.

**ASGR1 marks a subset of hepatocytes arising during HLC differentiation**

Using our HLC differentiation protocol we found that ASGR1 is present in a small number of cells after the third differentiation stage (the immature hepatocyte stage, “IMH”), and is more prevalent at the final stage of differentiation. This is in contrast to the expression pattern of albumin (ALB, a marker of functional hepatocytes), which is expressed at the IMH stage as well as the MH stage (Figures 5-2A and 5-3A). This makes ASGR1 preferable for prospective isolation of mature HLCs, in addition to the fact the ALB is not known to be expressed on the cell surface.
Figure 5-2. ASGR1 marks a subset of ALB+ cells within heterogeneous hepatocyte-like cell (HLC) differentiation cultures.

(A) Flow cytometry analysis showing the percentage of ALB and ASGR1 positive cells after the IMH and differentiation stages. ASGR1+ cells primarily arise during the HLC stage.

(B) Flow cytometry analysis showing co-expression of ASGR1 with HNF4A and ALB, illustrating that ASGR1+ cells have reached the desired differentiation state (n = 2 independent differentiations per analysis). Error bars represent s.e.m. Asterisks indicate statistical significance by Student’s t-test. *, P < 0.05.
Figure 5-3.

(A) Flow gating strategy and controls related to Figure 5-2A
(B) Gating controls for Figure 5-2B
(C-D) Summary of the percent albumin positive and surface-ASGR1 positive cells following HLC differentiation of four hPSC lines across multiple independent experiments.
We suspected that ASGR1 marks a more restricted population of HLCs as the percentage of ASGR1+ cells is almost always lower than the percentage of ALB+ cells in published reports as well as in our differentiations with several cell lines (Takayama et al., 2014 and Figures 5-3C and 5-3D). To confirm that ASGR1+ cells present at the end of HLC differentiation are in fact a subset of hepatocyte-like cells, as performed intracellular flow cytometry by co-staining ASGR1 with other markers. We found that ASGR1+ cells occur within a subpopulation of differentiated cells that express the hepatocyte-lineage marker HNF4A, as well as ALB (Figures 5-2B and 5-3B). These results show that ASGR1 is an ideal candidate marker for prospective isolation of hepatocytes following HLC differentiation.

**Enrichment of differentiated hepatocytes based on ASGR1 surface expression**

We next investigated the potential utility of ASGR1 for prospective hepatocyte isolation. We differentiated multiple hESC and hiPSC lines representing a range of HLC differentiation propensities and characterized the expression of hepatocyte markers among surface ASGR1+ cells. FACS analysis following HLC differentiation of four hPSC lines showed that a large proportion of surface ASGR1+ cells were also ALB+, even when the overall differentiation efficiency was extremely low (2.97% ASGR1+ cells, Figures 5-4A, 5-4B, 5-5A). Similar results were obtained when the expression of AAT, an additional marker of functional hepatocytes, was assessed (Figure 5-5B).
Figure 5-4. Enrichment of hepatocytes from HLC differentiation cultures by surface ASGR1 FACS.

(A) Four different hPSC lines were differentiated to HLCs, resulting in a range of differentiation efficiencies as expected (percent surface ASGR1+ cells: 2.97% to 48.28% by flow cytometry). The percentage of cells expressing the hepatocyte marker albumin (ALB) among unsorted HLCs, surface ASGR1- cells, and surface ASGR1+ cells was quantified by intracellular flow cytometry. Compared with unsorted HLCs and ASGR1- cells, a majority of surface ASGR1+ cells differentiated from all four hPSC lines were ALB+ (percent ALB+ cells > 85.28%).

(B) Summary of results shown in part A. Mean percent ALB+ cells by flow cytometry, among unsorted HLCs, surface ASGR1- cells, and surface ASGR1+ cells (n = 4 differentiations). Error bars represent s.e.m. Asterisks indicate statistical significance by Student’s t-test. *, P < 0.05.

(C) Heatmap summarizing qRT-PCR results, showing relative expression level in ASGR1+ cells compared with unsorted HLCs.
Figure 5-5. ASGR1 FACS gating strategy and controls; AAT flow cytometry analysis, related to Figure 5-4.

(A) ASGR1 FACS gating strategy used for experiments shown in Figure 5-4 and all ASGR1 FACS for RNA preparation.

(B) Two different hPSC lines were differentiated to HLCs, resulting in distinct differentiation efficiencies as expected (percent surface ASGR1+ cells: 29.79% and 46.55% by flow cytometry). The percentage of cells expressing the hepatocyte marker alpha alpha-1 antitrypsin (AAT) among unsorted HLCs, surface ASGR1- cells, and surface ASGR1+ cells was quantified by intracellular flow cytometry. Compared with unsorted HLCs and ASGR1- cells, a majority of surface ASGR1+ cells differentiated from both hPSC lines were AAT+ (percent AAT+ cells > 80%). Mean percentage of AAT+ cells in each population are summarized in the accompanying graph (n = 2 differentiations). Error bars represent s.e.m.
Next we analyzed the expression of several hepatocyte-specific genes in unsorted HLCs and FACS-isolated ASGR1+ cells from multiple differentiations of 3 representative hPSC lines. The average differentiation efficiency of the cell lines used in these experiments ranged from 4.2% to 26% as measured by percentage ASGR1+ cells (Figure 5-6A). During these experiments we noted that although there was substantial inter cell-line variability in differentiation efficiency, the differentiation efficiency of each cell line was generally consistent between experiments. Particular hPSC lines generally differentiated well (e.g. FHS-1 hiPSC), while others differentiated with a lower efficiency (e.g. 1016 hiPSC), as expected based on prior studies of HLC differentiation propensity (Takayama et al., 2014). For gene expression analysis we chose hepatocyte marker genes representing a range of expression patterns with respect to expression level in fetal vs. adult liver tissue (Su et al., 2004). These ranged from genes expressed predominately in fetal liver (alpha fetoprotein, AFP) to predominately in adult liver (apolipoproteins A1, APOA1). We found that expression of hepatocyte marker genes was significantly higher in ASGR1+ cells compared with matched unsorted HLCs (Figure 5-4C and 5-6B). This result was particularly pronounced for inefficient differentiations. The expression of more fetal expressed genes was in some cases diminished – particularly in the case of very efficient differentiates – while these same genes were expressed more highly in ASGR1+ cells from inefficient differentiations. The latter preponderance likely reflects the presence of an abundance of non-hepatic lineage cells in HLC differentiation cultures from inefficient differentiations. We also found that expression of marker genes representing the mesoderm and ectoderm lineages (NKX2-5 and PAX6, respectively) was generally
undetectable in ASGR1+ cells, but present to some degree in unsorted HLCs (data not shown). Collectively, these results strongly suggest that ASGR1 FACS results in the separation of hepatocyte-like cells from other cell types following HLC differentiation. They also suggest that ASGR1 FACS could potentially be useful for studies requiring the differentiation of multiple hPSC lines that have varied differentiation prophesies.
Figure 5-6. Detailed results of ASGR1-FACS and hepatocyte marker gene expression analysis, related to Figure 5-4

(A) Percentage of ASGR1+ cells following HLC differentiations of 3 different hPSC lines (n = 6 biological replicates from two independent differentiations per cell line). Shown are the differentiations used for ASGR1-FACS and subsequent gene expression analysis.

(B) Analysis of hepatocyte marker gene expression by qRT-PCR. Unsorted HLCs were compared to matched ASGR1+ cells from 6 HLC differentiations (2 differentiations per cell line). Shown are mean expression levels relative to the RPLP0, normalized to mean level in unsorted HLCs for each differentiation (n = 3 – 5 paired biological replicates per differentiation). Error bars represent s.e.m.
Figure 5-6 (continued)
Global transcriptional profiling of ASGR1+ cells

We next sought to confirm the identity of ASGR1+ cells using global gene expression profiling. We analyzed the transcriptional profiles of matched samples of ASGR1+ cells and unsorted HLCs differentiated from a hESC line (HUES9) and an hiPSC line (“1016”) as well as primary human hepatocytes (PHHs) and HepG2 hepatoma cells. Based on unbiased hierarchical clustering of all expressed genes, the global expression profiles of ASGR1+ cells were found to be distinct from those of matched unsorted HLCs (Figure 5-7A). There was greater correlation between the expression profiles of ASGR1+ cells from different hPSC lines ($r^2 = 0.959$) than there was between unsorted HLCs and ASGR1+ cells from the same cell line ($r^2 = 0.944$, $P < 0.05$, Figure 5-7B). In addition, the expression profiles of ASGR1-sorted cells and PHHs were slightly more highly correlated than were the expression profiles of HLCs and PHHs ($r^2 = 0.896$ vs. 0.866, $P < 1 \times 10^{-4}$). These results support the conclusion that ASGR1+ cells are at least as similar to hepatocytes as unsorted HLCs.
Figure 5-7 Microarray gene expression profile analysis of ASGR1+ cells, unsorted HLCs, primary human hepatocytes, and HepG2 hepatoma cells

(A) Hierarchical clustering of ASGR1+, HLC, and PHH samples based on all genes measured by transcriptional microarray and expressed above background. Clustering was performed using inverse Pearson correlation and Ward linkage.

(B) Heatmap showing pairwise Pearson correlation values for ASGR1+, HLC, and PHH samples based on the same expression data used in part A. Yellow, orange, and red color denotes lower, intermediate, and higher correlation respectively.

(C) Row-centered heatmap of hierarchical clustering performed on all genes differentially expressed between ASGR1+ cells and HLCs at a 5% FDR. Probe sets are colored based on average expression, with blue indicating lower expression and red indicating higher expression relative to all samples.

(D) Functional enrichment analysis of genes differentially expressed in ASGR1+ cells relative to HLCs. Statistical overrepresentation analysis was performed on the set of genes significantly upregulated in ASGR1+ cells vs. HLCs. Shown are the GO biological process terms most significantly overrepresented in the set of upregulated genes.

(E) Gene set enrichment analysis (GSEA) showing that a majority of genes from a liver-enriched gene set are more highly expressed in ASGR1+ cells relative to HLCs (normalized enrichment score: 3.5). Microarray gene expression data was arranged based on greater average expression in ASGR1+ cells (red color) or unsorted HLCs (blue color). Vertical black bars represent genes within the liver-enriched gene set. Out of 437 liver-enriched genes, 346 (79.2%) were higher in ASGR1+ cells.
Figure 5-7 (continued)
To further examine the effects of ASGR1 sorting, we performed differential expression analysis of the microarray data comparing ASGR1+ cells with unsorted HLCs. We performed hierarchical clustering and heatmap visualization of all genes (probesets) differentially expressed at a false discovery rate (FDR) of 5% (Figure 5-7C) as well as of the most highly differentially expressed genes (> 2 fold difference in expression between ASGR1+ cells and HLCs, Figure 5-8A). From this analysis we observed that ASGR1+ cells cluster more closely to PHHs than unsorted HLCs, while unsorted HLCs cluster more closely to HepG2 cells than PHHs. To further characterize the gene expression profile of ASGR1+ cells, a paired-sample design was used for differential expression analysis. This type of analysis enabled the identification of genes differentially expressed in ASGR1+ cells relative to matched HLCs differentiated from two different hPSC lines. 766 genes were found to be more than 2 fold differentially expressed between ASGR1+ cells and HLCs. Of these, 330 genes were significantly more highly expressed in ASGR1+ cells vs. unsorted HLCs. Functional enrichment analysis of the genes more highly expressed in ASGR1+ cells was performed using the PANTHER classification system (Mi et al., 2013). Statistical overrepresentation analysis revealed overrepresentation of a number of hepatocyte-related gene ontology (GO) biological processes (BP) in ASGR1+ cells vs. unsorted HLCs (Figure 5-7D). Many of the overrepresented processes related to key metabolic functions performed by the liver.
Figure 5-8. Differential expression analysis of microarray gene expression data shows that ASGR1+ cells share a distinct transcriptional profile and are more similar to adult hepatocytes compared with unsorted HLCs

(A) ASGR1+ cells and HLCs differentiated from two different hPSC lines were compared with HepG2 hepatoma cells and adult primary human hepatocytes (PHH) using Affymetrix U219 microarrays. Shown is a row-centered heatmap of hierarchical clustering performed on genes differentially expressed more than 2 fold between ASGR1+ cells and HLCs at a 5% FDR. Probe sets are colored based on average expression, with blue indicating lower expression and red indicating higher expression relative to all samples.

(B) Row-centered heatmap and hierarchical clustering of a panel of genes related to characteristic hepatic functions across the same samples as in part A. Gene expression levels are colored based on average expression level, with blue indicating below average expression and red indicating above average expression relative to all samples.
Figure 5-8 (continued)
Next, we assembled a panel of hepatocyte genes representing important categories of hepatic function: synthetic function (including the production of coagulation factors), energy metabolism (including lipid and carbohydrate metabolism and lipoprotein processing), bile production and metabolism, and detoxification and drug metabolism (including metabolism of xenobiotics). The majority of these genes were expressed more highly in ASGR1+ cells than in unsorted HLCs (Figure 5-8B). Hierarchical clustering based on the expression of these hepatic functional genes suggested that ASGR1+ cells are more similar to PHHs compared with unsorted HLCs. It should be noted that this trend is particularly pronounced for genes related to energy metabolism and systemic function in comparison with genes related to detoxification. This reflects the fact that ASGR1 isolation of HLCs is designed to address the problem of HLC impurity and differentiation variability, and is not aimed at substantially improving the characteristics of the actual HLCs, which are known to be relatively lacking in the expression of cytochromes and related genes.

Finally, we further verified that ASGR1+ cells are more similar to hepatocytes than unsorted HLCs using gene set enrichment analysis (GSEA, Subramanian et al., 2005). We found that 346 of 437 liver-enriched genes (79.2%) were more highly expressed in ASGR1+ cells (Figure 5-7E). Collectively these results suggest that ASGR1 FACS can be used to isolate true HLCs from heterogeneous HLCs culture.

**Replating ASGR1-enriched HLCs**

The findings presented above demonstrate that ASGR1 FACS isolation of hepatocytes from HLC differentiations could be used for gene expression studies.
(further confirmed in the following chapter of this thesis), which could be useful for certain types of in vitro genetic experiments. However, live cells in culture would be required for studies of cellular phenotypes. Therefore, we investigated the possibility of replating HLCs following enrichment of ASGR1+ cells by magnetic-activated cell sorting (MACS). We chose to use MACS for these initial experiments because we believed that cell would be more viable following the MACS-enrichment procedure than following FACS, especially considering that a conjugated ASGR1 antibody optimized for FACS was not initially.

Preliminary flow cytometry analysis suggested that, as expected, ASGR1-MACS following HLC differentiation results in the enrichment of ASGR1+ cells (data not shown). We next attempted to replate ASGR1 MACS-enriched HLCs and found that they adhered to standard collagen-coated 24-well plates. We examined the expression of lineage and functional hepatocyte markers in replated HLCs differentiated from hESCs and hiPSCs and found that the majority of the replated cells were positive for these markers by immunocytochemistry staining (Figure 5-9A). This was particularly striking considering that standard HLC differentiation cultures typically exhibit a patchy heterogeneous appearance upon immunostaining reflecting the presence of other cell types. This result was replicated with several hPSC lines. Of note, replated cells possessed morphologic features of primary hepatocytes on bright field microscopy; for example, binucleated cells were observed (Figure 5-10).
Figure 5-9. HLCs can be replated following ASGR1-MACS enrichment and retain hepatocyte characteristics.

(A) Immunocytochemistry staining for hepatocyte markers in primary human hepatocytes and replated HLCs. Primary human hepatocytes were plated as a monolayer; HLCs differentiated from FHS-2 iPSCs and HUES9 hESCs were replated following ASGR1-MACS enrichment. Hepatocyte marker expression by replated HLCs is more homogenous than is generally observed in standard HLC cultures. CK18, cytokeratin 18 (keratin 18, type I); ALB, albumin; AAT, (alpha-1 antitrypsin, SERPINA1). Staining was performed 72 hours after replating. Scale bars, 100 µm.

(B) Representative assays of hepatocyte cellular functions. Assays were performed at the following time points: HLC, day 28 of differentiation; replated HLC, 72 hours after replating on day 25 of differentiation; PHH, 48 hours after plating. For urea assay: HLCs were differentiated from HUES9 hESCs and 1016 hiPSCs ($n = 4$ per cell line for HLC and $n = 6 – 8$ per cell line for replated HLC); cryopreserved PHHs were thawed and plated ($n = 4$). Number of samples for other assays and additional details TBD. Error bars represent s.e.m. Asterisks indicate statistical significance by Student’s t-test. **, $P < 0.01$; ***, $P < 0.001$. 
Figure 5-9 (continued)

A

<table>
<thead>
<tr>
<th>Primary Hepatocytes</th>
<th>Replated HLCs (iPSC)</th>
<th>Replated HLCs (hESC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HNF4A</td>
<td>CK18</td>
<td>Albumin</td>
</tr>
</tbody>
</table>

B

- **HLC**
- **Replated HLC**
- **Primary hepatocytes**

**Albumin secretion**

**Urea secretion**

**CYP3A4 activity**
Figure 5-10.

**Figure 5-10 Bright field image of replated HLCs**
hESCs were differentiated to HLCs, enriched by ASGR1-MACS, and replated on collagen coated wells of a 24-well plate. Image was taken 72 hours after replating. Replated cells have morphologic features of hepatocytes; several binucleated cells are apparent (white arrows).
Lastly, we assessed representative hepatocyte cellular functions in replated HLCs in comparison with unsorted HLCs and PHHs plated as a simple monolayer. We found that albumin secretion and CYP3A4 activity (representing secretory and detoxification functions of the liver respectively) were significantly increased in replated HLCs vs. standard unsorted HLCs (Figure 5-9B). As anticipated based on prior characterization of hPSC-derived HLCs, these activities were substantially higher in PHHs. Collectively these results support the feasibility of replating HLCs following ASGR1 MACS-enrichment to produce a substantially more homogenous culture of HLCs.
DISCUSSION

The results presented in this study strongly suggest that prospective isolation of HLCs from heterogeneous HLC cultures based on ASGR1 surface expression is a viable solution to the problem of variable and incomplete HLC differentiation. We have shown that ASGR1, an established liver-specific protein, is expressed on a subset of ALB+ HLCs upon HLC differentiation. ASGR1 FACS can be used to isolate a population of cells that express hepatocyte functional markers (ALB and AAT) even when the overall differentiation efficiency is low. Therefore, ASGR1+ sorting addresses both the impurity of HLC differentiation cultures as well as inter-line variability in differentiation efficiency – both extremely significant limitations to performing in vitro genetic studies using HLCs.

ASGR1 FACS could facilitate studies of hepatic gene expression that would otherwise be compromised by the presence of a varying number of non-hepatocytes in HLC cultures. This would certainly be the case for studies involving the differentiation of multiple hiPSC lines, which will each, differentiate at a different efficiency. We are actively investigating the possibility of conducting this type of “next-generation association study” in which a cohort of hiPSCs are differentiated into HLCs, purified by ASGR1 FACS and then analyzed by gene expression and metabolomics profiling. To enable this type of study, conjugated ASGR1 antibodies can be used to simplify the ASGR1 sorting process (Figure 5-11).
Figure 5-11. Testing fluorophore-conjugated ASGR1 antibodies

Fluorophore-conjugated ASGR1 monoclonal antibodies for FACS were tested by surface flow cytometry analysis of HepG2 cells. Equivalent concentrations of isotype control and primary antibodies were used. The BD ASGR1-PE antibody produced the largest shift in the ASGR1+ population in the experiment shown.
In addition, we demonstrated the possibility of replating HLCs after ASGR1 enrichment. With further optimization this approach could facilitate phenotypic studies including functional assays on homogenous HLCs derived from multiple hPSC lines. Given the successes of advance culture systems in preserving the viability and functional properties of primary human hepatocytes in culture (Khetani and Bhatia, 2008; Khetani et al., 2013), we believe that the same improvements could be achieved with replated HLCs. Specially, following ASGR1 MACS-enrichment, HLCs would be plated onto MPCC plates. Supporting the feasibility of this approach, the MPCC technique has now been successfully adapted for culture of commercially produced hPSC-derived HLCs (iCell hepatocytes, Cellular Dynamics International; Berger et al., 2015), facilitating comprehensive assessment of drug toxicity using HLCs (Ware et al., 2015). Whether this method can be generally adapted to HLCs differentiated from multiple hPSC lines remains to be seen. Success in this area could significantly expand the potential uses of HLCs for \textit{in vitro} disease modeling.

In conclusion, the results presented in this report suggest that FACS isolation of ASGR1$^+$ cells is an effective approach for isolating HLCs from heterogeneous HLC differentiation cultures. This approach could generally be applied to isolate HLCs differentiated from hPSCs, much like similar sorting approaches are now commonly used to purify other hPSC-derived cell types such as neurons.
METHODS

Cell culture

Undifferentiated hPSC lines were cultured as described in Chapter 3 of this thesis. Two hESC lines (HUES1 and HUES9; Cowan et al., 2004) and two hiPSC lines (1016 and FHS-1) were used. HUES1 and HUES9 cells are part of the NIH hESC registry; 1016 and FHS-1 were derived using retroviral and Sendai viral reprogramming respectively. Cryopreserved primary adult human hepatocytes (PHHs) from a metabolically well characterized lot (HMCPMS lot no. Hu8138) were thawed and cultured as a simple monolayer following the manufacturer’s instructions (Life technologies).

Differentiation of hESCs and iPSCs into HLCs

Media preparation. All media preparation should be carried out in a tissue culture hood under aseptic conditions. Media should be freshly made and not used if stored for more than two weeks. Basal medium: RPMI-1640 + B27 supplement + penicillin/streptomycin (100 U/mL penicillin, 100 µg/mL streptomycin, final concentration). Use basal medium to make definitive endoderm (DE), hepatic endoderm (HE) and immature hepatocyte (IMH) media. DE medium: basal medium with 100 ng/ml Activin A and 3 µM CHIR99021. HE medium: basal medium with 5 ng/ml basic fibroblast growth factor (bFGF), 20 ng/ml bone morphogenic protein 4 (BMP4), and 0.5% DMSO. IMH medium: basal medium with 20 ng/ml hepatocyte growth factor (HGF) and 0.5% DMSO. Mature hepatocyte (MH) medium: hepatocyte basal medium with singleQuots added as well as 20 ng/ml HGF, 20 ng/ml oncostatin M, 100 nM dexamethasone, and
0.5% DMSO.

**Plating and differentiation.** Differentiation in six well plates was performed as follows:

1. Coat plates with Matrigel.
2. Split ES/iPS cells and plate at a density of $1.4 \times 10^5 \text{ cells/mL}$ in mTESR with 4μM of the ROCK inhibitor Y27632. Plate 1 mL per well. Plating density may have to be optimised for each different cell line (Day 1).
3. Day 2: Start treating cells with DE medium. Remove mTESR and replace with 1 ml DE medium per plate.
4. Days 3–4: Change DE medium daily. Cells will become spiky and ES cells should disappear or die off. At the end of this stage, you will have definitive endoderm cells.
5. Day 5: Start treating cells with HE medium for 5 days. Replace DE medium with 1 mL HE medium.
6. Days 6–9: Change HE medium daily. At the end of this stage, you will have hepatic endoderm cells.
7. Day 10: Start treating cells with IMH for 5 days. Replace HE medium with 1 mL IMH medium per well.
8. Days 11–14: Change IMH medium daily. At the end of this stage you will have immature hepatocyte-like cells.
9. Day 15: Start treating cells with MH medium for 10–12 days. Replace IMH medium with 1 mL MH medium
10. Days 16–24: Change MH medium daily. At the end of this stage you will have mature hepatocyte-like cells.

**Immunocytochemistry**

Immunocytochemistry (ICC) was performed as described in Chapter 3 of this thesis. The following primary and secondary antibodies were used for ICC staining: AAT (MA1-90438; Thermo scientific), ALB (A80-129A; Bethyl), ASGR1 (sc-52623; Santa Cruz), CK18 (ab82254; Abcam), HNF4A (ab92378; Abcam), Donkey Anti-Goat-IgG Alexa Fluor 555 (Life technologies), Donkey Anti-Mouse-IgG Alexa Fluor 488 (Life technologies), Donkey Anti-Rabbit-IgG Alexa Fluor 488 (Life technologies). All primary and secondary antibodies were used at dilutions of 1:250 and 1:1000, respectively. Hoechst (1:5000 dilution) (Invitrogen) was used for nuclear staining. Visualization of HNF4A, ALB, and ASGR1 ICC staining in unsorted HLCs was done using confocal microscopy. ICC staining of replated HLCs and primary human hepatocytes was visualized by inverted microscopy (Eclipse Ti; Nikon) using the NIS-Elements software package (Nikon).

**Intracellular flow cytometry**

For flow cytometry analysis, differentiated cells were fixed and stained using the BD Cytofix/Cytoperm Kit (BD Biosciences), following the manufacturer’s instructions. The primary and secondary antibodies described above were used at experimentally optimized dilutions. The following isotype control antibodies were used: Mouse IgG1 (Santa Cruz), Rabbit IgG (ab172730; Abcam), Goat IgG (sc-2028; Santa Cruz). All
isotype control antibodies were used at dilutions comparable to the corresponding primary antibody. Cells were analyzed using an LSRII cytometer and FlowJo software.

**ASGR1 FACS of HLCs with intracellular albumin staining**

1. Collect HLCs (MH day 12): wash with PBS and treat for 40min with accutase or 15min with trypsin (0.25%) followed by inactivation with FBS

2. Spin down (1000rpm, 5min, RT), wash with PBS, count cells

3. Resuspend in blocking buffer (5% HI FBS in PBS), 10million cells per 1mL

4. Block at 4deg for 20min on rocker

5. Distribute cells to 1.5mL tubes for staining: 50-100uL (0.5-1million cells) per tube minimum

6. Add conjugated ASGR1 primary or isotype control antibody (or no antibody for unstained control) and mix by tapping tube or gently pipetting
   
   a. BD ASGR1-PE antibody (5uL/test) 5uL per 1million cells (5uL antibody per 100uL cells [final volume 100-120uL])
   
   b. BD mouse IgG1k-PE isotype control antibody (20uL/test) 20uL per 1million cells (20uL antibody per 100uL cells [final volume 100-120uL])

7. Incubate at 4deg in dark for 30-45min on rocker

8. Spin down at 4deg, 250-300xg, 3-4min and carefully remove supernatant with pipette

9. Wash cells 2x with 0.5-1mL blocking buffer per tube

10. Resuspend cell pellet by tapping tube

11. Add 250uL BD Cytofix/Cytoperm solution per 1million cells and mix gently
12. Incubate at 4°deg in dark for 20 min on rocker.

13. Spin down at 4°deg, 350xg, 4 min and carefully remove supernatant (cells will be difficult to pellet at this stage).

14. Wash cells 2x with 1mL of 1X BD Perm/Wash solution (10X solution diluted 1:10 in distilled water).

15. Resuspend cells in Perm/Wash solution (100uL/1 million) and mix by tapping tube or gently pipetting.

16. Add primary or isotype control antibody (or no antibody for secondary antibody only control) for intracellular staining and mix gently: Bethyl goat polyclonal ALB antibody (1mg/mL stock) diluted 1:100 (1uL antibody per 100uL cells).

17. Incubate at 4°deg in dark for 1-2 hr.

18. Repeat spin/wash steps 17-18.

19. Resuspend cell pellet by tapping tube.

20. Add Perm/Wash solution (100uL/1 million cells) containing secondary antibody and mix gently: ALB: Life tech. Alexa Fluor 488 donkey anti-goat conjugated antibody (2mg/mL stock) diluted 1:500 (0.2uL antibody per 100uL cells).

21. Incubate at 4°deg in dark for 45 min-1 hr.

22. Repeat spin/wash steps 17-18.

23. Resuspend cells in flow cytometry buffer (1% FBS in PBS), 0.5mL per 1 million cells and analyze on LSRII cytometer using GFP/FITC and PE-Texas Red channels.
ASGR1 FACS of HLCs for RNA isolation

1. Perform FACS on day 25 to day 27 of differentiation.

2. Wash cells with PBS.

3. Trypsinize for 20 minutes. (Use 1mL per well of a 6-well).

4. After 20 minutes transfer trypsin and detached cells to 15mL conical tube (1 tube per sample, usually 3 wells of 6-well plate per sample). Move tubes to 4 degree.

5. Add 1mL 1:3 diluted Accutase to each well and treat for 20 additional minutes.

6. Cells will not detach on their own, but will visibly round up. They can be dissociated from the plate by pipetting up and down a few times and carefully scraping the plate with pipette tip.

7. Transfer cells to a 15mL conical (adding to previously collected cells).

8. Add extra 1mL PBS to each well to collect remaining cells.

9. At this point cells can be counted.

10. Spin down 5mins at 1k rpm at RT.

11. Resuspend cells in blocking buffer (PBS + 4% donkey serum) 1mL per 10 million cells.

12. Transfer 0.5 – 1 million cells to separate tubes for control staining (e.g. unstained, secondary antibody only, and isotype control). For isotype control replace primary with isotype antibody, perform secondary stain as normal. For secondary only antibody do not use a primary antibody and perform secondary stain as normal. For unstained control, leave out both antibodies.

13. From here on cells should be kept on ice or in the cold (4C).

15. Spin down for 4mins at 400rcf at 4C.

16. Resuspend in 100uL blocking buffer plus primary antibody (1:50 to 1:100 dilution) per million cells (ASGPR1 primary antibody: clone 8D7 sc-52623, Santa Cruz Biotech).

17. Stain for 45-60 mins at 4C.

18. Add approximately 900uL blocking buffer per 5 million cells.

19. Spin down 4mins 400rcf at 4C.

20. Remove liquid, resuspend in 100uL blocking buffer plus secondary antibody (1:1000 dilution) per 1M cells (Alexa Fluor488 secondary antibody, donkey anti-mouse, Life technologies).

21. Secondary antibodies are light sensitive, from here on out work in the dark.

22. Stain for 30-60 minutes at 4C.

23. Add approximately 900uL blocking buffer per 5 million cells.

24. Spin down 4 mins 400rcf at 4C.

25. Wash by removing liquid and resuspending in 500uL blocking buffer per sample.

26. Spin down 4 mins 400rcf at 4C.

27. Resuspend in blocking buffer (200uL per sample if less than 1 million cells, otherwise 400uL per 5 million cells) and transfer to FACS tube.

28. Perform FACS using a BD FACSARia II system:
   a. P1 - Gate on live cells – FSC vs SSC
   b. P2 – Gate on single cells – FSC-A vs FSC-W
   c. P4 - Gate on positive ASGPR – SSC vs PE
d. Sort minimum 100,000 cells for RNA – into 900 uL of LS trizol or 1 mL standard trizol

RNA isolation and qRT-PCR

RNA isolation and reverse transcription were performed as described in Chapter 3 of this thesis. qPCR was performed using TaqMan Gene Expression Master Mix (Applied Biosystems and TaqMan Gene Expression Assays (Life Technologies) for the following genes: RPLP0 (reference control gene), AFP, CYP3A7, FGA, ALB, SERPINA1, TF, and APOA1.

Microarray gene expression profiling

Array setup. Total RNA from each human cell type was extracted with Trizol (Invitrogen) and purified using the Clean & Concentrator-5 kit (Zymo Research). Labeled complementary DNA was hybridized to Affymetrix U219 gene expression arrays (Affymetrix). Arrays were scanned and converted into GeneChip probe result files (CEL) using Affymetrix Expression Console software.

Preprocessing and hierarchical clustering. Microarrays were normalized and background corrected using the Robust Multi-Array (RMA) method in the Bioconductor affy package in R v3.1.2. Normalized array values were reported on a log₂ scale and probesets with low expression values across all samples (log₂ intensity < 2.5) were filtered out. Unsupervised hierarchical clustering was performed based on Pearson correlation and Ward linkage using the hclust function in R. The statistical significance of clustering results was estimated by multiscale bootstrap resampling using the pvclust
Standard differential expression analysis and heatmap generation. Analysis was performed using Transcriptome Analysis Console software (Affymetrix). False discovery rate (FDR) adjusted P values for comparisons of average expression by cell type were calculated based on the Benjamini-Hochberg method and 5% FDR cutoff was used. Heatmaps were created using the heatmap.2 function in R, with accompanying dendrograms drawn based on Euclidean distance.

Paired-sample differential expression analysis, functional enrichment testing of differentially expressed genes, and gene set enrichment analysis. To assess gene expression differences between matched HLC and ASGR1+ samples differentiated from two different hPSC lines, the R/Bioconductor limma package was used to fit linear models for each gene, utilizing a paired-sample design. Moderated t statistics, log fold change, and P values were calculated. To correct for multiple hypothesis testing, false discovery rate adjusted P values were calculated by the Benjamini-Hochberg method. Probesets were filtered by FDR less than 5% and fold change greater than 2 (1079 probesets). Probesets were mapped to gene symbols. Functional enrichment analysis was performed using the PANTHER resource. Standard gene set enrichment analysis (GSEA; Subramanian et al., 2005) was performed using GSEA software (Broad Institute) and a list of liver-enriched genes (Yang et al., 2011).

Analysis of published gene expression data

Published microarray gene expression data was obtained from BioGPS.org and the GEO database. Analysis and heatmap generation was conducted using GENE-E
software (Broad Institute).

**ASGR1 MACS and replating of HLCs**

Cells were differentiated on 10cm plates and MACS enriched on day 25 of differentiation using the following procedure:

1. Dissociate cells using accutase
   
   a. About 3mL per plate
   
   b. About 20-30min

2. Add 3mL blocking buffer (10% FBS in PBS) to inactive accutase

3. Gently roll cell-roller across plates 2-3 times in horizontal and vertical direction

4. Transfer half volume of cells to 15mL tube

5. Gently scrape cells off plate by scarping glass pipette up and down in the vertical direction across the plate

6. Pipette a few times to collect cells, transfer to 15mL tube

7. Add 4mL additional blocking buffer to plate to collect remaining cells

8. Gently spin cells down: 3min, 1000rpm

9. Carefully remove supernatant and gently re-suspend cells in 5mL cold blocking buffer

10. Block for 15-20min on ice

11. Gently spin cells down

12. Carefully remove supernatant and gently re-suspend cells in 2mL cold blocking buffer
13. Add ASGR1 primary antibody (1:150 dilution): 13.33uL antibody per 2mL blocking buffer

14. Incubate 1hr on ice: gently invert tube 1-2 times to mix cells every 10-15min

15. Gently spin cells down and carefully remove supernatant

16. Gently re-suspend cells in 1.9mL cold MACS buffer

17. Add anti-mouse-IgG microbeads (1:20 dilution): 100uL beads per 2mL final volume in MACS buffer

18. Incubate 30-45min on ice

19. Warm bottle of MACS buffer to RT for use when running MACS

20. Add 8mL cold MACS buffer to each tube

21. Gently spin cells down and carefully remove supernatant

22. Gently re-suspend cells in 1mL cold MACS buffer

23. Pass cells through MACS filter into new 15mL tube

24. Wash filters with 0.5mL MACS buffer

25. Place 15mL tubes containing cells into top row of MACS chill-rack (left to right)

26. Place new 15mL tubes in chill-rack below each sample in rows 2 and 3 for waste and elute respectively

27. Setup MACS machine bottles

   a. Attach MACS buffer bottle to front right position

   b. Make sure there is enough MACS buffer for the run (50-60mL?)

   c. Make sure waste is empty (front left position)

   d. Make sure there is enough wash buffer (back right position)

   e. Make sure there is enough EtOH (back left position)
28. Startup MACS machine
   a. Turn on machine
   b. Run the following (?): Sep., ‘wash now’, ‘q-rinse’
29. Place chill-rack in MACS machine with samples facing machine
30. Run possels_s program; make sure last wash is sleep mode
31. Spin down cells: 5min, 1000rpm
32. Re-suspend cells in HLC re-plating media (500mL/re-plated well) and plate to wells of collagen-coated 24-well plate:
   a. 2-4 wells per 10cm plate, depending on size of cell pellet
   b. HLC re-plating media:
      i. HBM media + HCM additives
      ii. HGF and Dex (normal HLC concentrations)
      iii. ROCK-I (hPSC concentration)
      iv. Gentamicin (1:1000)
      v. Note: do not add Onc-M or DMSO
33. Wrap plate in parafilm and spin: 20min, 100g, at 37deg
34. Carefully remove parafilm and move plate to incubator
35. Change media the next day (HLC re-plating media without ROCK-I)

**Cellular assays of hepatocyte functions**

All cellular assay results were normalized to number of live cells determined by hemocytometer cell counting with trypan blue staining. Albumin secretion was quantified using enzyme-linked immunosorbent assay (ELISA) as described in Chapter 3 of this
thesis. Urea production was quantified using the QuantiChrom Urea Assay Kit (BioAssay Systems), following the manufacturer’s instructions. Cytochrome activity was quantified using the P450-Glo CYP3A4 Assay (Luc-PFBE) Cell-Based/Biochemical Assay (Promega), following the manufacturer's instructions.

**Statistical analysis**

Performed as described in Chapter 4 of this thesis.
REFERENCES


Tohyama S, Hattori F, Sano M, Hishiki T, Nagahata Y, Matsuura T, Hashimoto H,


Chapter 6: Modeling the Effects of a Common Noncoding Genetic Variant Within the 1p13/SORT1 Locus Using hPSCs
SUMMARY

Common DNA variants within the human 1p13 locus are highly associated with plasma LDL-C level and risk of myocardial infarction. Experimental characterization resulted in the discovery of a putative causal mechanism linking genotype to phenotype at this locus. Genotype at rs12740374 is associated with altered hepatic expression of the SORT1 gene, which in turn influences lipoprotein metabolism and plasma LDL-C concentration. In this thesis chapter, we used a hPSC-based in vitro model system to show that the minor allele of rs12740374 is directly responsible for elevated SORT1 expression in human hepatocyte-like cells. These results demonstrate that hPSCs can be used to characterize the effects of common non-coding genetic variants associated with human disease predisposition.
INTRODUCTION

Translating GWAS findings into mechanistic insight using hPSCs

Genome wide association studies (GWAS) have identified numerous genomic loci harboring common genetic variants associated with complex human diseases and risk factors\(^1\). As introduced in Chapter 1 of this thesis, these discoveries are an important step toward identifying novel genes and molecular pathways that underlie disease predisposition. In turn, newly appreciated causal genes and pathways represent potentially promising therapeutic targets. However, considerable experimental characterization is required to translate human genetics findings into molecular insight\(^2\). Modeling human disease \textit{in vitro} using human pluripotent stem cells (hPSCs) has been proposed as an innovative strategy to address this challenge. Although mainly a theoretical idea at first, hPSC disease modeling has become an increasingly viable research strategy due to advances in human induced pluripotent stem cell (iPSC) technology, the progression of genome editing techniques, and further refinement of differentiation methods\(^3\)\(^-\)\(^7\).

hPSCs have been used to model many different diseases in a variety of differentiated cell types and culture systems\(^8\)\(^-\)\(^9\). However, there are very few examples of the application of hPSC disease modeling specifically to study the effects of noncoding disease-associated genetic variation. Yet, a majority of disease-associated single nucleotide polymorphisms (SNPs) identified by GWAS are located in noncoding regions – 88% of GWAS SNPs catalogued by 2009 were intergenic or intronic\(^1\). Computational analysis suggests that up to 10% of the noncoding human genome is under purifying selection, reinforcing the notion that noncoding DNA regions harbor a
large number of important functional elements\textsuperscript{21}. Notably, common DNA variants that regulate gene expression generally act in a cell type dependent fashion and highly significant disease-associated variants are often located in predicted cell type specific enhancer elements\textsuperscript{22-23}. Therefore, there is a significant need for functional characterization of noncoding variants in relevant human cell types. Effectively utilizing hPSCs for this purpose would expand the boundaries of hPSC disease modeling and potentially yield important insight into the mechanistic underpinnings of disease.

As described in Chapter 1 of this thesis, hPSC disease modeling can take one of two general forms: unmatched or isogenic. In the latter approach, targeted genome editing is used to introduce, delete, or modify a particular DNA sequence of interest in an established hPSC line, generating modified cell lines and matched wild-type controls, which can then be differentiated into a mature cell type for characterization. This isogenic approach is generally a rigorous way to evaluate the function of a specific gene or DNA variant and may be particularly important for studies of noncoding genetic variants, which may exert more subtle phenotypic effects than rare variants or mutations underlying monogenic disorders\textsuperscript{*}. However, as more sophisticated hPSC techniques become available, advanced disease modeling involving cohorts of unmatched iPSCs or genome editing in multiple iPSC lines may become increasingly practical and informative (discussed further in Chapter 7 of this thesis).

A straightforward and precise method for characterizing noncoding DNA variants

\textsuperscript{*} GWAS, by design, identify genetic variants with relatively small effects on the associated trait. However, this does not diminish the potential utility of GWAS for identifying genes and pathways of clinical importance. A well-known illustration of this point is the fact that unbiased GWAS identified variants near the HMG-CoA reductase (HMGCR) gene. Although these variants are associated with only a modest increased myocardial infarction (MI) risk, HMGCR is known to be the principle target of statin medications, which are effective medications for reducing MI risk\textsuperscript{10}.  

215
is to use genome editing to either disrupt or introduce the suspected causal allele in hPSCs (Figure 6-1). This approach could be used to fulfill the "genetic version of Koch’s postulates," establishing a causal role for a specific noncoding DNA variant with respect to a disease phenotype. Identifying causal DNA variants within GWAS loci is an important endeavor because doing so can provide much-needed inroads for further studies, ultimately leading to the discovery of causal genes connected to the loci in question.
Figure 6-1. hPSC disease modeling applied to the study of common genetic variants

(Left) hPSCs, including human iPSCs, can be used to investigate the effects of disease-associated common genetic variants in relevant cell types \textit{in vitro}. However, if unmatched hPSC lines are used, phenotypic differences observed between cell lines cannot necessarily be attributed to a specific DNA variant due to differences in genetic background. Moreover, subtle phenotypic effects of a specific DNA variant may be obscured by overall cell-line differences not related to the disease of interest.

(Right) To avoid potential confounding effects resulting from unmatched genetic backgrounds, genome editing is used to generate hPSC lines with targeted genetic modifications, which are otherwise genetically matched (i.e. presumed to be otherwise isogenic). Using different genome editing techniques, a disease-associated DNA variant can either be deleted, or replaced by an alternate variant, depending on the experimental design.
Back to the future: hPSC disease modeling of the 1p13 locus

Common DNA variants within the human 1p13 locus are highly associated with plasma low-density lipoprotein cholesterol (LDL-C) level, coronary heart disease (CHD), and myocardial infarction (MI) risk\textsuperscript{11-13}. Single nucleotide polymorphisms (SNPs) present in this locus are among the most highly MI-associated common variants in the genome – inheriting two protective alleles of a tag SNP in the 1p13 locus (rs599839) is associated with an approximately 40\% reduction CHD risk\textsuperscript{12}. Experimental characterization of this locus, including studies in hPSC-derived hepatocyte-like cells (Chapter 3 of this thesis), resulted in the identification of sortilin 1 (SORT1) and elucidation of its role as a negative regulator of plasma LDL-C level\textsuperscript{14-16}.

In addition, a putative causal mechanism was discovered, which connects genotype to phenotype for this locus\textsuperscript{14}. Several of the most highly LDL-C-associated SNPs in the 1p13 locus are located in a noncoding region of the locus between the CELSR2 and PSRC1 genes. Genotype at these tightly linked SNPs is associated with altered hepatic expression of SORT1 – the likely causal gene for the locus – as well as CELSR2 and PSRC1 (Figure 6-2A). \textit{In vitro} characterization ultimately implicated a particular SNP, rs12740374, as the likely causal variant for the locus\textsuperscript{14}. The minor allele of rs12740374 is located in the highly conserved second position of a consensus CEBPA transcription factor-binding site, which is disrupted by the major allele of rs12740374 (Figure 6-2B). Notably, this CEBPA site and neighboring sequences are poorly conserved in the mouse genome, so it could only be studied in an endogenous setting using human cells\textsuperscript{14}. Collectively, these findings support the following mechanism connecting genotype to phenotype at the 1p13 locus: CEBPA binding to the
rs12740374 minor allele in hepatocytes leads to increased SORT1 expression, which in turn modulates lipoprotein metabolism, reducing plasma LDL-C concentration and MI risk\textsuperscript{14,17}.
Figure 6-2. 1p13 cholesterol-associated locus, showing the location of genes, results of prior eQTL analysis, and consensus CEBPA binding site involving rs12740374

(A, top) Diagram of 1p13 locus showing locations of genes and a region containing highly MI-associated SNPs, which includes rs12740374.

(A, bottom) Published expression quantitative trait locus (eQTL) analysis of the 1p13 locus showing association between the rs12740374 minor allele and increased expression of SORT1, CELSR2, and PSRC1 (“eQTL genes”) in liver but not adipose tissue. Expression of other genes in the locus (“non-eQTL genes:” SARS, MYBPHL, PSMA5) was not associated with rs12740374 genotype in all three tissues examined.

(B) Prior studies demonstrated that the minor allele of rs12740374 results in the creation of a CEBPA transcription factor binding site and the major allele disrupts the site. rs12740374 is located in a highly conserved position of the CEBPA consensus sequence. Of note, this site is not found in the orthologous DNA region in mice. CEBPA consensus motif logo obtained from the JASPAR database.
**Figure 6-2 (continued)**

### A

**Chromosome 1p13 locus**

![Diagram of chromosome 1p13 locus with SNPs and genes](image)

- **Gene Position**: Chromosome 1p13 locus
- **Genes**: SARS, CELSR2, PSRC1, SORT1, MYBPHL, PSMAS, SYPL2
- **SNPs**: rs12740374 major and minor

![Gene expression bar charts for liver (N=960), subcutaneous adipose (N=433), and omental adipose (N=520)](image)

- **Liver Expression**: Red = homozygote major (MM), Blue = homozygote minor (mm)
- **Adipose Expression**: Red = homozygote major (MM), Blue = homozygote minor (mm)

### B

**CEBPA sequence motif**

- **Sequence Motif**: CTAGGCAT
- **Mouse Sequence**: TGGCATGGTGGCCCTGAG

- **rs12740374 major**: TGGCTCGGCTGCCCTGAG
- **rs12740374 minor**: TGGCTCGGCTGCCCTGAG
- **Canonical C/EBP site**: CTAGGCAT
- **Mouse**: TGGCATGGTGGCCCTGAGG

---

221
The connection between rs12740374, SORT1, and MI risk is one of very few examples of a mechanism linking a common noncoding variant to a causal gene and a disease-related phenotype\textsuperscript{19}. However, it is suspected that similar mechanisms underlie a number of additional GWAS loci that have yet to be well characterized\textsuperscript{10,20}. Therefore, the connection between rs12740374 genotype and SORT1 expression represents an opportunity to examine the feasibility of hPSC modeling of common genetic variation by investigating an \textit{a priori} hypothesis. With this in mind, the studies presented in this thesis chapter were designed around the following objectives: (1) show that genotype at rs12740374 directly effects SORT1 expression in human hepatocyte-like cells, confirming the direction in which SORT1 activity could be therapeutically altered to reduce LDL-C levels and MI risk\textsuperscript{24}; (2) establish the feasibility of modeling common noncoding human genetic variation using an \textit{in vitro} hPSC model system.
RESULTS

Targeting rs12740374 in hPSCs using CRISPR/Cas9 genome editing

We chose the most straightforward approach for modeling the effects of rs12740374: begin with a hPSC line that is homozygous minor at rs12740374 and use genome editing to mutate (“knock-out”) the putative CEBPA transcription factor binding site created by the rs12740374 minor alleles (Figure 6-1 and 6-2B)\(^\dagger\). Although an alternative strategy would be to specifically introduce the rs12740374 major allele into a homozygous minor cell line – or visa versa, demonstrating “sufficiency” of the minor allele – we chose the more straightforward “gene disruption” approach for this proof-of-concept study\(^\dagger\). We genotyped rs12740374 in several hPSC lines and identified cell lines that were homozygous minor for rs12740374, including the FHS-1 human iPSC line (data not shown). We chose the FHS-1 iPSC line for the studies described in this thesis chapter because it could be efficiently differentiated into hepatocyte-like cells (HLCs) \textit{in vitro} (results shown in Chapter 5 of this thesis). At least minimal differentiation efficiency is essential to the present study because we hypothesize that genotype at rs12740374 effects \textit{SORT1} expression specifically in HLCs (Figure 6-2A)\(^\dagger\).

We differentiated FHS-1 iPSCs into HLCs to confirm its differentiation ability and explore the connection between cell identity and gene expression within the 1p13 locus. We used an optimized directed differentiation protocol that is routinely used to generate

\(^\dagger\) This knock-in strategy, termed “gene correction,” can be achieved using genome editing (demonstrated in Chapter 4 of this thesis), but is almost always less efficient than the “gene disruption” approach used in this thesis chapter. Issues related to the efficiency of gene correction are discussed in Chapter 7 of this thesis.
“hepatocyte-like cells” (HLCs) from hPSCs. The protocol is completed in approximately 25 days and includes 4 sequential differentiation stages, each named after the cell type generated in the corresponding stage: definitive endoderm (DE), hepatic-specified endoderm (HE), immature hepatocyte (IMH), and hepatocyte-like cell (HLC). Using the FHS-1 iPSC line, we determined the expression level of notable genes in pluripotent cells as well as each stage HLC differentiation. Consistent with successful differentiation, expression of the hepatocyte marker genes ASGR1 and ALB was significantly elevated at the final stage of differentiation (Figure 6-3A) relative to all other stages. Similarly, CEBPA – which encodes CEBPA the established liver-enriched transcription factor – was expressed predominantly at the IMH and HLC stages (Figure 6-3A). As described previously, genotype at rs12740374 and tightly linked SNPs is associated with altered expression of CELSR2, PSRC1, SORT1 in liver, but not in adipose tissue (Figure 6-2A). We found that CELSR2 and PSRC1 were expressed most highly in pluripotent cells and expression at lower levels in the later stages of HLC differentiation. Conversely, SORT1 expression was highest at the HLC stage, lowest during the DE and HE stages, and intermediate in pluripotent cells (Figure 6-3A). These expression patterns are consistent with published gene expression data from HLC differentiation of hESCs (Figure 6-3B). The significant differences in CEBPA, CELSR2, PSRC1, and SORT1 gene expression at each differentiation stage underscore the importance of studying gene regulatory mechanisms in the proper cellular context.

‡ hPSC-derived hepatocytes are referred to as hepatocyte-like cells because they possess many of the cellular characteristics of hepatocytes in vivo, but unlike adult hepatocytes they generally retain expression of alpha-fetoprotein (AFP), a marker of fetal hepatocytes.
Figure 6-3. Expression pattern of hepatocyte marker genes and 1p13-locus genes upon differentiation of the FHS-1 human iPSC line into hepatocyte-like cells (HLCs)

(A) qRT-PCR analysis of gene expression at each stage of HLC differentiation for the FHS-1 hiPSC line, which was used to generate rs12740374 genome-edited clones. Genes chosen for analysis included hepatocyte markers (ASGR1 and ALB) to gauge the progress of differentiation, and genes related to the 1p13 locus eQTL result (CEBPA, CELSR2, PSRC1, and SORT1). Mean expression levels relative to the RPLP0 were normalized to mean level at the HLC stage of differentiation (n = 3 wells per stage). Error bars represent s.e.m.

(B) Row-centered heatmap showing expression levels of genes included in part A at each stage of hESC HLC differentiation from published microarray data. Red denotes higher than average expression and blue denotes lower than average expression. Gene expression trends were similar to those observed during differentiation of FHS-1 hiPSCs.
Figure 6-3 (continued)

A

ASGR1

ALB

CEBPA

CELSR2

PSRC1

SORT1

Normalized gene expression

PSC  DE  HE  IMH  HLC

HLC differentiation
(DeLaForest et al.)

B

HLC  PSC  DE  HE  IMH  HLC

ASGR1

ALB

CEBPA

CELSR2

PSRC1

SORT1

row min  row max
Using standard CRISPR design criteria described in Chapter 2 of this thesis, we chose a CRISPR target site (protospacer sequence) that overlaps rs12740374 and generated the corresponding custom guide-RNA (gRNA) construct. In CRISPR/Cas9 genome editing, short deletion or insertion mutations (indels) typically occur 3 base pairs upstream (5') of the protospacer adjacent motif (PAM)\textsuperscript{30}. Based on this tendency, the CRISPR that we designed would be expected to introduce mutations overlapping rs12740374 (Figure 6-4A).

We performed CRISPR/Cas9 genome editing in the FHS-1 hiPSC line (homozygous minor for rs12740374). We obtained several mutant clones harboring biallelic indel mutations overlapping rs12740374. The overall efficiency for this CRISPR was approximately 66% (the proportion of isolated hPSC clones with at least one confirmed mutant allele near the target site, determined by sanger sequencing). This is representative of the extremely high efficiency that can be routinely achieved with CRISPR/Cas9 genome editing using the optimized procedure described in Chapter 2 of this thesis (see Chapter 7 for additional discussion of CRISPR/Cas genome editing). We focused our attention in particular on 4 mutant clones with deletions that disrupted substantial portions of the CEBPA consensus sequence including rs12740374 (Figure 6-4A). Subsequent differentiation experiments were performed with these 4 mutant clones as well as 2 wild-type clones and unmodified FHS-1 iPSCs (a total of 3 wild-type cell lines).
Figure 6-4. rs12740374 mutant HLCs exhibit 1p13-locus gene expression changes reflecting human eQTL results

(A) Generation of rs12740374 “knockout” and control hiPSC clones with CRISPR/Cas9 genome editing of an rs12740374 homozygous minor hiPSC line (FHS-1). The location of rs12740374, the CRISPR target site, and other notable features are shown. Deletions in the two alleles of each mutant clone are indicated; inserted bases present on one allele of clone D are shown in lowercase. The parental cell line and 6 genome-edited clones were used for subsequent studies (3 wild-type and 4 mutant cell lines in total).

(B) Percentage ASGR1+ cells by FACS and ALB gene expression level in ASGR1+ cells by qRT-PCR (n = 18, wild-type; n = 20, mutant; 1 – 3 independent differentiations per cell line). HLCs were differentiated from 3 wild-type and 4 mutant hiPSC lines. Mean gene expression level relative to the RPLP0 reference gene was normalized to mean level in wild-type cell lines. To adjust for variability between independent experiments, prior to calculating mean expression levels, the expression level of each sample was normalized to the mean expression level of the wild-type cell lines from the same experiment (n = 7 – 20 independent samples per experiment). Error bars represent s.e.m. Asterisks indicate statistical significance by Student’s t-test. *, P < 0.05.

(C) Expression of 1p13 locus genes by qRT-PCR in ASGR1+ cells (n = 18, wild-type; n = 20, mutant; 1 – 3 independent differentiations per cell line). Cells were isolated by FACS following HLC differentiation of 3 wild-type and 4 mutant hiPSC lines. Mean gene expression levels relative to the RPLP0 reference gene were normalized to mean level in wild-type cell lines. To adjust for variability between independent experiments, the expression levels of each sample were first normalized as described in part B. Error bars represent s.e.m. Asterisks indicate statistical significance by Student’s t-test. *, P < 0.05; **, P < 0.01; ***, P < 0.001.
Figure 6-4 (continued)

A

C/EBP site; rs12740374 minor allele; CRISPR target: PAM

<table>
<thead>
<tr>
<th></th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type</td>
<td>GGCTCGGTCGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Parental FHS-1(WT)</td>
<td>GGCTCGGTCGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Parental FHS-1(WT)</td>
<td>GGCTCGGTCGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone A (WT)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone A (WT)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone B (WT)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone B (WT)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone C (del11)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone C (del11)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone D (del19)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone D (del13;ins3)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone E (del20)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone E (del19)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone F (del27)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
<tr>
<td>Clone F (del11)</td>
<td>GGCTCGGCTGCCCTGAGTTGCTCAATCAAGCAGAGGT</td>
</tr>
</tbody>
</table>

B

![Graph showing percent ASGR1+ cells and normalized gene expression for wild-type and mutant conditions.](image)

C

![Graph showing normalized gene expression for SARS, CELSR2, PSRC1, MYBPHL, SORT1, PSMA5.](image)
The rs12740374 minor allele is required for the maintenance of elevated SORT1 gene expression in human hepatocyte-like cells

Expression quantitative trait locus (eQTL) analysis of the 1p13 locus performed on human tissue samples\(^\text{14}\) suggested that the minor allele of rs12740374 is associated with increased expression of several genes ("eQTL genes:" CELSR2, PSRC1, SORT1) in liver but not adipose tissue, while the expression of other nearby genes was unchanged ("non-eQTL genes:" SARS, MYBPHL, PSMA5) in all three tissues. This effect appears to be the result of CEBPA recruitment when the rs12740374 minor allele is present\(^\text{14}\). Based on these findings, we hypothesize that the minor allele of rs12740374 directly augments SORT1 expression in human hepatocytes. We expect to obtain results with our hPSC-derived model system that are consistent with the proposed \textit{in vivo} interaction between rs12730374 and SORT1. Specifically, expression of SORT1 and the other eQTL genes should be lower in rs12740374 mutant HLCs compared with wild-type controls, while expression of the non-eQTL genes should be unaffected.

We differentiated the mutant and control iPSC lines described above into HLCs and isolated ASGR1+ cells by FACS\(^\text{§}\). We analyzed the results of three independent differentiation experiments. As expected, a range of differentiation efficiencies were observed among the cell lines (\textbf{Figure 6-5A}); the average proportion of ASGR1+ cells was lower for differentiations with the mutant cell lines compared with with wild-type (\textbf{Figure 6-4B}). However, upon FACS isolation of ASGR1+ cells and gene expression

\(^\text{§}\) Extensive characterization described in Chapter 5 of this thesis showed that ASGR1+ cells are more similar than unsorted HLCs to primary hepatocytes. Moreover, FACS isolation of ASGR1+ cells following HLC differentiated of multiple cell lines yielded more comparable cells across differentiations as determined by hepatocyte marker expression.
analysis, the average level of *ALB* expression in mutant HLCs was remarkably similar to wild-type (Figure 6-4B and 6-5B).
Figure 6-5. Detailed results of ASGR1-FACS and 1p13-locus gene expression analysis with rs12740374 genome-edited clones, related to Figure 6-4

(A) Percentage of ASGR1+ cells following HLC differentiations of the parental hiPSC line and 6 genome-edited clones (n = 3 – 8 biological replicates from 1 – 3 independent differentiations per cell line). Shown are the differentiations used for ASGR1-FACS and subsequent gene expression analysis.

(B) qRT-PCR analysis of 1p13-locus gene expression in ASGR1+ HLCs differentiated from wild-type and rs12740374 genome-edited (mutant) hiPSC lines. Genes analyzed include eQTL (CEBPA, CELSR2, PSRC1, and SORT1) and non-eQTL(SARS, MYBPHL, PSMA5) as well as the hepatocyte marker ALB. Shown are normalized mean expression levels relative to the RPLP0 reference gene (n = 3 – 8 biological replicates from 1 – 3 independent differentiations per cell line). To adjust for variability between independent experiments, prior to calculating mean expression levels, the expression levels of each sample were normalized to the mean expression level of the wild-type cell lines from the same experiment (n = 7 – 20 independent samples per experiment). Note: the normalized expression level of clone B is equal to 1 because it was the only WT clone from its experiment. Error bars represent s.e.m.
Figure 6-5 (continued)

A

Percent ASGR1+ cells

Mean percentage ASGR1+

B

Normalized gene expression

ALB

wild-type

mutant

CELSR2

SARS

PSRC1

MYBPHL

SORT1

PSMA5
Finally, we examined the expression levels of eQTL and non-eQTL genes in the 1p13 locus by qRT-PCR. For gene expression analysis we chose to use *RPLP0* (which encodes “ribosomal protein, large, P0”) as a reference gene, because among genes commonly used for qRT-PCR standardization, *RPLP0* is expression relatively consistently across an array of human tissues types\(^3\). We analyzed the results of 3 independent differentiation experiments; we found that as expected, the average gene expression levels of *CELSR2*, *PSRC1*, and *SORT1* were lower in mutant HLCs relative to wild-type HLCs (*P* < 0.001; Figure 6-4C). The average gene expression levels of the non-eQTL genes were not significantly different between mutant and wild-type HLCs (*P* > 0.05; Figure 6-4C). Although the average expression level of *MYBPHL* was lower in mutant HLCs relative to wild-type, its expression level varied considerably across all cell lines (Figure 6-5B) and was generally expressed a low level in HLCs relative to all other genes analyzed (data not shown). Notably, the expression levels in each cell line exhibited relatively little variability across all differentiation experiments, reflecting a reproducible gene expression phenotype. Overall, these results are consistent with the original 1p13 locus eQTL findings. Furthermore, they demonstrate that the minor allele of rs12740374 is directly responsible for elevated *SORT1* expression in human HLCs.
DISCUSSION

In this thesis chapter, we used an hPSC-based model system to show that the minor allele of rs12740374 is directly responsible for elevated \textit{SORT1} expression in human hepatocyte-like cells (HLCs). This result supports the mechanism, uncovered previously\textsuperscript{14}, connecting genotype to phenotype at the 1p13 locus. It also further implicates \textit{SORT1} as the causal gene for the 1p13 locus, connecting altered \textit{SORT1} expression directly to CHD predisposition\textsuperscript{11,13}. Although it is well established that increased \textit{SORT1} expression or activity could decrease LDL-C level – thereby possibly reducing the risk of MI – an ideal drug target should be directly linked to disease risk as well as a specific risk factor. Therefore, establishing a direct connection between rs12740374 and \textit{SORT1} expression lends support to \textit{SORT1} as a therapeutic target for the prevention of MI.

The studies presented in this thesis chapter illustrate a novel application of the hPSC disease modeling approach. By confirming an \textit{a priori} hypothesis, these results demonstrate the feasibility of using hPSCs to characterize common non-coding genetic variants associated with human disease. This work suggests that similar experiments could be conducted using hPSCs and genome editing to investigate noncoding variants identified by GWAS whose functional significance has yet to be determined. Finally, the cell lines and methods presented in this thesis chapter could serve as the basis for a more advanced disease modeling study. Specifically, a cohort of human iPSCs generated from individuals with different rs12740374 genotypes could be differentiated into HLCs, purified by ASGR1 FACS and then analyzed for gene expression differences correlated to genotype. This type of \textit{“in vitro association study”} remains an untested
approach and will likely include a higher level of experimental variability. Therefore, in this type of study the wild type and genome-edited cell lines described in this thesis chapter could be useful as reference points for evaluating changes in \textit{SORT1} expression.
METHODS

The following methods were performed as described in Chapter 5 of this thesis: cell culture; differentiation of hESCs and iPSCs into HLCs; ASGR1 FACS of HLCs for RNA isolation.

Genome editing and confirmation of mutant cell lines

CRISPR/Cas9 genome editing was performed as described in Chapter 2 of this thesis. Screening of genome edited clones was performed by PCR of the target site, gel electrophoresis, and Sanger sequencing of PCR products. Ambiguous mutant sequences were confirmed by TA cloning of PCR products followed by Sanger sequencing of TA clones. Genome editing efficacy was calculated as number of isolated hESC clones containing at least one on-target mutant allele divided by the total number of hESC clones screened.

RNA isolation and qRT-PCR

RNA isolation and reverse transcription were performed as described in Chapter 3 of this thesis. qPCR was performed using TaqMan Gene Expression Master Mix (Applied Biosystems and TaqMan Gene Expression Assays (Life Technologies) for the following genes: RPLP0 (reference control gene), ALB, ASGR1, CEBPA, CELSR2, MYBPHL, PSMA5, PSRC1, SARS, and SORT1.

Analysis of published gene expression data

Analysis of published microarray gene expression data was performed as
described in Chapter 5 of this thesis.

Statistical analysis

Performed as described in Chapter 4 of this thesis.
REFERENCES


Chapter 7: Implications, Discussion, and Future Directions
INTRODUCTION

The work presented in this thesis demonstrates the application of hPSC disease modeling for experimentally characterizing genetic findings underlying cardiovascular disease (CVD) predisposition. The availability of human genetic findings related to complex human disease represents an opportunity to discover novel genes and molecular pathways that contribute to the pathogenesis of common diseases. Fulfilling this promise requires careful evaluation of disease-associated genetic variants in an appropriate model system. As discussed throughout this thesis, hPSCs are a uniquely valuable resource for research of this nature. By in vitro differentiation, hPSCs provide access to disease-relevant cell types; moreover, advances in genomic modification techniques have further unlocked the potential of hPSCs as the basis of an experimental model system. In addition, iPSC reprogramming has made it possible to generate hPSCs from individuals representing a wide range of genetic backgrounds and phenotypic traits.

As described in Chapter 1 of this thesis, there is a significant need for new therapies to prevent CVD, in particular coronary artery disease (CAD) and myocardial infarction (MI). Although substantial progress has been made in understanding the pathogenesis of CVD and developing preventative strategies and treatments approaches, CVD remains a leading cause of death. Furthermore, the prevalence of obesity and diabetes – risk factors for CVD – has increased in an alarming fashion. Despite the promise of hPSCs as transformative tools for biomedical research, substantial technical obstacles have impeded efforts to realize this potential.
A significant portion of the work presented in this thesis was dedicated to addressing and overcoming two key challenges: (1) the difficulty of making targeted genetic modifications in hPSCs; (2) the confounding effect of variability in hPSC differentiation to hepatocyte-like cells (HLCs), which significantly limited the utility of hPSCs for more sophisticated experimental studies pertaining to this cell type. The first challenge – the technical difficulty of genetically modifying hPSCs – has been met by impressive progression centered on the development of genome editing technology. In 2009, the ability to introduce targeted genetic modifications in hPSCs was extremely limited by the inefficiency of homologous recombination, making the use of this important technique very impractical and inflexible. The genome editing technique emerged as a potential solution to this problem, but the unique challenge of culturing hPSCs mandated considerable methodological optimization to enable the efficient application of genome editing for hPSCs. Initially, genome editing in hPSCs was laborious and limited by inefficient DNA delivery and the difficulty of high-throughput expansion of hPSC clones for screening purposes. To address these challenges, we developed the methods described in Chapter 2 of this thesis. These methods of efficient hPSC genome editing have now been widely used to generate modified hPSCs in a rapid and straightforward fashion. This has facilitated the generation of many hPSC disease models, such as those described in Chapters 3 and 4 of this thesis, among other applications.

The second obstacle introduced above relates to a problem that is often encountered in hPSC disease modeling studies: the challenge of detecting consistent cellular phenotypes in mature cell-types differentiated from hPSCs in vitro. This
difficulty is often the result of variable, incomplete, or inefficient hPSC differentiation with available protocols. Many creative solutions have been developed by investigators to address this challenge depending on the needs of their individual research studies. In Chapter 5 of this thesis we address this challenge in the context of hPSC-differentiation to HLCs, where variability in differentiation efficiency can significantly limit the use of these cells for certain functional studies. We developed and carefully validated a simple and readily applicable strategy for isolating HLCs from heterogeneous differentiation cultures based on the expression of an endogenous hepatocyte-specific cell-surface protein, ASGR1. This approach does not require extra manipulation of the hPSCs and does not involve the use of constructs for selection or identification of the target cell type – strategies that have been used successfully, but are somewhat less straightforward and scalable. Considerations pertaining to this challenge, as well as alternative solutions have recently been reviewed\textsuperscript{1-2}.

Utilizing the methods described above, especially genome editing, we generated isogenic disease models to investigate several instances of genetic variation underlying CVD risk, in those influencing plasma lipid concentrations. We used these models to address specific hypotheses pertaining to the functions of the \textit{SORT1} and \textit{ANGPTL3} genes, as well as the noncoding SNP rs12740374 in the 1p13 locus. Each of these genetic factors was initially linked to altered plasma lipid levels and CVD risk by human genetic studies\textsuperscript{3-6}, although further characterization was needed to clarify their specific functional roles. We performed this characterization using the isogenic hPSC disease modeling approach (Chapters 3, 4, and 6 of this thesis respectively). These studies demonstrate the characterization of three very different examples of disease-associated
genetic variation that emerge from next-generation human genetic studies. Briefly, 
SORT1 (Chapter 3 of this thesis) was identified by GWAS of blood lipid levels and 
subsequent in vitro characterization\(^{3-4}\), although uncertainty remained regarding its 
functional role in human cells\(^7\). Similarly, human genetic studies found that ANGPTL3 – 
the focus of Chapter 4 of this thesis – is a regulator of plasma lipid concentrations\(^{3,5}\), 
unlike SORT1 however, ANGPTL3 underlies a monogenic condition (FCH), which is 
inherited in a monogenic fashion\(^6\). The studies presented in Chapter 4 of this thesis 
were designed to investigate its contribution to the regulation of human LDL-C levels 
and ultimately provided insight into potential pitfalls of hPSC disease modeling. Finally, 
in Chapter 6 of this thesis, we focused on a research topic that is most centrally related 
to the overall objective of this thesis – how might hPSC disease modeling be applied to 
investigate common noncoding DNA variants identified by GWAS of disease-related 
traits. This is an especially important question because GWAS have identified many 
such variants, which are of unknown functional significance, but are tantalizing 
signposts that likely indicate the present of a novel disease-related gene, which has yet 
to be identified\(^8\). We applied the ASGR1-sorting approach described in Chapter 5 of 
this thesis to test an a priori hypothesis related to rs12740374\(^4\), clarifying its connection 
to SORT1 (the likely causal gene for the locus), and demonstrating the feasibility of 
studying a common noncoding SNP using hPSCs.

In summary, in this thesis we described the need for functional evaluation of 
human-genetic findings related to complex human diseases and discussed the potential utility of hPSCs for performing this type of characterization. We also described methods 

\(^*\) Common genetic variation near ANGPTL3 are also associated with blood lipid levels\(^3\), suggesting that common as well as rare variation in ANGPTL3 effect these levels in humans.
for addressing two major challenges facing hPSC disease modeling research, in particular an optimized approach for hPSC genome editing. Lastly, we demonstrated the application of isogenic hPSC disease modeling to address hypotheses arising from different types of human genetic studies. The results of this research should inform and enable a future stage of hPSC disease modeling.
FUTURE DIRECTIONS: THE 1P13 LOCUS AND HPSC MODELS OF GENETIC VARIATION

The connection between rs12740374 and SORT1 expression\(^4\) is a unique example of a long-range genetic interaction in which a disease-related genetic variant exerts a tissue-specific influence on the expression of a gene located many kilobases away on the same chromosome. As introduced above, experiments described in Chapter 6 of this thesis provide direct evidence in support of this mechanism, namely that the rs1240374 minor allele causes an increase in SORT1 gene expression in human hepatocyte-like cells (HLCs). These findings raise the possibility that similar mechanisms may be at play at other loci implicated by GWASs of disease-related traits. These studies also illustrate how isogenic hPSC modeling could be used to investigate genetic connections of this nature.

Having established an hPSC-based model of rs12740374 function, we have the opportunity to further examine the mechanism connecting rs12740374 to the regulation of SORT1 expression. Elucidating these details will provide a more complete molecular understanding of the rs12740374-SORT1 connection and may inform future studies aimed at characterizing the effects of non-coding DNA variants identified by GWAS. Furthermore, rs12740374 genome-edited cell lines could be used in additional proof-of-concept experiments related to human iPSC-based disease modeling. With this in mind, several experiments are proposed below.

The 1p13 locus: regulation at a distance

The experiments described in the Chapter 6 of this thesis demonstrate that the
putative C/EBP transcription factor binding site created by the rs12740374 minor allele functions to elevate the expression of SORT1 in HLCs. These experiments represented a test of sufficiency; HLCs differentiated from hiPSCs in which the rs12740374 minor allele had been deleted exhibited decreased SORT1 gene expression compared with matched wild-type control HLCs. Although this result and the results presented in Musunuru et al. (2010) firmly establish the connection between rs12740374 and SORT1 expression in human hepatocytes, a test of sufficiency would satisfy the remaining condition of the “genetics version of Koch’s Postulates” (introduced in Chapter 1 of this thesis) and lend additional support to the rs12740374-SORT1 causal connection. This could be accomplished in a straightforward fashion using genome editing and the methods described in this thesis. Specifically, genome editing could be used to introduce (“knock-in”) the rs12740374 minor allele into an hPSC line of the rs12740374 homozygous major genotype.

As described in Chapter 2 and illustrated in Chapters 3 and 4 of this thesis, a specific DNA variant can be introduced in a single selection-free genome-editing step. Of note, although technically feasible, this process is inefficient in comparison to the generation of indel mutations, which can be routinely achieved with high efficiency. Once generated, heterozygous rs12740374-minor knock-in cell lines could be re-targeted to generate homozygous knock-in cell lines, or differentiated directly to HLCs to assess the expression of SORT1. As described above, this experiment would test the sufficiency of the rs122740374 minor allele in terms of its ability to increase SORT1 expression and would serve as a compliment to the test of necessity described in Chapter 6 of this thesis. More generally, the ability to introduce specific DNA variants
into hPSCs by CRISPR/Cas9-mediated homology-directed repair will be of critical importance to future research aimed at characterizing specific DNA variants identified by human genetics studies. Therefore, additional methods or significant optimization of current techniques will ultimately be needed in order to increase the efficiency of knock-in generation and facilitate more advanced isogenic disease modeling in the future.

Returning to the proposed mechanism of action of rs12740374, our results support the following model: in the presence of the rs12740374 minor allele, a C/EBP transcription factor binds at the site of rs12740374 in hepatocytes, ultimately producing increased SORT1 expression. Although these aspects of the proposed model are well validated, it is currently unclear how binding of a C/EBP transcription factor at rs12740374 actually enhances the expression of SORT1, especially in light of the fact that rs12740374 is located many kilobases away from SORT1. Specifically, the SORT1 promoter region is more than 100kb away from rs12740374 in the chromosome 1p13 region. We suspect that there is a physical interaction that brings the rs12740374 site into proximity with the SORT1 promoter and that this interaction between distant regions of the chromosome occurs in cis. We propose several experiments that could be used to directly test the following hypothesis: a C/EBP transcription factor and the rs12740374 minor allele function via a long-range chromosomal interaction. This could be assessed using the chromosome conformation capture (3C) method and the set of isogenic rs12740374 mutant and control cell lines described in Chapter 6 of this thesis. Briefly, rs12740374 mutant and wild-type control hiPSCs would be differentiated into HLCs; genomic DNA would than be prepared for 3C from these cells as well as undifferentiated mutant and control hiPSCs. If successful, this 3C experiment would
indicate whether or not a long-range interaction exists between the rs12740374 region and the SORT1 promoter region. Furthermore, if there were an interaction, this experiment would suggest whether or not the interaction is cell-type specific (hPSCs vs. HLCs) and/or dependent on the presence of the rs12740374 site (mutant vs. wild-type control). In preliminary experiments we have designed and optimized a 3C assay for the 1p13 locus that can be used to detect potential interaction between the SORT1 promoter region and the region near rs12740374 (data not shown).

Provided that the 3C experiment described above demonstrates the existence of a long-range interaction at the 1p13 locus, additional studies could be performed to determine the importance of C/EBP transcription factor protein activity in this mechanism. These experiments would be performed using a dominant-negative C/EBP protein (“A-C/EBP”). This protein has been previously described and was used in the initial studies that demonstrated the enhancer activity of the rs12740374 minor allele. Expression of A-C/EBP in HLCs would be achieved by means of lentiviral transduction; this method was used to expression exogenous SORT1 in SORT1 mutant HLCs in Chapter 3 of this thesis. To assess whether C/EBP transcription factor binding is required for the establishment or maintenance of a 1p13 chromosomal interaction, the 3C experiment described above would be repeated, in this case with overexpression of A-C/EBP. We hypothesize that elevated SORT1 expression requires C/EBP transcription factor activity as well as the presence of a long-range interaction, although we suspect that the C/EBP transcription factor is not itself involved in establishing or maintaining the long-range interaction. Therefore, we expect that A-C/EBP expression would produce a decrease in SORT1 expression specifically in HLCs and that there will
be no change in the long-range interaction. The former would be assessed by qRT-PCR in wild-type HLCs with and without A-C/EBP expression. If successful, the experiments proposed above will establish whether the genetic interaction between rs12740374 and SORT1 expression is the result of a direct long-range chromosomal interaction and will also indicate whether C/EBP transcription factor activity is necessary for this interaction.

**Isogenic controls for iPSC cohort experiments**

As introduced in Chapter 1 of this thesis, in comparison with unmatched iPSC lines, isogenic mutant and control hPSC lines generated by genome editing are less likely to exhibit undesired phenotypic variability due to genetic background differences. This type of variability could significantly confound the interpretation of phenotypic assays performed with a set of iPSC lines. On the other hand, an iPSC-based experimental design could be used to examine the effects of genetic variants in unique and effective ways. For example, a cohort of human iPSCs could be differentiated into cell types that would otherwise be difficult to obtain (i.e. neurons or airway epithelial cells); this would make it possible to perform expression quantitative trait locus (eQTL) analyses *in vitro* on cell types that would otherwise be impossible or impractical to obtain from human tissue samples. However, the feasibility of this approach remains largely unknown. Towards this goal, the two distinct hPSC disease-modeling study designs (isogenic and iPSC-based/unmatched) could be combined to leverage the unique advantages of each approach. The isogenic disease modeling approach could be used to confirm whether a genotype-phenotype correlation identified with unmatched
iPSC lines can be attributed directly to a particular gene or DNA variant. Conversely, once a phenotype is identified using isogenic disease modeling, a set of iPSC lines derived from different individuals could then be used to establish whether the phenotype is also observed using unmatched cell lines (i.e. cell lines with different genetic backgrounds, including distinct genotypes at a specific variant of interest). If a phenotype is observed to be associated with a particular genotype, individual iPSC lines that exhibit extreme or unexpected phenotypes could be identified from within the set of cell lines and analyzed further. The goal of this analysis would be to identify modifier genes that interact genetically with the primary variant of interest. These hypothetical experiments illustrate ways in which isogenic disease modeling can be used in conjunction with a cohort of iPSC lines.

As a proof-of-concept, this combined approach could be applied in future experiments involving the 1p13 locus. Introduced briefly in the discussion section of Chapter 6 of this thesis, the set of isogenic rs12740374 mutant and wild-type cell lines discussed in that chapter could serve as “control” cell lines in an “iPSC cohort” experiment. In this type of experiment a set of iPSC lines from individuals of different rs12740374 genotypes would be differentiated into HLCs, and SORT1 expression level would be assessed. In light of the strong association between rs12740374 genotype and SORT1 expression level observed in human liver samples⁴, one would expect to see a similar association in the corresponding in vitro experiment. Isogenic rs12740374 mutant and wild-type control cell lines could also be differentiated and analyzed in the same experiment. The magnitude of association between rs12740374 genotype and SORT1 expression level should be at least as great in HLCs differentiated from the
isogenic cell lines as in HLCs differentiated from the unmatched iPSC lines. Thus, in this way the genome-edited cell lines would serve as “positive control” cell lines within an iPSC cohort experiment.
DISCUSSION

Genome editing technology

The studies described in this thesis raise several issues that warrant further discussion and suggest avenues for future research. Principle among the former are concerns related to potential off-target effects resulting from genome editing, which have arisen as a persistent counterbalance to considerable enthusiasm surrounding genome editing technology. The CRISPR/Cas system is an especially promising genome editing platform, yet as with other genome editing technologies, questions have been raised about its specificity and whether introducing genome-editing components into cells causes a significant number of unintended genomic modifications that could confound research studies and preclude their application for clinical uses.

In many ways the CRISPR/Cas system exemplifies the desired characteristics of a genome editing technology, including for applications in hPSCs. As a unique RNA-guided nuclease, the CRISPR/Cas system is more straightforward to customize than protein-based engineered nucleases (ZFNs and TALENs) and offers greater flexibility as well as increased activity in many systems. For these reasons the CRISPR/Cas system has rapidly become a prominent genome editing technology. General attributes of three commonly used genome-editing technologies are summarized in Table 7-1. The following discussion will primarily focus on CRISPR genome editing, although it must be said that the other platforms are viable options depending on the needs of a particular application.
Table 7-1. Summary of commonly used technologies for genome editing in human pluripotent stem cells

Key: major advantage, major disadvantage, notable application, emerging area

<table>
<thead>
<tr>
<th>TECHNOLOGY</th>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
</table>
| **ZFNs**  | • First widely used technology for genome editing  
• Successful applications demonstrated in a variety of systems and contexts  
• Moderate adaptability – fusion to other functional domains (especially for gene regulation)  
• Increased specificity by cooperative binding (obligate heterodimers)  
• Relatively small size  
• First therapeutic application of genome editing in clinical trials | • Challenging design (somewhat non-modular)  
• Challenging construction, despite available resources  
• Relatively greater toxicity in some systems  
**Off-target effects:**  
Some degree of off-target effects observed (comparable to TALENs) |
| **TALENs** | • Fairly straightforward design (somewhat flexible spacing and binding site design criteria)  
• Modular assembly  
• Availability of genome-wide libraries  
• Moderate adaptability – fusion to other functional domains (especially for gene regulation)  
• Increased specificity by cooperative binding (obligate heterodimers)  
• Successful applications demonstrated for hPSC disease modeling | • Multiple-step construction  
• Sequence complexity and large size (although sequence optimized versions are available and viral delivery has been demonstrated)  
• Difficult to multiplex  
**Off-target effects:**  
Some degree of off-target effects documented in several systems |
| **CRISPR/Cas9 (General)** | • Simple design and construction  
• Customized gRNA rather than protein  
• Relatively simple design criteria (target site requirements) | • Moderate size of Cas9 protein  
• However, smaller than ZFN and TALEN in total size (monomer vs. dimer) |
<table>
<thead>
<tr>
<th>CRISPR/Cas9 (General, continued)</th>
<th>(\textbf{CRISPR/Cas9})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Extremely high DSB efficiency</strong> (facilitating indel creation especially; HDR rate may be comparable to alternatives)</td>
<td>o Too big for convenient viral delivery – specifically, standard spCas9, but smaller versions are available (e.g. saCas9)</td>
</tr>
<tr>
<td><strong>Flexibility</strong> – amenable to:</td>
<td>• Target site flexibility (limited primarily by PAM locations)</td>
</tr>
<tr>
<td>o Multiplexing (simultaneous editing at multiple sites or single locus for deletions; even straightforward library construction and delivery)</td>
<td>o Generally only a concern for targeting of a specific DNA variant</td>
</tr>
<tr>
<td>o Inducible activation (e.g. RNA transfection into inducible-Cas9 stable lines) – “iCRISPR”</td>
<td>o Can be addressed by use of alternate PAMs or Cas9 from other species</td>
</tr>
<tr>
<td>o Multiple versions of Cas9 available (nickase, nuclease-dead)</td>
<td>• Significant concerns regarding off-target effects (see below)</td>
</tr>
<tr>
<td><strong>Extensive adaptability</strong> – functional domains can be linked to Cas9 or to gRNA via RNA binding protein</td>
<td><strong>Off-target effects:</strong></td>
</tr>
<tr>
<td>o Gene regulation (activation, repression, inhibition)</td>
<td><strong>Substantial off-target effects seen in some cell types</strong> (particularly commonly used cancer cell lines); however, very few examples of confirmed off-target effects in hPSC despite several studies</td>
</tr>
<tr>
<td>o Epigenetic modification</td>
<td></td>
</tr>
<tr>
<td>o Intracellular imaging/marking</td>
<td></td>
</tr>
<tr>
<td>• Additional options for increased specificity</td>
<td></td>
</tr>
<tr>
<td>o Cas9 nickase (see additional details below)</td>
<td></td>
</tr>
<tr>
<td>o Truncated gRNA (tru-gRNA) – can also be applied to Cas9 nickase</td>
<td></td>
</tr>
<tr>
<td>o Cas9 from other species and Cas9 with enhanced specificity</td>
<td></td>
</tr>
<tr>
<td>o Adapted to require cooperative binding (Cas9-FokI fusion); drawback – large size</td>
<td></td>
</tr>
<tr>
<td>• Binding specificity and off-target effects relatively well-studied compared to ZFNs and TALENs</td>
<td></td>
</tr>
<tr>
<td>o Multiple whole-genome sequencing studies</td>
<td></td>
</tr>
<tr>
<td>CRISPR/Cas9</td>
<td></td>
</tr>
</tbody>
</table>
| (General, continued) | o Published characterization of targeting specificity using different approaches and cell lines (including "guide-seq")
• High efficiency genome editing in hPSCs with scarce documented cases of off-target effects (see cons)
• Has been used to facilitate one-step bi-allelic gene targeting ("iKO")
• Has been used for in vitro screens
• Successful applications for in vivo genome editing (including multiplexed, germ line or somatic) |
| CRISPR/Cas9 | |
| (Single nickase) | • HDR favored over NHEJ at some loci
• Decreased off-target activity demonstrated (compared with fully active Cas9) |
| CRISPR/Cas9 | • NHEJ can still occur with some efficiency (including at off-target sites)
• Generally inefficient compared with other CRISPR approaches |
| Off-target effects: | Generally lower frequency relative to fully active Cas9 |
| CRISPR/Cas9 | |
| (Paired nickase) | • More efficient than single nickase for HDR and NHEJ; less off-target than single gRNA/Cas9 (however, possibly more off-target effects due to 2nd gRNA – see off-target effects discussion)
• Decreased off-target effects compared to single gRNA/Cas9 (see note in cons section) |
| Off-target effects: | • Spacing requirement limits target selection somewhat
• Possible additional off-target effects (perhaps even synergistic due to second gRNA) |
| Generally lower frequency relative to fully active Cas9 with single gRNA; however, the degree of additional off-target effects (potentially even synergistic) resulting from 2nd gRNA is not well established |
Potential off-target effects of genome editing

The possible impact of off-target effects resulting from CRISPR/Cas genome editing can be considered by dividing the topic into two important sections: detecting and subverting. Two other areas that are actively being investigated can be represented by the terms “predicting” and “preventing.” Although these are promising areas of ongoing development with respect to genome editing technology, they will not be discussed in detail in this chapter because issues related to ‘detecting’ and ‘subverting’ are more directly related to the studies described in this thesis.

The use of genome editing, to create hPSC disease models, has brought considerations of the potential confounding effects of off-target mutations to the forefront – especially considering the fact that a central purpose of the “isogenic” disease modeling approach is to avoid potentially confounding factors. Initially, several studies were published which caused considerable alarm by showing somewhat lax binding specific of CRISPRs and many instances of off-target mutagenesis, sometimes at high frequency. Although concerning, much of this work was conducted with transformed cell lines, including 293T cells, which may very well have a diminished capacity for DNA repair. Therefore, the generalizability of some of these results was unclear.

Efforts were undertaken to identify (detect) examples of off-target mutagenesis occurring in hPSCs, in the context of actual genome editing and disease modeling. A number of genome edited hPSC clones were analyzed by genomic DNA sequencing using methods ranging from targeted PCR sequencing to exome sequencing of significant depth, and whole-genome sequencing with moderate depth. The outcomes
of these analyses were presented in several publications\textsuperscript{11-12}. Collectively, these studies uncovered evidence of occasional off-target mutations attributable to TALENs and ZFNs, but found few or no examples of mutations that were likely caused by off-target CRISPR binding\textsuperscript{12}. However, notably, these studies observed a significant number of genetic differences between cell lines – even between hPSC clones generated from the same parental cell line – that were probably not the result of off-target nuclease activity, but instead likely arose during the derivation and expansion of the hPSC lines. This finding is consistent with previous research, which concluded that a significant number of single nucleotide variants (among other types of variation) arise spontaneously in culture\textsuperscript{13-14}. These results ultimately indicated that in the context of hPSC disease modeling, clonal heterogeneity between cell lines may be a more significant concern than the possibility of off-target effects resulting from engineered nucleases – although it should be noted that a relatively small number of clones and ZFNs/TALENs/CaRISPRs were included in these studies.

Given that most hPSC genome editing strategies (including the approach outlined in this thesis) include at least one step of hPSC sub-cloning as well as several passages in culture, genetic heterogeneity introduced by the genome editing process should be considered in the design and interpretation of experiments, irrespective of the likelihood of encountering nuclease-related off-target effects. Studies suggest that single-cell clone prior to generating a set of genome edited clones may reduce the number of background heterogeneity\textsuperscript{15-16}, although this may also increase the chances of encountering the emergence of karyotypic abnormalities with are well known to arise during hPSC culturing and sub-cloning\textsuperscript{17}. 

260
Controlling confounding factors in isogenic hPSC disease modeling

In light of findings described above, it would be prudent to plan strategies to subvert potentially confounding factors that arise during isogenic hPSC disease modeling, regardless of how they occur. Fortunately there are many experimental straggles that can be used for this purpose. Based on the relative simplicity of CRISPR/Cas genome editing, multiple mutant and control hPSC clones can be generated using distinct CRISPRs (different guide-RNAs) and included in phenotypic studies. Generating clones with different CRISPRs eliminates the possibility that an observed phenotype is due to an off-target mutation present in several clones, because off-target effects would be related to the different target sequences of the CRISPRs. Moreover, background mutations arising from culturing and sub-cloning should differ somewhat across multiple clones included in a particular study, so clonal heterogeneity could not fully account for a phenotype that is consistently observed with several different clones. Changing phenotypes in particular clones over time, or differences in growth rate or morphology between clones, could signal the presence of background mutations or other potential confounding factors; therefore, these clones should be excluded from further studies. In addition, whole-genome sequencing of individual clones can be performed to identify potential differences in genetic background, although this may be impractical due to time and cost constraints and could be uninformative if sequencing depth is insufficient.

Finally, perhaps the most informative strategy for mitigating or overcoming confounding effects in disease modeling studies is the classic approach of genetic reconstitution. If an observed phenotype is primarily due to a targeted genetic
modification being evaluated, reversing this modification should correct the observed phenotype. This can be accomplished by overexpression the wild-type copy of a gene that was mutated, or repairing the original mutation or modification through a second of genome editing. Both approaches have successfully been applied in hPSC studies, including in the studies presented in Chapter 3 of this thesis\textsuperscript{18-19}. In summary, in the context of hPSC disease modeling, efforts to detect off-target mutations resulting from genome editing have dampened concerns about nuclease-related mutagenesis and informed a variety of studies for overcoming potential confounding effects in general.

Overall, tremendous advancement in genome editing technology has enabled the development of isogeneic hPSC disease models particularly via the optimized stagey described in Chapter 2 of this thesis and similar approaches. The disease modeling studies contained in this thesis are a testament to the potential utility of this approach, which continues to be applied in many interesting and increasingly advanced ways to investigate the genetic underpinnings of human disease.
CONCLUSIONS AND FUTURE PROSPECTS

The overarching goal of this thesis research is the translation of emerging human genetics findings into molecular insights related to the pathogenesis of cardiovascular disease. Successfully doing so would not only improve our understanding of human pathophysiology, it would enable the development of new methods for the prevention or treatment of common human diseases. Perhaps at no previous time in the history of biomedical science has there been such an abundance of opportunities to advance our understanding of human disease and ultimately develop new treatments, effective preventative strategies, or even cures in the case of irreversible conditions. These opportunities are the result of continued advancement in many scientific disciplines related to biomedical research – especially cellular and molecular biology and genetics.

The studies performed in this thesis are derived from the fields of human genetics and pluripotent stem cell research. The pace of technological advancement in these fields over the past two decades has been striking. The availability of the complete human genome sequence and ultimately the development of next-generation DNA sequencing, genotyping, and bioinformatics technologies have provided unprecedented access to the genetic framework of human biological processes, including disease predisposition. Therefore, the challenge is to extract from this expanse of genetic data specific information that is useful for the purpose of predicting, preventing, or treating disease.

Progress in the field of pluripotent stem cell research has been similarly impressive. The first hESC lines were isolated less than 20 years ago. Since that time, hPSC research has progressed significantly, both in predictable and unexpected
directions. This success is in part attributable to a strong foundation of earlier developmental biology and pluripotency research.

Although potential research and clinical applications of hPSCs were apparent even before these cells were actually isolated and cultured, a large number of technical barriers had to be overcome to enable the practical utilization of hPSCs for research purposes and work toward the possibility of hPSC-based cellular-replacement therapies. These advances include: substantial improvements in basic culturing methods; the creation of protocols for *in vitro* differentiation of many mature cell types from hPSCs; fundamental characterization of hPSCs at the molecular, transcriptional, and epigenetic levels. In addition, several surprising and transformative discoveries and technical developments have emerged. Two especially important developments were (1) iPSC reprogramming of somatic cells by defined factors and (2) genome editing of human cells. The discovery and refinement of iPSC reprogramming has made it possible to model the vast genetic diversity of the human population *in vitro* in a way that was previously almost unimaginable. Similarly, genome-editing technology has provided an ability to make targeted genetic perturbations at the genomic DNA level – or even transcriptional or epigenetic changes – with unprecedented precision. The 2012 Nobel Prize in Physiology or Medicine was awarded to Sir John B. Gurdon and Shinya Yamanaka in recognition of the importance of iPSC reprogramming; perhaps a Nobel Prize will ultimately be awarded for the discovery of the CRISPR/Cas system and its adaptation for genome engineering.

Although genome editing approaches have substantial utility across practically all biological systems and even as future therapies, the combination of genome editing
technology with hPSCs is particularly powerful. Together, hPSCs and genome editing possess an almost ideal set of attributes as an *in vitro* model system for characterizing the genetic underpinnings of human disease. Specially, features related to hPSCs in general include: (1) an intact human genome – potentially of any desired genetic background via the use of human iPSCs – which enables the evaluation of next-generation genetic findings in the most relevant genetic context; (2) the ability to generate many different mature cell types, including those profoundly effected by human disease and/or otherwise unavailable for characterization; (3) the ability to investigate the possible contribution of cellular development or differentiation to disease pathogenesis. Furthermore, features related to genome editing of hPSCs include: (1) the ability to rigorously – and with a reasonable level of efficiency – draw causal connections between genes or genetic variants and molecular or cellular phenotypes using classical experiment principles (Koch’s postulates); (2) the ability to advance hPSC research by generating reporter cell lines or performing high-throughput knock-out screens to better understand pluripotency, cellular identify, and *in vitro* differentiation to facilitate future studies.

In summary, the scientific community – in particular the field of hPSC research – is better equipped than ever before to enter a new era of biomedical discovery.
REFERENCES


