Investigating Genetic Risk Factors of Coronary Artery Disease Using Genome Editing

Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

by

David Graham McVey BSc (Hons)

Department of Cardiovascular Sciences
University of Leicester

August 2015
Investigating Genetic Risk Factors of Coronary Artery Disease Using Genome Editing

David Graham McVey

Abstract

Genome-wide association studies (GWAS) have identified the genetic loci associated with many complex diseases including coronary artery disease (CAD). The challenge now is to elucidate the biological and cellular pathways affected by disease-associated loci. In order to fully understand the functional mechanisms, the causal genetic variants need to be identified. The majority of GWAS loci lack candidate genes, and may instead be located in regulatory regions, making the functional effects of specific variants difficult to appreciate. Recently, genome editing techniques have become available that allow targeted alteration of the genome, producing isogenic cell lines that differ only at the site of interest.

In this study, recombinant adeno-associated virus (rAAV) genome editing was established and used to investigate potentially functional disease-associated variants in the 1p13 and 9p21 CAD loci. Evidence from previous work suggests that 1p13 (rs12740374) and 9p21 (rs10811656 and rs10757278) single nucleotide polymorphisms (SNPs) affect transcription factor binding, leading to dysregulation of local genes. Specific alteration of these SNPs using this technique enabled the examination of these hypotheses directly.

The 1p13 study has provided evidence to support the hypothesis that rs12740374 is the causal SNP at this locus. We observed genotype-dependent effects upon C/EBPα binding and the expression of four 1p13 genes (SORT1, CELSR2, PSRC1 and MYBPHL).

Our examination of the 9p21 SNPs showed that these variants are capable of influencing STAT1 binding, but local gene expression was not affected. This suggests that variation of just rs10811656 and rs10757278 is insufficient to affect gene expression, and that other pathways may be involved.

The first study to utilise the rAAV technique to examine non-coding, regulatory SNPs, this work demonstrates that isogenic cell lines produced by rAAV genome editing allow for the quantification of subtle, genotype-specific effects. This work suggests that this adaptable technology may be beneficial for other studies examining the genetics of complex diseases.
Acknowledgements

I would like to thank, and express my appreciation of, my supervisors, Professor Sir Nilesh Samani, Dr Tom Webb and Dr Dave Lodwick for their support and guidance throughout this thesis. I would also like to thank Dr Veryan Codd for sharing her technical expertise. I am grateful for the opportunity to work with, and learn from, such excellent academics.

I would also like to thank the staff and fellow students in the Department of Cardiovascular Sciences, past and present, who have made my time within the department highly enjoyable, have provided help, support and advice, and have become great friends.

I also wish to thank and dedicate this thesis to my mum and grandparents who have provided life-long support, encouragement and motivation, and have been instrumental in the completion of this thesis.
# Table of Contents

Abstract ........................................................................................................................................... i  
Acknowledgements ............................................................................................................................ ii  
Table of Contents ............................................................................................................................... iii  
List of Tables ....................................................................................................................................... ix  
List of Figures ..................................................................................................................................... xi  
Abbreviations ..................................................................................................................................... xiii  
Chapter 1: Introduction ...................................................................................................................... 1  
1.1.1 Overview of Coronary Artery Disease ...................................................................................... 1  
1.1.2 Coronary Artery Disease: Pathophysiology of Atherosclerosis and Clinical Impact .......... 1  
1.2 Coronary Artery Disease Risk Factors ....................................................................................... 5  
1.3 The Genetics of Coronary Artery Disease .................................................................................... 5  
1.3.1 Assessment of Coronary Artery Disease Heritability ............................................................ 5  
1.3.2 The Genetic Causes of Coronary Artery Disease: Familial Hypercholesterolaemia & Candidate Gene Studies .......................................................................................... 6  
1.3.3 Identifying Novel Loci Associated with Coronary Artery Disease: Linkage Studies .......... 8  
1.3.4 Identifying Novel Loci Associated with Coronary Artery Disease: Genome-Wide Association Studies (GWAS) .......................................................................................... 9  
1.3.4.1 Introduction to GWAS ........................................................................................................ 9  
1.3.4.2 Coronary Artery Disease GWAS ....................................................................................... 10  
1.3.4.3 Understanding GWAS Associations and the Challenges of Functional Analyses ........ 12  
1.4 Functional Analysis of Disease-Associated Variants Using Genome Editing ..... 13  
1.4.1 “Protein Directed, Protein Mediated” Genome Editing Technologies ............................... 16  
1.4.1.1 Zinc Finger Nucleases ........................................................................................................ 17  
1.4.1.2 Transcription Activator-Like Effector Nucleases (TALENs) ........................................... 19  
1.4.2 “RNA Directed, Protein Mediated” Genome Editing Technology: CRISPR/Cas9 .............. 21  
1.4.3 “DNA Directed, Cell Mediated” Genome Editing Technologies ......................................... 24  
1.4.3.1 Overview of rAAV and the rAAV Genome Editing Procedure ..................................... 24  
1.4.3.2 Advantages and Limitations of rAAV Genome Editing .................................................. 25  
1.4.3.3 Example Uses of rAAV Technology ................................................................................... 28  
1.5: Experimental Approaches Taken in this Work ....................................................................... 31
1.6 Aims and Objectives ........................................................................................................... 32
Chapter 2: Materials and Methods .......................................................................................... 34
  2.1: Cell Lines and Culture Techniques .................................................................................. 34
  2.1.1: EA.hy926 ....................................................................................................................... 34
  2.1.2: HCT116 ......................................................................................................................... 35
  2.1.3: HEK293T ........................................................................................................................ 35
  2.1.4: HeLa-S3 .......................................................................................................................... 35
  2.1.5: HepG2 ................................................................................................................................ 35
  2.1.6: Mesenchymal Stem Cells (MSCs) .................................................................................. 35
  2.1.7: THLE-3 ........................................................................................................................... 35
  2.1.8: Passaging Procedures ..................................................................................................... 36
  2.1.9: Cell Counting .................................................................................................................. 36
  2.1.10: Cell Line Cryopreservation ........................................................................................... 36
  2.1.11: Cell Line Ageing .......................................................................................................... 37
  2.2: Parental Cell Line Characterisation ................................................................................... 38
  2.2.1: Proliferation Analysis Using SRB Assay ........................................................................ 38
  2.2.2: G418 Death Curve ......................................................................................................... 39
  2.2.3: rAAV2-GFP Infection .................................................................................................... 40
  2.2.4: Optimisation of Diagnostic PCR Assays for 1p13 and 9p21 Projects ............................. 40
  2.2.4.1 GoTaq PCR Reaction Mix ............................................................................................. 41
  2.2.4.2: Standard Touchdown PCR Protocol ......................................................................... 41
  2.2.4.3: Standard Gradient PCR Protocol ................................................................................. 42
  2.2.4.4: Standard Gel Electrophoresis Protocol ...................................................................... 42
  2.2.4.5: EA.hy926 9p21 Genomic Instability Determination ..................................................... 43
  2.2.5: SNP Genotyping by Sanger Sequencing ........................................................................ 44
  2.2.5.1: SNP Genotyping PCRs ................................................................................................. 44
  2.2.5.2: Sanger Sequencing Protocol ...................................................................................... 46
  2.2.6 SNP Genotyping by Microarray ...................................................................................... 47
  2.3: rAAV2 Production and Titre Estimation .......................................................................... 47
  2.3.1: Virus Production ............................................................................................................. 47
  2.3.2 Virus Titre Determination ................................................................................................. 49
  2.4: Generation of Isogenic Cell Lines by Genome Editing: Virus Infection, Antibiotic Selection and Cre Recombination of the Selection Cassette ........................................ 51
  2.4.1 Cell Line Infection .......................................................................................................... 51
2.4.2: Infected Cell Selection ........................................................................................................... 52
2.4.3 Selected Cell Colony Counting ............................................................................................... 53
2.4.4 Cell Consolidation and DNA Isolation .................................................................................... 53
2.4.5 Identification of Successfully Targeted Cell Pools by Diagnostic PCR Screen .......................... 54
2.4.6 Expansion of Diagnostic PCR Positive Pools .......................................................................... 57
2.4.7 Single Cell Dilution .................................................................................................................. 57
2.4.8 pML-Cre Plasmid Preparation and Selection Cassette Removal by pML-Cre Transfection ....... 59
2.4.8.1 pML-Cre Plasmid Transformation and Purification ............................................................... 59
2.4.8.2: Selection Cassette Removal by pML-Cre Transfection ....................................................... 60
2.4.9 PCR Screening of pML-Cre Transfected Single-Cell Derived DNA Samples ....................... 61
2.4.10 SNP Genotyping by Restriction Endonuclease Digestion ....................................................... 64
2.5 RNA Isolation Using Qiagen RNeasy Mini Kit ........................................................................... 65
2.6 SuperScript III Reverse Transcription ......................................................................................... 66
2.7 SensiFAST cDNA Synthesis Reverse Transcription ..................................................................... 67
2.8 Design and Optimisation of qPCR Assays ................................................................................... 67
2.9 1p13 Gene Expression Analysis .................................................................................................. 70
2.10 9p21 Gene Expression Analysis ................................................................................................ 72
2.10.1 Interferon-γ Treatment ........................................................................................................... 72
2.10.2 RNA Isolation of Interferon-γ Treated Cells With Zymo Direct-Zol RNA Miniprep kit ......... 73
2.10.3 Reverse Transcription of RNA Derived from Interferon-γ Treated Cells ......................... 73
2.10.4 9p21 qPCR Assays and Analysis ......................................................................................... 73
2.10.4.1 qPCR Reaction Setup ........................................................................................................ 74
2.10.4.2 qPCR Data Analysis .......................................................................................................... 74
2.11 1p13 Chromatin Immunoprecipitation (ChIP) ......................................................................... 78
2.12 9p21 Chromatin Immunoprecipitation (ChIP) ......................................................................... 81
2.13 ChIP-qPCR Analysis of Transcription Factor Binding at the 1p13 and 9p21 Loci ....................... 82

Chapter 3: Parental Cell Line Characterisation and Optimisation of rAAV Genome Editing Protocols ................................................................................................................................. 85
3.1 Introduction .................................................................................................................................... 85
3.1.1 Overview of rAAV Genome Editing ....................................................................................... 85
3.1.2 rAAV Vector Design Considerations and the Importance of Parental Cell Line Characterisation ................................................................................................................................. 85
4.2.8 Quality Control of Genome-Edited and “Aged” Controls: Proliferation Comparison ................................................................. 130
4.2.9 Assessment of SNP rs12740374 Functionality: Quantifying C/EBPα Transcription Factor Binding by Chromatin Immunoprecipitation-qPCR (ChIP-qPCR) .................................................................................. 133
4.2.10 Assessment of SNP rs12740374 Functionality: 1p13 Gene Expression Analysis ......................................................................... 134
4.3 Discussion ................................................................................................................................. 136
Chapter 5: Functional Analysis of the 9p21 Locus Using rAAV-Mediated Genome Editing ........................................................................ 143
5.1 Introduction ............................................................................................................................. 143
5.1.1 Overview of the 9p21 Coronary Artery Disease-Associated Locus ............................................. 143
5.1.2: Functional Analysis of the 9p21 CAD-Associated Locus ....................................................... 145
5.1.2.1 CDKN2A & CDKN2B ........................................................................................................... 145
5.1.2.2 ANRIL ................................................................................................................................. 147
5.1.2.3 Summary of the 9p21 Genes & Their Association With CAD .............................................. 148
5.1.3 Putative Functional Variants at 9p21 .................................................................................... 149
5.1.4 Hypothesis and Experimental Methodology ........................................................................ 152
5.2 Results ........................................................................................................................................ 153
5.2.1 Vector Design ....................................................................................................................... 153
5.2.2 Efficiency of Targeting and Cre Recombination .................................................................... 154
5.2.2.1 Generation of Heterozygous HeLa-S3 Cell Lines ................................................................. 156
5.2.2.2 Generation of Heterozygous HCT116 Cell Lines ................................................................. 161
5.2.2.3 Generation of Homozygous Risk HCT116 Cell Lines ......................................................... 165
5.2.2.4 Complete List of Clonal Cell Lines Obtained for Functional Analysis of the 9p21 SNPs rs10811656 and rs10757278 ...................................................................................................................... 168
5.2.2.5 Confirmation of Successful Targeting by Sequencing rs10811656 and rs10757278 ................................................................................................................................ 170
5.2.2.6 Microarray Genotyping of HeLa-S3 and HCT116 Cell Lines ................................................ 173
5.2.2.7 rs10811656 & rs10757278 Targeting in Mesenchymal Stem Cells (MSCs) ............... 173
5.2.3 Proliferation Comparison: HeLa-S3 ..................................................................................... 173
5.2.4 Proliferation Comparison: HCT116 ....................................................................................... 175
5.2.5 Assessment of SNP rs10811656 & rs10757278 Functionality: Quantifying STAT1 Transcription Factor Binding by Chromatin Immunoprecipitation-qPCR (ChIP-qPCR) ................................................................................... 177
List of Tables

Table 1.1: The Current, Published Loci Associated with CAD by GWAS .................. 11
Table 2.1: G418 Dilution Series used for G418 Death Curve .................................. 39
Table 2.2: Standard Touchdown PCR Protocol for use with GoTaq Hot Start DNA Polymerase .......................................................... 41
Table 2.3: Standard Gradient PCR Protocol for use with GoTaq Hot Start DNA Polymerase .......................................................... 42
Table 2.4: Cycling Conditions for PCR Assays to Assess EA.hy926 9p21 Genomic Instability ........................................................................... 44
Table 2.5: rs12740374 Sequencing PCR Cycling Conditions .................................. 45
Table 2.6: rs10811656/rs10757278 Sequencing PCR Cycling Conditions ................ 45
Table 2.7: Sanger Sequencing Cycling Protocol ....................................................... 46
Table 2.8: G418* Virus Titre Assay Standard Curve ................................................ 50
Table 2.9: G418 Virus Titre Assay Cycling Conditions ............................................. 51
Table 2.10: Cell Line Seeding Densities and G418 Selection Concentrations ...... 52
Table 2.11: Cycling Conditions and Expected Bands for 1p13 and 9p21 Diagnostic PCR Assays .......................................................... 56
Table 2.12: Expected Band Sizes (with Primer Combinations) and Primer Sequences for Cre PCR Screening .................................................. 63
Table 2.13: Cycling Conditions for 1p13 Cre PCR Assay ......................................... 63
Table 2.14: Cycling Conditions for 9p21 Cre PCR Assay ......................................... 64
Table 2.15: SNP Genotyping by Restriction Digestion Procedure .............................. 65
Table 2.16: Primer Sequences of 1p13 qPCR Assays with 36B4 Reference Assay ...... 70
Table 2.17: 1p13 qPCR Assay Final Primer Concentrations and Cycling Conditions... 72
Table 2.18: 9p21 qPCR Assay Primer Sequences .................................................... 76
Table 2.19: qPCR Reaction Mixes and Cycling Conditions for 9p21 Assay .......... 77
Table 2.20: Primer Sequences for 1p13 and 9p21 ChIP-qPCR Assays ...................... 83
Table 3.1: SNP Genotype of Cell Lines and Primary MSCs Extracted from Multiple Umbilical Cord Donors, Determined by Sanger Sequencing ......................................... 102
Table 4.1: 1p13 SNPs Associated with CAD and LDL Cholesterol Levels .......... 108
Table 4.2: Summary of Published 1p13 CAD/LDL GWAS Associations .................. 109
Table 4.3: Liver eQTL Summary for 1p13 SNPs and CELSR2, PSRC1 and SORT1 .. 116
Table 4.4: Monocyte and Macrophage eQTLs for 1p13 Genes - Best eSNPs .......... 117
Table 4.5: GWAS SNP eQTL p-values for SORT1, CELSR2, and PSRC1 for Monocytes and Macrophages ................................................. 117
Table 4.6: Summary of Transduction Experiments Targeting 1p13 in HepG2 Cells... 126

Table 4.7: List of HepG2 Lines Obtained via rAAV Genome Editing of rs12740374 128

Table 5.1: Summary of 9p21 Targeting Attempts in Order to Obtain Heterozygous HeLa-S3 Cells from Parental, Homozygous Non-Risk Cells............................... 159

Table 5.2: Summary of Single Cell Dilution Procedure and Observed Efficiency for Heterozygous HeLa-S3 Cell Line Development ........................................... 160

Table 5.3: pML-Cre Recombination Procedure Summary and Observed Efficiencies for HeLa-S3 Heterozygous Cell Lines......................................................... 160

Table 5.4: Summary of 9p21 Targeting Attempts in Order to Obtain Heterozygous HCT116 Cells from Parental, Homozygous Non-Risk Cells ................................ 163

Table 5.5: pML-Cre Recombination Procedure Summary and Observed Efficiencies for HCT116 Heterozygous and LoxP Control Cell Lines .................................. 163

Table 5.6: Procedural Summary of Cre Transfection and Subsequent Screening PCR Assay Efficiency for HCT116 Homozygous Risk and Aged Control Lines ............ 167

Table 5.7: List of HeLa-S3 Lines Obtained by Genome Editing and Aged Controls .. 168

Table 5.8: List of HCT116 Heterozygous Lines Obtained by Genome Editing and Aged Controls.................................................................................................. 169

Table 5.9: List of HCT116 Homozygous Risk Lines Obtained by Genome Editing and Aged Controls ................................................................................................. 170

Table 5.10: Calculated Doubling Time of HCT116 Cell Lines from SRB Proliferation Assay Data ........................................................................................................ 177
List of Figures

Figure 1.1: Diagram of Atherosclerotic Development and Plaque Rupture .................. 2
Figure 1.2: Homologous Recombination (HR) and Non-Homologous End Joining (NHEJ) Repair Pathways ................................................................. 16
Figure 1.3: Structure and DNA Binding of Zinc Finger Nucleases and TALENs ....... 20
Figure 1.4: CRISPR/Cas9 Genome Editing System with a Single gRNA ............... 23
Figure 3.1: Overview of rAAV Genome Editing Procedure ................................ 93
Figure 3.2: Diagnostic PCR and Cre PCR Screening Strategies .......................... 94
Figure 3.3: Sulphorhodamine B (SRB) Cell Proliferation Assay for EA.hy926, HeLa-S3 and HepG2 cells over 6 days Post-Seeding ............................................. 97
Figure 3.4: Fluorescence Microscopy of rAAV2-GFP Transduced Cells ............ 99
Figure 3.5: G418 Death Curve Phase Contrast Microscopy ................................ 101
Figure 3.6: Sequence Trace Data for HCT116, HeLa-S3 and HepG2 DNA for the Loci of Interest ........................................................................................................ 103
Figure 3.7: PCR Analysis Indicates Aged EA.hy926 Cells Display Genomic Instability at 9p21 .......................................................... 105
Figure 4.1: The Genes and SNPs of Interest in the 1p13 Region ......................... 110
Figure 4.2: Vector Design for rs12740374 Knock-In, with Annotation .............. 121
Figure 4.3: 1p13 Left Homology Arm Diagnostic PCR Screen of HepG2 Cells .... 123
Figure 4.4: 1p13 Cre PCR Screening of Targeted HepG2 Cells Transfected with pML-Cre ................................................................. 125
Figure 4.5: Parental and Targeted HepG2 Genotyping of rs12740374 by Sanger Sequencing ......................................................................................... 129
Figure 4.6: Proliferation Assay for a Subset of the HepG2 Genome-Edited and “Aged” Control Lines Compared to Parental Cells ........................................ 132
Figure 4.7: rs12740374 Genotype Influences C/EBPα Binding ....................... 133
Figure 4.8: Expression Analysis of SORT1, CELSR2 and PSRC1: Comparison of rs12740374 Homozygous Risk and Heterozygous Isogenic Cell Lines ........ 135
Figure 4.9: Expression Analysis of SARS, MYBPHL, PSMA5, and WDR47: Comparison of rs12740374 Homozygous Risk and Heterozygous Isogenic Cell Lines ......... 135
Figure 5.1: Genes and Key SNPs of the 9p21 Region ......................................... 144
Figure 5.2: Schematic Representation of 2nd Generation 9p21 Targeting Vector with shRNA Directed Against the G418 Resistance Gene ....................... 155
Figure 5.3: HeLa-S3 9p21 Left Homology Arm Diagnostic PCR ...................... 161
Figure 5.4: HeLa-S3 9p21 Cre Screening PCR ................................................ 161
Figure 5.5: 9p21 Left Homology Arm HCT116 Diagnostic PCR for Generation of Heterozygous (Single Knock-in) Cells ................................................................. 164
Figure 5.6: 9p21 Cre PCR Screening of Heterozygous HCT116 Cells ...................... 164
Figure 5.7: HCT116 9p21 Left Homology Arm Diagnostic PCR for Generation of Homozygous Risk (Double Knock-in) Lines .............................................. 166
Figure 5.8: Representative Agarose Gel Image of HCT116 9p21 Cre PCR for Generation of Homozygous Risk (Double Knock-in) Lines .............................................. 166
Figure 5.9: Sanger Sequencing of Parental (Homozygous Non-Risk) and Heterozygous HeLa-S3 Lines to Confirm SNP Genotype Editing .................................................. 171
Figure 5.10: Sanger Sequencing of Parental, Heterozygous and Homozygous Risk HCT116 Lines to Confirm SNP Genotype Editing .................................................. 172
Figure 5.11: Proliferation Assay for a Subset of the HeLa-S3 Genome-Edited and “Aged” Control Lines Compared to Parental Cells .................................................. 174
Figure 5.12: Proliferation Assay for a Subset of the HCT116 Genome-Edited and “Aged” Control Lines Compared to Parental Cells .................................................. 176
Figure 5.13: Relative STAT1 Binding at 9p21 SNPs of Interest in HeLa-S3 and HCT116 cells ................................................................................................................. 178
Figure 5.14: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for IRF-1, CDKN2A (p16) and CDKN2A (p14/ARF) .................. 180
Figure 5.15: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for CDKN2B, ANRIL (exons 1-5) and ANRIL (exons 4-6) ........... 181
Figure 5.16: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for ANRIL (exons 17-18), ANRIL (exons 18-19) and MTAP ........ 182
Figure 5.17: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for Total IFN-α and IFN-β ......................................................... 183
**Abbreviations**

* may refer to a gene or the corresponding protein; standard nomenclature is applied within the text (all capitals, italicised = human gene; all capitals, not italicised = human protein; initial capital, remaining lower case, italicised = mouse gene; initial capital, remaining lower case, not italicised = mouse protein).

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>A549</td>
<td>Human lung carcinoma cell line</td>
</tr>
<tr>
<td>AAT</td>
<td>Alpha-1 antitrypsin</td>
</tr>
<tr>
<td>AAV</td>
<td>Adeno-associated virus</td>
</tr>
<tr>
<td>ABCA1*</td>
<td>ATP-binding cassette, sub-family A, member 1</td>
</tr>
<tr>
<td>ADFP*</td>
<td>Adipose differentiation-related protein</td>
</tr>
<tr>
<td>ADIPOR1*</td>
<td>Adiponectin receptor 1</td>
</tr>
<tr>
<td>AMD</td>
<td>Age-related macular degeneration</td>
</tr>
<tr>
<td>ANRIL*</td>
<td>Antisense RNA in the INK4 locus (CDKN2BAS)</td>
</tr>
<tr>
<td>AP</td>
<td>Angina pectoris</td>
</tr>
<tr>
<td>APOB*</td>
<td>Apolipoprotein B</td>
</tr>
<tr>
<td>ARF*</td>
<td>Alternate reading frame (alternative name for CDKN2A variant p14)</td>
</tr>
<tr>
<td>aSMC</td>
<td>Aorta-derived smooth muscle cells</td>
</tr>
<tr>
<td>ATCC</td>
<td>American Type Culture Collection</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial artificial chromosome</td>
</tr>
<tr>
<td>BEGM</td>
<td>Bronchial epithelial cell growth media</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair(s)</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine serum albumin</td>
</tr>
<tr>
<td>BsiHKAI</td>
<td>Restriction enzyme from Bacillus stearothermophilus</td>
</tr>
<tr>
<td>bZIP</td>
<td>Basic leucine zipper</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>C/EBP</td>
<td>CCAAT/enhancer binding protein family or specific protein domain</td>
</tr>
<tr>
<td>C/EBPα*</td>
<td>CCAAT/enhancer binding protein alpha</td>
</tr>
<tr>
<td>C/EBPβ*</td>
<td>CCAAT/enhancer binding protein beta</td>
</tr>
<tr>
<td>CAD</td>
<td>Coronary artery disease</td>
</tr>
<tr>
<td>CARD8*</td>
<td>Caspase recruitment domain family member 8</td>
</tr>
<tr>
<td>Cas</td>
<td>CRISPR-associated sequence, protein component of the CRISPR system</td>
</tr>
<tr>
<td>Cas9</td>
<td>CRISPR-associated sequence 9, protein component of the CRISPR system of Streptococcus pyogenes</td>
</tr>
<tr>
<td>CASP14 *</td>
<td>Caspase 14</td>
</tr>
<tr>
<td>CBX7*</td>
<td>Chromobox homologue 7</td>
</tr>
<tr>
<td>CDK6*</td>
<td>Cyclin dependent kinase 6</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>Cyclin dependent kinase inhibitor 2A, p14/alternate reading frame variant</td>
</tr>
<tr>
<td>CDKN2A (p16)*</td>
<td>Cyclin dependent kinase inhibitor 2A, p16 variant</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>CDKN2B (p15)*</td>
<td>Cyclin dependent kinase inhibitor 2B, p15 variant</td>
</tr>
<tr>
<td>CDKN2BAS*</td>
<td>Cyclin-dependent kinase inhibitor 2B anti-sense (ANRIL)</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CELSR</td>
<td>Cadherin, EGF LAG Seven-Pass G-Type Receptor; may refer to protein family or protein domain unique to this family</td>
</tr>
<tr>
<td>CELSR2*</td>
<td>Cadherin, EGF LAG Seven-Pass G-Type Receptor 2</td>
</tr>
<tr>
<td>CEU</td>
<td>Northern Europeans from Utah; a HapMap project population</td>
</tr>
<tr>
<td>CF</td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td>CFH*</td>
<td>Complement factor H</td>
</tr>
<tr>
<td>CFTR*</td>
<td>Cystic fibrosis transmembrane conductance regulator</td>
</tr>
<tr>
<td>CHART</td>
<td>Capture hybridisation analysis of RNA targets</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>ChIP-qPCR</td>
<td>Chromatin immunoprecipitation-quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin immunoprecipitation-sequencing</td>
</tr>
<tr>
<td>Chr</td>
<td>Chromosome</td>
</tr>
<tr>
<td>CK1ε*</td>
<td>Casein kinase 1 epsilon</td>
</tr>
<tr>
<td>c-Myc*</td>
<td>V-myc avian myelocytomatosis viral oncogene homologue</td>
</tr>
<tr>
<td>COL1A1*</td>
<td>Collagen, type I, alpha 1</td>
</tr>
<tr>
<td>CRCs</td>
<td>Conditionally reprogrammed cells</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered regularly interspaced palindromic repeats</td>
</tr>
<tr>
<td>crRNA</td>
<td>CRISPR ribonucleic acid</td>
</tr>
<tr>
<td>CsCl</td>
<td>Caesium chloride</td>
</tr>
<tr>
<td>C&lt;sub&gt;T&lt;/sub&gt;</td>
<td>Cycle threshold</td>
</tr>
<tr>
<td>CXCL10*</td>
<td>Chemokine C-X-C motif ligand 10</td>
</tr>
<tr>
<td>DHR59*</td>
<td>Dehydrogenase/reductase (SDR) family member 9</td>
</tr>
<tr>
<td>DHS</td>
<td>DNaseI hypersensitivity site</td>
</tr>
<tr>
<td>DKI</td>
<td>Double knock-in</td>
</tr>
<tr>
<td>DMD</td>
<td>Duchenne muscular dystrophy</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco's modified Eagle media</td>
</tr>
<tr>
<td>DMRTA1*</td>
<td>Doublesex- and Mab-3-related transcription factor A1</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethylsulphoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNase</td>
<td>Deoxyribonuclease</td>
</tr>
<tr>
<td>dNTPs</td>
<td>Deoxyribose nucleoside triphosphates</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double stranded DNA</td>
</tr>
<tr>
<td>DSB</td>
<td>Double strand break</td>
</tr>
<tr>
<td>DSBR</td>
<td>Double strand break repair</td>
</tr>
<tr>
<td>DUT*</td>
<td>Deoxyuridine triphosphatase</td>
</tr>
<tr>
<td>E1FAY*</td>
<td>Mammalian target of rapamycin (mTOR)</td>
</tr>
<tr>
<td>E2F1*</td>
<td>E2F transcription factor</td>
</tr>
<tr>
<td>EA.hy926</td>
<td>Hybridoma of primary human endothelial cells and A549 (lung carcinoma cells)</td>
</tr>
<tr>
<td>EB1*</td>
<td>End binding protein 1</td>
</tr>
</tbody>
</table>
HPRT*  Hypoxanthine-guanine phosphoribosyltransferase
HR  Homologous recombination
hRP65*  Human retinal pigment epithelium-specific protein, 65kDa
hTERT*  Human telomerase reverse transcriptase
HUES  Human embryonic stem cells
Huh7  Human liver hepatocellular carcinoma cell line
HUVEC  Human umbilical vein endothelial cell
IFNA21*  Interferon-α 21
IFNGR1*  Interferon-γ receptor 1
IFNGR2*  Interferon-γ receptor 2
IFN-α*  Interferon-α
IFN-α21*  Interferon-α 21
IFN-β*  Interferon-β
IFN-γ*  Interferon-γ
IgG*  Immunoglobulin G
IMS  Industrial methylated spirit
INK4a*  Inhibitor of CDK4 A
INK4b*  Inhibitor of CDK4 B
iPSC  Induced pluripotent stem cells
IRF-1*  Interferon regulatory factor 1
ITR  Inverted terminal repeat
JAK*  Janus kinase
kb  Kilobases
KI  Knock-in
Kif2a*  Kinesin heavy chain member 2A
KO  Knock-out
LB  Luria broth
LCLs  Lymphoblastoid cell lines
LDL  Low density lipoprotein(s)
LDL-C  Low density lipoprotein cholesterol
LDLR*  Low-density lipoprotein receptor
lncRNA  Long, non-coding ribonucleic acid
LOD  Logarithm of the odds
LOD*  Lipoprotein lipase
mASS*  Murine argininosuccinate synthase
MCAK*  Mitotic centromere-associated kinesin
MDM2*  Mouse double minute 2 protein, human homologue, E3 ubiquitin ligase
MEDI*  Mediator complex subunit 1
MEF2A*  Myocyte enhancer factor 2A
MEFs  Mouse embryonic fibroblasts
MI  Myocardial infarction
miR-21*  Micro ribonucleic acid-21
miR-99a*  Micro ribonucleic acid-99a
miR-449a* Micro ribonucleic acid-449a
miRNA Micro ribonucleic acid
MMP13* Matrix metallopeptidase 13
MOI Multiplicity of infection
mOTC* Murine ornithine carbamoyltransferase
MSC Mesenchymal stem cell
MTA Methylthioadenosine
MTAP* Methylthioadenosine phosphorylase
mTOR* Mammalian target of rapamycin (E1FAY)
MYBPHL* Myosin binding protein H-like
NaCl Sodium chloride
NHEJ Non-homologous end joining
NHS National Health Service
NIHR National Institute for Health Research
NO Nitric oxide
NPC Nuclear pore complex
nt Nucleotide(s)
NTC No template control
p22phox* Human neutrophil cytochrome b light chain
p300* p300 E1A binding protein
pAAV rAAV vector plasmid
PAM Protospacer-adjacent motif
PBS Phosphate-buffered saline
PCR Polymerase chain reaction
PCSK9* Proprotein covertase subtilisin/kexin 9
PDCD4* Programmed Cell Death 4
PGK* Phosphoglycerate kinase 1 promoter, from Mus musculus
pML-Cre Cre-expressing plasmid with pML backbone
PNACL Protein and Nucleic Acids Chemistry Laboratory, University of Leicester, UK
PPARD* Peroxisome proliferator-activated receptor delta
PRC* Polycomb repressive complex
PRC2* Polycomb repressive complex 2
PSMA5* Proteasome (prosome, macropain) subunit, alpha type, 5
PSRC1* Proline/Serine rich coiled-coil 1
p-value Probability value
qPCR Quantitative polymerase chain reaction
R Reverse primer
r² Correlation coefficient squared; coefficient of determination
rAAV Recombinant adeno-associated virus
rAAV2 Recombinant adeno-associated virus serotype 2
Rat1-R12 Rat fibroblasts, clone R12 (from Rattus norvegicus)
RFP* Red fluorescent protein
RNA Ribonucleic acid
RNAi          RNA interference
RNase         Ribonuclease
rs            Reference SNP ID
RT            Reverse transcriptase/reverse transcription
SARS*         Seryl-tRNA synthetase
SCD           Single cell dilution
SDS           Sodium dodecyl sulphate
SDSA          Synthesis-dependent strand annealing
SEM           Standard error of the mean
sgRNA         Synthetic guide ribonucleic acid
shRNA         Small hairpin ribonucleic acid
siRNA         Small interfering ribonucleic acid
SNAP          SNP Annotation and Proxy Search, Broad Institute, USA
SNP           Single nucleotide polymorphism
SORT1*        Sortilin-1
SRB           Sulphorhodamine B
ssDNA0        Single stranded DNA
STAT1*        Signal transducer and activator of transcription 1
SUZ12*        Suppressor of Zeste homologue 12 (PRC2 subunit)
T             Thymine
TALE          Transcription activator-like effector
TALEN         Transcription activator-like effector nuclease
TCA           Trichloroacetic acid
TF            Transcription factor
TFBS          Transcription factor binding site
THLE-3        SV40 T-antigen transformed human liver epithelial cells, clone 3
Tm            Melting temperature
TMEM167B*     Transmembrane protein 167B
TNTC          Too numerous to count
tracrRNA      trans-activating CRISPR ribonucleic acid
Tris-HCl      Trisaminomethane hydrochloride
UCSC          University of California Santa Cruz
VAMP3*        Vesicle-associated membrane protein 3
VP1/2/3       Viral protein 1/2/3
VSMCs         Vascular smooth muscle cells
WDR47*        WD repeat domain 47
WHO           World Health Organisation
WJ-MSC        Wharton's jelly-derived mesenchymal stem cell(s)
WT            Wild-type
WTCCC         Wellcome Trust Case Control Consortium
YRI           Yoruba in Ibadan, Nigeria; a HapMap project population
ZeoR          Zeocin resistance gene
ZFN           Zinc finger nuclease
α-MEM         Minimum essential media, α modification
Chapter 1: Introduction

1.1.1 Overview of Coronary Artery Disease
Coronary artery disease (CAD) is a significant source of mortality worldwide. In 2011, approximately 7.4 million deaths were attributed to CAD across the globe, accounting for 13% of all deaths in this year (World Health Organisation [WHO], 2015). Not only does CAD cause a substantial number of deaths each year, the economic cost of caring for patients is high. It is estimated that CAD costs the UK NHS £1.8 billion per year (based upon data from 2009). Extrapolating to account for additional costs, such as informal care and, for those patients of working age, lost time from work, the cost to the UK economy has been estimated to be £6.7 billion per annum (Townsend et al. 2012).

The coronary arteries are the vessels that supply the myocardium with blood. In CAD patients, atherosclerosis of the coronary arteries over time eventually leads to the development of plaques. The initial stages of atherosclerosis have been observed in infants, with more complex lesions developing in young adults (reviewed by Stary, 2000). Clinical manifestation of this disease, such as myocardial infarction (MI) and angina pectoris (AP), occur many years (often decades) after the atherosclerotic process begins.

1.1.2 Coronary Artery Disease: Pathophysiology of Atherosclerosis and Clinical Impact
The coronary arteries are composed of three distinct layers (described by Libby et al. 2011, shown in figure 1.1a). The lumenal surface of human coronary arteries comprises a single layer of endothelial cells which reside on a basement membrane. Beneath the basement membrane, vascular smooth muscle cells (VSMCs) are located. These three sub-layers are referred to collectively as the tunica intima. This layer provides a barrier between the blood and the other layers of the coronary artery. Further, the endothelial cells in this layer help regulate haemodynamics in the vessel, through secretion of multiple factors, including vasodilators, vasoconstrictors, and anti-thrombotic factors (reviewed by Cines et al. 1998).

The second layer is composed principally of VSMCs within an extracellular matrix. This layer, termed the tunica media, is responsible for maintaining vascular tone. The VSMCs in this layer respond to vasodilatory and vasoconstrictory stimuli in order to maintain appropriate blood pressure within the coronary artery.
Figure 1.1: Diagram of Atherosclerotic Development and Plaque Rupture. A) The normal coronary artery is composed of three layers; the tunica intima, tunica media and adventitia. The tunica intima comprises a single layer of endothelial cells on a basement membrane, with a layer of vascular smooth muscle cells (VSMCs) underneath. The tunica media is composed primarily of VSMCs in an extracellular matrix, whilst the adventitia is composed of many cell types, including mast cells and fibroblasts. B) Receptor expression on the surface of activated endothelial cells permits immune cells, such as monocytes, to adhere to the endothelial cells. The monocytes migrate into the VSMC layer of the intima and differentiate into macrophages. These macrophages take up lipid, resulting in foam cell formation. C) Accumulation of foam cells and cytokine-induced VSMC migration and proliferation from the tunica media leads to plaque formation in the intima. Apoptotic foam cells and macrophages release lipid and cholesterol crystals into the plaque, forming a necrotic core. The plaque is stabilised by a fibrous cap, composed principally of collagen, which is produced by the VSMCs. D) Rupture of the fibrous cap releases the contents of the necrotic core into the lumen, inducing thrombosis formation that may lead to a myocardial infarction. Taken from Libby et al. (2011).
The third layer, the adventitia (reviewed by Majesky et al. 2011), is composed of a number of cell types and structures. These include fibroblasts, mast cells, nerves, macrophages, adipocytes, pericytes, T cells, haematopoietic progenitor cells and microvessels (vasa vasorum). The adventitia serves multiple purposes. The vasa vasorum provides the tunica media with a blood supply, whilst the immune and progenitor cells have been shown to be involved in vascular remodeling. Further, the adventitia may also interact with the surrounding tissue in which the coronary artery resides, mediating signalling between them.

Atherosclerosis (reviewed by Lusis, 2000; Homeister & Willis, 2010; Libby et al. 2011; Hopkins, 2013) results from injury to the endothelium due to either mechanical shear stress or chemical factors. This results in inflammation that drives progression of the atherosclerotic plaque.

Atherosclerosis is initiated by the adhesion of monocytes and other immune cells, such as lymphocytes, to the endothelium. This occurs due to the expression of cell surface markers that are recognised by antigens on the surface of the adhering cells. This activation has been shown to be caused by endothelial dysfunction. In healthy arteries, the endothelium helps maintain vascular tone by secreting vasodilators (such as nitric oxide, NO) and vasoconstrictors, including endothelin and angiotensin II (Davignon & Ganz, 2004). Alterations to these homeostatic mechanisms elicit a dysfunctional endothelial phenotype. Such dysfunction, initiated by shear stress, dyslipidaemia and other cell-damaging molecules (such as those present in cigarette smoke), results in the up-regulation of cell surface markers that permit monocyte attachment.

Monocytes attached to the endothelial cells infiltrate into the VSMC layer of the intima. Here, the monocytes differentiate into macrophages that are capable of accumulating lipid. Over time, the lipid-laden macrophages develop into foam cells that remain in the intima (figure 1.1b).

Lesions develop further due to a combined increase in proliferation and migration of VSMCs from the tunica media into the intima. This is likely driven, at least in part, by endothelin and angiotensin II stimulation of proliferation, as well as macrophage-derived cytokines. Concomitant with the increased VSMC number, synthesis of extracellular matrix components (e.g. collagen, elastin) is increased. The additional extracellular matrix constituents form a fibrous cap over the plaque.
During the development of the plaque, some of the foam cells and VSMCs undergo apoptosis due to the deleterious environment. This causes the release of lipids and necrotic substances, as well as crystallised cholesterol, leading to the formation of a lipid-rich necrotic core underneath the fibrous cap (figure 1.1c).

Stability of the plaque is maintained by the fibrous cap; damage to the fibrous cap can cause the release of pro-thrombotic factors from the plaque into the lumenal blood which can trigger severe clinical manifestations of disease. The fibrous cap can be weakened through shear stress of blood flow, or from within by macrophage-derived proteins. Such proteins include inducers of VSMC apoptosis (preventing further collagen deposition and permitting the weakening of the cap) as well as enzymes, such as collagenase, that directly degrade the fibrous cap constituents. Further, macrophages within the plaque produce pro-coagulant molecules that are responsible for promoting the thrombosis observed upon plaque rupture (figure 1.1d shows the thrombotic outcome of plaque rupture due to reduced integrity of the fibrous cap).

There are 2 principle clinical manifestations of CAD that can occur. Myocardial infarction (heart attack) is possibly the best-known complication of this disease. This event occurs due to fibrous plaque rupture and subsequent thrombosis, which can lead to ischaemia due to reduced blood flow. If the ischaemia is sufficiently extensive and prolonged, cardiomyocyte death is possible; this can be fatal. Symptoms of an MI include prolonged (>20 minutes) chest pain that is typically non-responsive to glycerine trinitrate spray, dyspnoea, fatigue, thready pulse, hypotension and bradycardia/tachycardia. Not all of these symptoms present in all patients (Camm & Bunce, 2009).

Angina pectoris (AP) is another pathology of CAD. There is a positive correlation between AP and risk of MI, and vice versa (Camm & Bunce, 2009). AP is characterised by symptoms similar to MI, but with reduced severity, and include chest pain that is often described as a “tightness” which may spread to the jaw or back, breathlessness and sweating. The symptoms of classical AP are induced by physical exertion or emotional distress, and manifest due to the reduced lumenal diameter caused by the presence of a plaque (AP is not induced by a thrombotic event). This means that the coronary arteries are unable to supply enough oxygen to the myocardium to meet the current demand, due to the volume of the plaque reducing the capacity of the vessel.
1.2 Coronary Artery Disease Risk Factors

CAD is a multifactorial disease; the interaction between various aspects of the environment, an individual’s behaviour, and genetic variation can alter the risk of developing CAD. There are two categories of risk factor for CAD; modifiable and non-modifiable. The former of these include those risk factors that can (at least to a degree) be reduced by changes in behaviour or with pharmacological therapies. These risk factors include hypercholesterolaemia, hypertriglyceridaemia, obesity and poor diet, sedentary behaviour, smoking, hypertension and diabetes mellitus. Non-modifiable risk factors include advancing age, male gender, post-menopausal status in females, and genetic factors, such as ethnicity and family history (reviewed by Yahagi et al. 2015).

1.3 The Genetics of Coronary Artery Disease

1.3.1 Assessment of Coronary Artery Disease Heritability

Family history is one of the most significant risk factors of CAD, suggesting genetic components can influence the risk of this disease. An early study to confirm this hypothesis and demonstrate that genetic factors can contribute risk to CAD was performed by Myers et al. (1990). In this work, it was observed that people with a parental history of CAD had a 29% increased risk of developing CAD.

Heritability (h^2), a measure of how much variation between individuals for a given trait is due to a genetic component, can be assessed using twin studies. This epidemiological approach utilises pairs of twins that are either monozygous or dizygous. The former twins share the same genome sequence, whereas dizygous twins share ~50% of their genome sequence. The assumption is that environmental factors are controlled for, as both twins were raised together. This therefore assumes nullification of an environmental component between the twins. Dizygous twins are preferred over standard siblings because the womb is shared by both twins during gestation, as is the environment during subsequent development. If a disease or trait has a genetic basis, the concordance rate will be higher for monozygotic twins than dizygotic twins.

Marenberg et al. (1994) utilised the twin study approach (with over 10,500 twin pairs) and compared the mortality from CAD in monozygous and dizygous twins. They observed a >2-fold increase in risk in male monozygous compared to dizygous twins. The difference in risk was even greater in females, with an almost 6-fold increased risk in monozygous twins. In this study, a heritability of 57% and 38% was calculated for males and females, respectively.
Additional studies have also replicated this finding, although the absolute value of heritability for both genders differs between studies. Wienke et al. (2001) observed $h^2$ values of 0.53 for males and 0.58 for females for CAD. Zdravkovic et al. (2002) calculated $h^2$ for males of between 0.45 and 0.69, whilst for females $h^2$ was 0.26-0.50. From these studies, heritability of CAD can be concluded to be between 40-60% (Björkegren et al. 2015).

Twin studies such as these are capable of providing estimates of the heritability of a trait, but they are unable to identify which loci or genes are involved in the heritable trait. To determine these, additional approaches must be used.

### 1.3.2 The Genetic Causes of Coronary Artery Disease: Familial Hypercholesterolaemia & Candidate Gene Studies

Familial hypercholesterolaemia (FH) is a condition characterised by elevated cholesterol levels that are inherited through families. Typical LDL cholesterol (LDL-C) levels observed in patients with this disease are $>4.9$mmol/L (recommended LDL-C is below 3mmol/L). In addition, patients usually have xanthomas (deposits of cholesterol under the skin) and early-onset CAD (Klose et al. 2014).

Khachadurian (1964) determined that this disease is inherited in an autosomal co-dominant manner; in this study, members of affected families were stratified according to plasma cholesterol level. Patients with cholesterol levels 2x the normal level were assigned as “heterozygous”, whilst “homozygous” patients exhibited cholesterol levels 4x higher than the normal plasma concentration. The frequency of heterozygotes is approximately 1/500; homozygosity is therefore uncommon, with frequencies of 1 in 1,000,000 (Soutar & Naoumova, 2007). Heterozygous patients present with coronary artery disease manifestations in early adulthood, whilst homozygotes experience these in childhood (Klose et al. 2014).

FH is caused by mutations in three principle genes; LDLR (low density lipoprotein receptor), APOB (apolipoprotein B) and PCSK9 (proprotein convertase subtilisin/kexin type 9). LDLR was identified as a candidate gene for FH given the observation of elevated LDL in patients. Subsequent analyses in patients and their families has identified over 1700 mutations in LDLR that cause FH, as listed in the database hosted by Leigh et al. (2008). Mutations in this gene prevent synthesis of functional LDLR protein, resulting in severely reduced LDL uptake and subsequent metabolism. This leads to elevated circulating LDL and the clinical symptoms of the disease. LDLR
mutations are, individually, rare, but due to the number of them account for a large number of FH cases.

APOB mutations have also been implicated in FH in some families by Innerarity et al. (1987). In these families, LDL particle binding to LDLR was affected. This group showed that the effect was caused by defective APOB. Soria et al. (1989) subsequently showed that a point mutation, causing an amino acid substitution (Arg3500Gln), was responsible for FH in these patients.

Linkage analysis of multiple families identified a locus on chromosome 1 (Varret et al. 1999) as associated with FH. This region contains 41 genes, one of which is PCSK9. Sequencing of this gene in multiple affected and unaffected family members identified a substitution mutation (S127R) that causes FH (Abifadel et al. 2003).

The studies of FH show that rare, large-effect variants in genes involved in cholesterol metabolism cause this disease. More common variants with a lower effect size could be contributing to CAD in genes known to be involved in CAD-related pathways. Indeed, candidate gene studies have identified CAD-associated variants. In this approach, cases and controls are compared for variants within candidate genes; genes are assigned as such based upon their known function.

An example of such work was presented by Topol et al. (2001), who genotyped SNPs in candidate genes in both cases and control families. SNP genotypes were compared between the two groups in order to identify any that were associated with CAD. This approach identified CAD-associated variants, including a missense variant in the thrombospondin-4 gene, as well as variants in thrombospondin-2 and thrombospondin-1.

Analysis of 71 candidate genes by genotyping 112 variants in a Japanese case/control cohort by Yamada et al. (2002) identified variants in four genes that were associated with MI. Interestingly, these associations appeared gender-specific. In men, variants in connexin-37 (C1019T) and in p22phox (C242T) were associated with disease, whereas in females, variants in plasminogen-activator inhibitor type 1 and stromelysin-1 were associated with MI. This indicates that some of the genetic risk factors that contribute to CAD may be gender-specific.

The candidate gene approach is a useful method by which confirmation of apparent involvement of variants in disease can be ascertained. However, there are several limitations to such a study methodology. The first, and arguably the most important, is the requirement of candidate genes, of known function, that could be affecting a
pathway involved in the disease. This is limiting because novel gene associations cannot be identified, and the requirement of knowledge of variants within those genes is also needed. Further, the polymorphisms identified as associated are possibly not the causative variants, leaving the actual mechanism by which a candidate gene is affected unfound.

The candidate gene approach for the investigation of complex disease suffers from high rates of false positive findings. This has been demonstrated by the inability to replicate the majority of associations in independent cohorts. Samani et al. (2007) attempted to replicate candidate gene associations as part of a genome-wide association study (see section 1.3.4). Of the 55 candidate genes examined, only 1 was replicated in this work, suggesting a high rate of false positive associations from CAD candidate gene studies. Subsequently, analysis of candidate gene replication studies from a number of diseases by Ioannidis et al. (2011) suggests that there are at least 20 false positive associations from candidate gene studies for every 1 true positive result. This value is variable and depends upon the disease and study considered, but does demonstrate that candidate gene studies appear to be prone to false positive associations, limiting the applicability of their findings.

1.3.3 Identifying Novel Loci Associated with Coronary Artery Disease: Linkage Studies

Linkage studies, involving families of affected and unaffected individuals, are another method by which genetic loci can be associated with disease. This method does not require a priori hypothesis, as the entire genome can be interrogated to identify loci associated with disease (at albeit low density). Genotyping of microsatellite markers is used to divide the genome up into segments; typically, these are ~770kb in length.

One example of linkage analysis being used to investigate CAD was presented by Samani et al. (2005). This study, utilising over 4000 people, identified markers on chromosome 2 that showed suggestive linkage to CAD, with a LOD score of 1.98. However, the markers failed to reach genome-wide significance.

A study by Wang et al. (2004) identified the chromosome 1p34-36 region as associated with premature MI. Interestingly, as noted by the authors of this publication, the connexin-37 gene is located within this region; this gene was identified by Yamada et al. (2002; section 1.3.2) as part of a candidate gene study.
A case of autosomal dominant CAD/MI has also been reported (Wang et al. 2003), with the genetic cause elucidated. A family consisting of 13 patients and 8 unaffected individuals were investigated by linkage analysis. A single locus (15q26) was identified, with a LOD score of 4.19. Of the 93 genes at this locus, Wang et al. (2003) studied MEF2A, a transcription factor involved in vascular development. Sequencing of the gene revealed a 21bp deletion, which resulted in the loss of 7 conserved amino acids from the nuclear localisation domain of the protein.

The microsatellite linkage method used to investigate CAD and MI has provided valuable clues regarding the genetic basis of this disease, but, as with the previously discussed approaches, linkage analysis has limitations. The principle issue with this technique is that, whilst genome-wide, the density of markers is not sufficient for refined investigation of a locus but rather a manner by which the genome can be broken up into ~770kb segments (on average). Each of these individual segments may contain many genes. Thus, the causative gene cannot be identified easily. Further, the nature of the causal polymorphism(s) cannot be determined. Additional functional interrogation of the loci identified by linkage is therefore necessary.

1.3.4 Identifying Novel Loci Associated with Coronary Artery Disease: Genome-Wide Association Studies (GWAS)

1.3.4.1 Introduction to GWAS

The development of the genome-wide association study (GWAS) has been an important facilitator in increasing our understanding of diseases with a genetic component. In this hypothesis-free methodology, many thousands of SNPs from across the genome are genotyped in large cohorts of cases and controls for the disease or trait of interest. Comparisons of allele frequency for all typed SNPs are made between the cases and controls, with any SNPs that reach genome-wide significance providing evidence for an association between the locus in which the SNP resides and the trait of interest.

The first GWAS was published in 2005 and investigated age-related macular degeneration (AMD) using a small number of cases and controls (96 and 50, respectively) (Klein et al. 2005). This study found a variant in the CFH (complement factor H) gene on chromosome 1. This variant was associated with a 7.4-fold increase in the risk of AMD in homozygous individuals.

Since this study, almost 2500 GWAS studies have been published (NIHR Catalog of Published Genome-Wide Association Studies, accessed August 2015; Welter et al.
2014; Hindorff et al. 2015). These studies have examined a diverse range of traits, including multiple diseases and related phenotypes, but also drug responses and traits such as hair colour and height.

1.3.4.2 Coronary Artery Disease GWAS

The first large-scale GWAS performed for CAD was carried out by the Wellcome Trust Case Control Consortium (WTCCC, 2007). In this publication, 2000 cases for multiple diseases and 3000 combined controls were genotyped for nearly 470,000 SNPs. A single locus was strongly associated with CAD \((p = 1.79 \times 10^{-14})\); the 9p21 region. Additionally, 6 further loci with moderate association were identified (1q43, 5q21, 6q25, 16q23, 19q12, 22q12) with \(p = 1.04 \times 10^{-4} - 5.72 \times 10^{-6}\).

Subsequently, multiple CAD GWAS have been performed by several groups. A total of 45 loci have been associated (at genome-wide significance) with CAD. A summary of these loci is given in table 1.1.

Recently, the CARDIoGRAMplusC4D Consortium (Deloukas et al. 2013) calculated the contribution of these 45 loci to additive genetic CAD variance. Assuming heritability to be 40%, these 45 loci contribute to 6% of the observed variation. Taking into account an additional 104 SNPs identified by false discovery rate analysis (at <5%), the total known genetic variation accounts for 10.6% of the genetic variation in CAD. This work was based upon a meta-analysis of over 22,000 cases and almost 65,000 controls.

CAD GWAS have found associations for loci that contain genes that are currently targeted therapeutically (e.g. lipid metabolism pathways). Novel loci have also been identified that do not contain known candidate genes, suggesting that they act through new pathways. These loci may present new strategies for therapeutic intervention.
<table>
<thead>
<tr>
<th>Band</th>
<th>Nearest Gene(s)</th>
<th>Lead/Proxy SNP(s)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p13.3</td>
<td>SORT1</td>
<td>rs599839</td>
<td>Samani et al. (2007)</td>
</tr>
<tr>
<td>1p32.2</td>
<td>PPAP2B</td>
<td>rs17114036</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>1p32.3</td>
<td>PCSK9</td>
<td>rs11206510</td>
<td>Myocardial Infarction Genetics Consortium (2009)</td>
</tr>
<tr>
<td>1q21.3</td>
<td>IL6R</td>
<td>rs4845625</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>1q41</td>
<td>MIA3</td>
<td>rs17464857</td>
<td>Samani et al. (2007)</td>
</tr>
<tr>
<td>2p21</td>
<td>ABCG5-ABCG8</td>
<td>rs6544713</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>2p11.2</td>
<td>VAMP5-VAMP8-GGCX</td>
<td>rs1561198</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>2p24.1</td>
<td>APOB</td>
<td>rs515135</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>2q22.3</td>
<td>ZEB2-AC074093.1</td>
<td>rs2252641</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>2q33.2</td>
<td>WDR12</td>
<td>rs6725887</td>
<td>Myocardial Infarction Genetics Consortium (2009)</td>
</tr>
<tr>
<td>3q23.3</td>
<td>MRA5</td>
<td>rs9818870</td>
<td>Erdmann et al. (2011)</td>
</tr>
<tr>
<td>4q31.22</td>
<td>EDNRA</td>
<td>rs1878406</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>4q32.1</td>
<td>GUCY1A3</td>
<td>rs7692387</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>5q31.1</td>
<td>SLC22A4-SLC22A5</td>
<td>rs273909</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>6p21.2</td>
<td>KCNK5</td>
<td>rs10947789</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>6p21.31</td>
<td>ANKS1A</td>
<td>rs17609940</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>6p24.1</td>
<td>PHACTRI</td>
<td>rs12526453</td>
<td>Myocardial Infarction Genetics Consortium (2009)</td>
</tr>
<tr>
<td>6q23.2</td>
<td>TCF21</td>
<td>rs12190287</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>6q25.3</td>
<td>SLC22A3-LPAL2-LPA</td>
<td>rs3798220</td>
<td>Myocardial Infarction Genetics Consortium (2009)</td>
</tr>
<tr>
<td>6q26</td>
<td>PLG</td>
<td>rs4252120</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>7p21.1</td>
<td>HDAC9</td>
<td>rs2023938</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>7q22.3</td>
<td>COG5</td>
<td>rs12539895</td>
<td>Peden et al. (2011)</td>
</tr>
<tr>
<td>7q23.2</td>
<td>ZC3HC1</td>
<td>rs11556924</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>8p21.3</td>
<td>LPL</td>
<td>rs264</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>8q24.13</td>
<td>TRIB1</td>
<td>rs2954029</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>9p21.3</td>
<td>CDKN2BAS1</td>
<td>rs1333049</td>
<td>Wellcome Trust Case Control Consortium (2007)</td>
</tr>
<tr>
<td>9q34.2</td>
<td>ABO</td>
<td>rs579459</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>10p11.23</td>
<td>KIAA1462</td>
<td>rs2505083</td>
<td>Erdmann et al. (2011)</td>
</tr>
<tr>
<td>10q11.21</td>
<td>CXCL12</td>
<td>rs501120</td>
<td>Samani et al. (2007)</td>
</tr>
<tr>
<td>10q23.31</td>
<td>LIPA</td>
<td>rs1412444</td>
<td>Wild et al. (2011)</td>
</tr>
<tr>
<td>10q24.32</td>
<td>CYP17A1-CNMM2-NT5C2</td>
<td>rs12413409</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>11q22.3</td>
<td>PDGFD</td>
<td>rs974819</td>
<td>Peden et al. (2011)</td>
</tr>
<tr>
<td>11q23.3</td>
<td>ZNF259-APOA5-APOA1</td>
<td>rs9326246</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>12q24.12</td>
<td>SH2B3</td>
<td>rs3184504</td>
<td>Gudbjartsson et al. (2009)</td>
</tr>
<tr>
<td>13q12.3</td>
<td>FLT1</td>
<td>rs9319428</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>13q34</td>
<td>COL4A1-COL4A2</td>
<td>rs4773144</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>14q32.2</td>
<td>HHIPL1</td>
<td>rs2895811</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>15q25.1</td>
<td>ADAMTS7</td>
<td>rs7173743</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>15q26.1</td>
<td>FURIN-FES</td>
<td>rs17514846</td>
<td>The CARDIoGRAMplusC4D Consortium (2013)</td>
</tr>
<tr>
<td>17p11.2</td>
<td>RAII-PEMT-RASD1</td>
<td>rs1293657</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>17q13.3</td>
<td>SMG6</td>
<td>rs216172</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>17q21.32</td>
<td>UBE2Z</td>
<td>rs46522</td>
<td>Schunkert et al. (2011)</td>
</tr>
<tr>
<td>19p13.2</td>
<td>LDLR</td>
<td>rs1122608</td>
<td>Myocardial Infarction Genetics Consortium (2009)</td>
</tr>
<tr>
<td>19q13.32</td>
<td>ApoE-ApoC1</td>
<td>rs2075650</td>
<td>Butterworth et al. (2011)</td>
</tr>
<tr>
<td>21q22.11</td>
<td>KCNE2 (Gene desert)</td>
<td>rs9982601</td>
<td>Myocardial Infarction Genetics Consortium (2009)</td>
</tr>
</tbody>
</table>

Table 1.1: The Current, Published Loci Associated with CAD by GWAS. Adapted from the findings of the CARDIoGRAMplusC4D Consortium (Deloukas et al. 2013).
1.3.4.3 Understanding GWAS Associations and the Challenges of Functional Analyses

The GWAS methodology benefits from being hypothesis-free, as no candidate genes or loci are needed. Further, the genome is typed at a high density using SNPs. The linkage disequilibrium between the most-significant SNP and the other typed/untyped SNPs can be used to identify the region in which the causal variant(s) are likely to be.

This approach, whilst providing much insight into the genetics of CAD, does have limitations. A GWAS is only capable of reporting an association between genotyped SNPs and the trait of interest. Because of this, a GWAS can only identify loci associated with disease, not specific, causal polymorphisms. This is because of linkage disequilibrium; SNPs that are inherited together through linkage create haplotypes. It is therefore difficult to classify SNPs as causal or not within a disease-associated haplotype because they are all inherited with approximately the same frequency as the lead SNP from the GWAS.

In some cases, identifying a genomic region by GWAS is sufficient to elucidate the affected gene because there is only a single gene in the locus. However, in the majority of loci, this is not the case, with multiple genes present. One may hypothesise which genes are likely candidates, but this means that causal genes may be overlooked in some cases. It is also possible that there is no likely candidate gene in a locus, or there are no genes at all. In this case, the causal effects are likely to be regulatory in nature.

Identifying the mechanisms and genes involved in such cases can be complicated by the presence of multiple annotated regulatory sites because the linkage disequilibrium at the locus means that the affected regulatory element is unknown.

Because of these confounding issues, a GWAS can only be used as a method by which to identify loci associated with disease. In most cases, further association and functional studies are required to identify the causal variant(s), and the genes/pathways affected.

Additional association methods that provide valuable information can be used in an attempt to identify the gene(s) regulated by SNPs in the locus of interest. This can be investigated by eQTL (expression quantitative trait loci) analysis, which associates gene expression with tag SNP genotypes in a genome-wide manner. Such work may reveal differential gene expression that is associated with a particular genotype of a tag SNP. eQTL data therefore identifies genes whose regulation may be affected by variants within a particular haplotype, but because tag SNPs are used for the analysis, causal
SNPs cannot be determined from an eQTL association. Rather, eQTL analyses facilitate understanding of the potential mechanisms through which causal variant(s) may act at a locus of interest. It is important to consider which tissues are used for eQTL analyses in order to obtain data relevant to the disease or phenotype of interest.

Determination of causality with respect to variants, such as SNPs or rare variants with high effects, and a particular trait is complicated by the linkage disequilibrium between the causal variant and other “bystander” variants in the haplotype. This issue is further confounded by variants elsewhere in the genome that are not inherited by linkage (e.g. on a different chromosome). Such background variation contributes to the heterogeneity between individuals. Thus, a method of investigating specific variants, in isolation from all other background variation (both inside and outside of the linkage disequilibrium block) is required.

1.4 Functional Analysis of Disease-Associated Variants Using Genome Editing

Genome editing (also referred to as gene engineering, gene editing or genome engineering) is a group of powerful technologies that can be used to modify the genome in a pre-designed manner in order to determine the effects of the alterations that have been made. For all of the genome editing approaches available, it is important to consider the cell models utilised. This is needed to ensure meaningful results are obtained from any functional studies performed after editing. Therefore, the cell models used should, whenever possible, be applicable to the disease or pathway of interest and closely recapitulate the in vivo functions of interest.

All genome editing technologies utilise cellular DNA repair pathways in order to modify the genome. The two primary repair pathways utilised are non-homologous end joining (NHEJ) and homologous recombination (HR), although homology-directed repair (HDR), similar to HR, is also used. Figure 1.2 illustrates the key processes involved in NHEJ and HR.

NHEJ is the pathway used to repair a double strand break (DSB) in cells that are not in S phase (no nearby homology donor is available). During NHEJ (reviewed by Lieber (2010) and Davis & Chen (2013)), ligation of complementary strands is performed by a complex of scaffold proteins, nucleases, polymerases and ligases. The proteins involved depend upon the nature of the DSB. In some cases, where the DSB does not produce complementary ends for direct ligation, nucleases resect out nucleotides as required on both strands in order to give a ligatable template. This results in an indel mutation
(because of this feature, NHEJ is said to be “error-prone”). The number of nucleotides added or lost varies. In genome editing of a gene, for example, such indels can produce a frameshift mutation that introduces an early stop codon, preventing translation of the resulting mRNA from the gene through nonsense-mediated decay. This effectively prevents expression of the gene from the edited chromosome. Targeting of both chromosomes would generate a complete knock-out, where no protein product from the gene is made.

HR (reviewed by Sung & Klein, 2006) is an alternative pathway to NHEJ, and is active when there is a nearby, homologous template molecule. Thus, this pathway is most active during S phase, when the sister chromatid is nearby. Similar to NHEJ, HR requires resection of DNA from the chromosome that has undergone a DSB. This is followed by strand invasion from the homologous donor molecule (only one strand) and subsequent DNA synthesis on the DSB chromosome using the donor as a template. Following this, two potential mechanisms occur, “double-strand break repair, DSBR” or “synthesis-dependent strand annealing, SDSA”. In DSBR the second DSB end is captured and DNA synthesis and ligation is used to repair the gap. This can be terminated by resolving the Holliday junctions produced by the strand invasions in a manner that results in crossing-over. Alternatively, in SDSA, the homologous donor is removed and DNA synthesis occurs using the repaired strand, followed by ligation. This method of HR does not produce any cross-over events. Because HR uses a template molecule for repair, and no nucleotides are lost, this repair mechanism is deemed “error-free”. However, whilst no nucleotides are lost, alterations may be made if the template molecule contains sequence differences that do not significantly disrupt the overall homology between the vector and the target locus. It is this feature of HR that makes it useful for genome editing.

HR is used in genome editing to make small changes (e.g. small indels/substitutions of SNPs) as well as larger alterations, such as insertion of markers/reporters (which may allow selection of targeted cells), promoters and large-scale deletions.

Homology-directed repair is also utilised during genome editing procedures. This repair pathway, which is poorly understood, utilises a small template (~100bp) that contains homology with the target locus (except for, for example, a single nucleotide difference). HR repair requires much larger template molecules (at least several hundred bp). In genome editing, a single-stranded oligonucleotide molecule is used as the repair template for HDR (Bedell et al. 2012). Off-target integrations and unintended on-target
integrations have been observed when using this method of repair (Radecke et al. 2010). HDR is used to make small changes to the target locus, such as indels, SNP genotype alterations or epitope tag insertions (Campbell et al. 2013).

Genome editing techniques can be grouped into 3 categories, based upon how the technology works. Whilst these three groups may be referred to in different ways, the following shall be used in this thesis: “Protein-Directed, Protein Mediated”, “RNA-Directed, Protein-Mediated” and “DNA-Directed, Cell Mediated”.

In the following sections, a discussion regarding the predominant genome editing methods is given, with a focus on how the approaches work and their advantages/disadvantages. Examples of uses of each technology shall also be given. An overview of rAAV genome editing is given in section 1.4.3, whilst an in-depth discussion of rAAV technology, including technical aspects of this method, is given in section 3.1. Particular attention to this technique has been made as this is the method chosen for use in the work presented in this thesis. Justification for this choice is provided in section 1.5.
Figure 1.2: Homologous Recombination (HR) and Non-Homologous End Joining (NHEJ) Repair Pathways. In HR, strand invasion of a sequence with significant homology to the region being repaired initiates Holliday junction formation and subsequent DNA synthesis. HR is terminated by Holliday junction resolution. NHEJ requires resection of non-compatible ends followed by ligation. Key proteins involved in both processes are given. Adapted from Kee & D’Andrea (2010).

1.4.1 “Protein Directed, Protein Mediated” Genome Editing Technologies
Two principle methods are assigned to this category. These techniques involve proteins that recognise the target DNA through pre-designed binding motifs within their structure. Such binding initiates DNA cleavage by the proteins. This is repaired using
either non-homologous end joining (NHEJ) or, if the cell is provided with one, a template molecule and the HR or HDR processes.

1.4.1.1 Zinc Finger Nucleases

The first technology to be created for genome editing was the zinc finger nuclease (ZFN) by Kim et al. (1996). Zinc finger nucleases are a fusion between the FokI endonuclease from *Flavobacterium okeanokoites* and zinc finger DNA binding domains. Zinc fingers are modular in structure, with each motif (comprising Tyr/Phe-X-Cys-X₂₄-Cys-X₃-Phe-X₅-Leu-X₂-His-X₃₅-His) binding to a specific triplet of bases. By modifying the variable amino acids (X in the motif sequence above), theoretically any DNA target can be recognised.

Two zinc finger nucleases are required per locus of interest. These are directed to target either strand of the region of interest. The two proteins bind the DNA and are separated by ~5-6bp of intervening DNA, across which the FokI domains lie. Figure 1.3A shows the structure formed by two zinc finger nucleases.

ZFN proteins are usually provided to the cell by transfection/electroporation of plasmids encoding the two ZFN proteins. The ZFN plasmids are created by cloning sequence-specific domains into the vector. Binding of two ZFNs to the target locus induces a double-strand break that is repaired by NHEJ. If a homologous donor DNA is provided (as a plasmid or an oligo), HR can occur and be used to make directed alterations to the target locus (instead of relying upon NHEJ to make random alterations). However, whilst HR can occur, it does not always do so. Thus, in some cases, despite a template being provided, NHEJ may still occur.

The zinc finger nuclease system for genome editing provides some benefits that suit the technology for some experimental designs. These include high efficiency for NHEJ-based edits, and, in NHEJ experiments, no selection step is required prior to screening because of the high efficiency. This means that cell lines can be obtained relatively quickly. In addition, ZFNs can be used to edit the embryos of zebrafish (Liu et al. 2014) and rodents, such as mice (Park et al. 2014) and rabbits (Flisikowska et al. 2011). The approach has even been used successfully with algae (Sizova et al. 2013) and plants (Tovkach et al. 2010), demonstrating the flexibility of this system.

However, this technology has multiple drawbacks. These include the potential need to redesign one or both ZFNs due to failed experiments, increasing both the cost of the experiment and the time required to generate edited cell lines. Further, in order to
achieve expression of the proteins in the cell line/type of interest, transfection or electroporation is required. This means that the cells chosen have to be amenable to such treatment, which can be stressful and reduce viability.

Additionally, whilst for some experimental approaches the absence of a selection step can be beneficial (i.e. when a high efficiency is expected in experiments where, for example, NHEJ is the desired result), in others (especially those relying upon the HR pathway), this is undesirable and selection may be needed. In these cases, selectable markers need to be included in the HR template.

A final issue with ZFNs is the potential for off-target cleavage and subsequent repair that may affect genomic sequences elsewhere. The regions likely to be affected by off-target effects can be predicted bioinformatically, but this is not ideal for several reasons. Firstly, confirmation of the absence of off-target effects is required by sequencing such predicted regions, increasing the cost of the procedure. Secondly, off-target ZFN activity could occur at loci not predicted bioinformatically; reliance upon such software could mean some off-target effects are not identified.

On a related note, when an approach requires HR of a donor template into the locus cleaved by the ZFNs, this occurs at a low frequency. There is therefore a high probability that if HR does occur, it only occurs to a single chromosome. In this case, the other chromosome may be cleaved by the ZFNs and repaired with NHEJ, changing the sequence at the locus of interest in a way that was not envisaged through indel mutation. For the majority of experiments, this would likely be a confounding factor affecting experimental data.

For experiments with NHEJ as the desired repair pathway, efficiency of targeting both chromosomes is high. However, the actual alteration made to them is different, and whilst this may be sufficient to knock-out (KO) the gene of interest, there may be additional effects induced by these differences (for example, creation or abolition of regulatory elements that lie within the gene sequence). The effects of this can be minimised by *in silico* analysis of the region and good design of the ZFNs so as not to affect known regulatory sites.

ZFNs have been used to repair the *CFTR* gene in induced pluripotent stem cells (iPSCs) from patients with cystic fibrosis (CF). In this work, Crane *et al.* (2015) isolated skin fibroblasts from CF patients, generated iPSCs from these cells and targeted *CFTR* to repair the mutations in the gene before differentiating the cells into epithelial cells;
repaired cells showed increased mature CFTR protein (a chloride channel) and improved protein function.

**1.4.1.2 Transcription Activator-Like Effector Nucleases (TALENs)**

TALENs are similar in structure and function to ZFNs, in that they employ a modular DNA binding domain and a FokI cleavage domain. Further, they too require binding of two TALEN molecules to the target DNA (one on either strand) in order to elicit cleavage.

TALENs are derived from transcription activator-like effector (TALE) proteins produced by *Xanthomonas* spp., a plant pathogen (Bonas et al. 1989). TALEs recognise specific plant gene sequences and increase expression of those genes (Kay et al. 2007). Recognition of the target locus is achieved through a modular DNA binding domain, each consisting of ~34 amino acid repeats (Herbers et al. 1992), two of which are variable (residues 12 and 13) and govern the nucleotide that particular motif binds to (Moscou & Bogdanove, 2009); a single pair of amino acids is specific to a particular base in the target DNA.

A complete TALEN (first created by Christian et al. 2010) typically contains an N-terminal nuclear localisation signal, the repeat variable di-residue (RVD) motifs and a C-terminal FokI domain (Campbell et al. 2013). A schematic diagram of TALEN-DNA interactions is given in figure 1.3B.

TALENs show similar advantages and disadvantages as ZFNs (see section 1.4.1.1). However, TALEN plasmids are easier to assemble than ZFN vectors due to the modularity of the repeat motif and the advent of cloning methods involving sequential ligation of multiple motifs at once, meaning that TALEN vectors can be assembled using approximately 4 ligation reactions, instead of one per nucleotide in the recognition sequence (Reyon et al. 2012).
Figure 1.3: Structure and DNA Binding of Zinc Finger Nucleases and TALENs. (A) Zinc finger proteins (ZFP) comprise multiple zinc finger domains, each of which recognises 3bp of DNA. Zinc finger nucleases (ZFN) also contain a FokI DNA cleavage domain. The binding of two ZFNs, on opposite DNA strands, permits double-strand cleavage of the DNA by the two overlapping FokI domains. (B) TALENs are composed of multiple TALE domains, each of which binds to a single base. Each TALE domain is identical, except at two adjacent residues. The amino acids at these two sites determine which nucleotide the TALE recognises. Like ZFNs, TALENs bind on opposite DNA strands and the FokI domain cleaves the DNA, causing a double-strand break. Adapted from Gaj et al. (2013).
Many groups have utilised TALENs for genome editing. Uhde-Stone et al. (2014) utilised TALENs and HR to knock-out a miRNA (miR-21) in HEK293 cells. This work involved biallelic targeting of the miR-21 gene with a pair of TALENs and a HR template encoding RFP and puromycin resistance. Clonal cells expressing RFP and surviving puromycin selection were screened to confirm insertion of the HR template into the genome at the target locus. In clones where this had occurred, a subset showed biallelic targeting of miR-21 by both HR and NHEJ repair. These cells were subsequently used to show that the loss of miR-21 expression permits increased expression of its target gene (PDCD4), resulting in reduced proliferation. Further, data from these lines suggests that loss of miR-21 is capable of affecting the expression of 17 other miRNAs (which did not share homology with miR-21 and therefore are highly unlikely to have been targeted with the TALEN pair).

1.4.2 “RNA Directed, Protein Mediated” Genome Editing Technology: CRISPR/Cas9

This category of genome editing methods currently contains a single technology – the CRISPR/Cas9 system. The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats, named by Jansen et al. 2002) were first discovered by Nakata et al. (1989) in Escherichia coli. Similar loci in additional species have been found, and have been compared by Mojica et al. (2000). In each species, repeats of 21-37bp are interspersed with similarly-sized, non-repetitive sequences (spacers). In addition, a 300-500bp region (termed the “leader region” by Jansen et al. 2002) is present on one side of the cluster. The repeats and leader region are different between species, but are conserved within species. Multiple CRISPR loci can be present throughout the genome, with the number varying between species. Jansen et al. (2002) identified 20 CRISPR loci in Mycobacterium jannaschii. This suggests that the CRISPR locus is a mobile element.

The number of repetitive regions also varies between species (124 were found in Mycobacterium thermoautotrophicum). Jansen et al. (2002) also identified genes that were associated with the CRISPR loci, which they termed “Cas” (CRISPR-Associated Sequence) genes.

Pourcel et al. (2005) discovered that the non-repetitive spacers are derived from viral sources. This led to the hypothesis of a form of bacterial immunity, protecting against invading phage viruses – this was confirmed by Barrangou et al. (2007). Transcription
of the CRISPR locus into RNA (crRNA) provides immunity to phages when the spacer is homologous to the infecting virus sequence (termed the protospacer). The protospacer may be DNA (Marraffini & Sontheimer, 2008) or RNA (Hale et al. 2009). Additionally, a protospacer adjacent motif (PAM) is required for the recognition of the protospacer. The PAM is located next to the protospacer in the genome of the invading virus, but is not located in the CRISPR locus (Mojica et al. 2009) and stimulates the cleavage of the target DNA by the Cas protein (Garneau et al. 2010).

Further understanding regarding the CRISPR/Cas mechanism was revealed by Deltcheva et al. (2011), who identified the trans-activating CRISPR RNA (tracrRNA). Complementary binding of the tracrRNA to the pre-crRNA and subsequent cleavage by RNaseIII (in Streptococcus pyogenes) causes maturation of the crRNA. Mature crRNA recognition of, and base-pairing to, phage DNA recruits a Cas protein to the heteroduplex, which is cleaved if the Cas-specific PAM is present.

Current CRISPR/Cas9 genome editing systems are based on the work by Bassett et al. (2013), in which the crRNA and tracrRNA molecules have been replaced with a single RNA called a guide RNA (gRNA). The gRNA is composed of a region homologous to the target region (~20bp) and a scaffold region that provides structure. The Cas protein Cas9 from S. pyogenes is frequently used (Cas9 PAM is NGG).

To use the system, the guide sequence (the sequence with homology to the target) is cloned into the remainder of the gRNA in an expression vector. This is then transfected or electroporated into cells alongside a Cas9 expression construct. Both components are expressed, with the gRNA inducing cleavage of the genomic DNA by directing Cas9 activity (figure 1.4). When designing the guide sequence, it is essential that it is designed adjacent to the PAM specific for the Cas protein used.

Following cleavage, NHEJ is used to repair the DSB in the absence of a repair template. HR or HDR can be used if a repair template is provided.

The advantages of this system over ZFN and TALEN methods are increased simplicity with respect to cloning – a single ~20bp cloning event is all that is required to create a new vector for use. CRISPR/Cas9 (also referred to simply as CRISPRs) show high NHEJ efficiency, and, like ZFNs and TALENs, can be used with HR and HDR (at lower efficiency compared to NHEJ; comparable to that observed with ZFN and TALENs). Work by Ramakrishna et al. (2014) shows that enrichment of CRISPR-edited cells can be performed using surrogate reporter vectors. Transfection of these vectors into cells alongside the CRISPR plasmid(s) enables CRISPR-mediated cleavage
of the vector. Indels produced by the resulting repair of the plasmid rescue an out-of-frame reporter gene, such as hygromycin resistance or red/green fluorescent proteins (RFP/GFP), allowing expression of the reporter and subsequent selection of CRISPR-expressing cells (which are likely to also be targeted).

The disadvantages of this approach are as for those given for ZFNs and TALENs (sections 1.4.1.1 and 1.4.1.2, respectively), with the potential for increased off-target effects without careful design/screening (due to the need for only ~20bp recognition sequences) (Koo et al. 2015).

The CRISPR/Cas9 system, since its first use as a genome editing method in 2013, has been widely used for a number of applications in a number of species. A recent example of CRISPR/Cas9 work was by Ousterout et al. (2015), who corrected mutations that cause Duchenne muscular dystrophy (DMD) in cell lines. This disease is caused by mutations that cause frameshifts, resulting in the absence of active dystrophin protein. DMD patient myoblast cells were taken and exons of the dystrophin gene were deleted by using multiple gRNAs. Deletion of non-essential exons was found to return the essential exons back into frame, restoring dystrophin protein expression.

![Figure 1.4: CRISPR/Cas9 Genome Editing System with a Single gRNA.](image)

The sgRNA (synthetic guide RNA) is composed of a scaffold sequence (red) and a recognition ("guide") sequence (blue) which is complementary to the target (blue bar). The Cas9 protein recognises this interaction (due to the tracrRNA part of the gRNA) and cleaves the target DNA (red arrows). The PAM (red bar) must be present adjacent to the target sequence for Cas9 cleavage. Adapted from Ran et al. (2013).
1.4.3 “DNA Directed, Cell Mediated” Genome Editing Technologies

Homologous recombination in mouse embryonic stem cells has facilitated the generation of murine models, harbouring many alterations to the genome. These include gene knock-outs and transgenic mice containing human genes. These mouse models are generated through homologous recombination of plasmid DNA into the target locus in murine embryonic stem (ES) cells, followed by transfer of the cells into a blastocyst and subsequent implantation into a female mouse (Hall et al. 2009).

The efficiency of this process is sufficient when generating mouse models, but is much lower when the technique is utilised in experiments involving human cells (efficiency may be low - <10^{-7}; Vasquez et al. 2001). An alternative method is to use viral-based vectors to deliver the HR template to the cells. One such virus, the recombinant adeno-associated virus (rAAV), has been shown to have efficiencies of between 0.4-13% (reviewed by Hendrie & Russell, 2005). The improvement in efficiency provided by rAAV technology permits the editing of human cells using HR.

1.4.3.1 Overview of rAAV and the rAAV Genome Editing Procedure

rAAV are modified forms of a naturally occurring virus, the AAV (Choi et al. 2005). AAV are a member of the parvoviridae group. With a ssDNA genome of 4.7kb in size, these viruses are unable to replicate in the absence of helper viruses (such as adenoviruses). Wild-type AAV contain two genes; rep and cap that encode the proteins required for AAV genome replication and capsid production, respectively. Rep encodes 4 proteins (Rep78/68/52/40) needed for replication, whilst cap produces 3 viral capsid proteins (VP1/2/3). Multiple proteins are synthesised from these two genes through the use of multiple promoters and start codons, as well as alternate splicing.

AAV genomes are packaged within an icosahedral capsid comprising a total of 60 capsid proteins (t = 1 symmetry), predominantly VP3 (Choi et al. 2005). The inverted terminal repeats (ITRs) at either end of the viral DNA are the only sequences required for encapsidation. The ITRs also direct integration of wild-type AAV into a specific locus on chromosome 19q13.3q-ter (Kotin et al. 1991) when no helper virus is available, due to sequence motifs within the region (Kotin et al. 1992).

Because only the ITR sequences are required for assembly of the virus, the DNA between the two ITRs can be replaced with human genomic sequences. These can then be integrated into the genome at the homologous locus via homologous recombination (Russell & Hirata, 1998).
There are several capsid serotypes, each with different uptake and trafficking behaviours, as well as affinities and specificities for different cell types (Keiser et al. 2011). For example, AAV1 and AAV6 show better transducability than AAV2 for muscle (Chao et al. 2000). Serotype 2 (rAAV2) is the most widely used and best characterised capsid (Choi et al. 2005).

rAAV vectors used for genome editing are composed of two homology arms identical to the target locus (except for any modifications that are to be made to the region). Between the two homology arms lies a selection cassette, which encodes a selectable marker (such as an antibiotic resistance gene), nested PCR primers for diagnostic PCR screening and two loxP sites. Homologous recombination of the vector at the target locus incorporates the alterations included in the homology arms, as well as the selection cassette, into the genome. Cells that have integrated the vector can be selected through antibiotic treatment, and targeted cells identified by diagnostic PCR screening. The selection cassette can be removed by Cre recombination, leaving a single loxP site (34bp) behind. The modifications made to the genome remain. Further details regarding the design of rAAV vectors and the technical procedure can be found in section 3.1.

1.4.3.2 Advantages and Limitations of rAAV Genome Editing
rAAV-mediated genome editing offers several advantages over other technologies. The first of these is the utilisation of homologous recombination at the target locus. This feature is beneficial in that a specific change can be made with a significantly reduced risk of unpredictable/undetectable off-target effects. With the other genome editing methods, there is a high probability of both on- and off-target DSB induction by NHEJ and low on-target HR efficiency. Off-target effects with alternative methods occur due to the short recognition sequences used; the likely site of such off-target effects can be predicted bioinformatically, but some may be missed. The long homology arms of rAAV vectors facilitate HR into the target locus and minimises non-specific HR. Off-target integration of rAAV vectors by, for example, NHEJ, can be detected through diagnostic PCR screening of antibiotic-selected cells without knowledge of the locus at which this has occurred.

rAAV vectors also provide a selectable marker that can be removed by Cre recombination. This helps in the identification of targeted cell lines by inhibiting
proliferation and inducing cell death of non-targeted cells, enriching cell populations with targeted cells and facilitating their successful identification by diagnostic PCR.

The rAAV system is highly flexible and can be used for a large number of purposes. For example, altering SNP genotypes, mutation introduction at specific nucleotides (in a pre-designed manner), gene knock-outs (KO), gene, tag, promoter or reporter insertions, and deletion of small or large regions of DNA are all possible with this technology.

However, as with all technologies, there are some limitations to the rAAV method. The first of these is the limited size of the vector, which limits the amount of DNA that can be inserted into target loci. This can be circumvented by the use of multiple vectors, but this increases the cost and time needed for such projects.

Also, whilst the efficiency of the method is comparable to the HR efficiencies observed for the other methods when using repair templates (0.1-1%; Hirata et al. 2002), for some projects the use of alternative methods may be preferential. For example, for gene KO studies, rAAV would have less efficiency than the ZFN, TALEN or CRISPR-based approaches because the NHEJ induced by these technologies is capable of causing a gene KO with higher efficiency than HR.

Another limitation of the approach is the need to use cell lines that are capable of long-term culture, are amenable to rAAV transduction and are not resistant to the chosen selectable marker. Therefore, extensive characterisation of potential cell lines is required prior to targeting. Of these parameters, the ability to culture the cells for a long time is the most limiting factor. Cells that have limited replication ability in culture, and thus reach replicative senescence quickly, are unsuitable for the approach. This is due to the long-term culture that is necessary for rAAV genome editing. This therefore limits the cell lines that one may use for experiments. In some cases, where basic cell biology is being explored, this is less of an issue. However, where complex, differentiated cell functions are being investigated, it may be difficult to use the correct cell model.

Whilst this may potentially be a limitation for some studies, it should be considered that any non-NHEJ approach taken with ZFNs/TALENs/CRISPRs would also require clonal expansion of cells (which may not be possible with the cell type of interest). NHEJ-based work, however, would likely be possible for the majority of cell lines/types in cases with high efficiency, as clonal line derivation may not be required. As cells capable of long-term culture are needed, cancer cells and immortalised cell lines can be used, and may be appropriate for some studies. However, stem cells such as iPSCs may be used as an alternative to cell lines. iPSCs are capable of long-term culture and can be
differentiated into a number of cell types, facilitating investigation of loci with suspected effects in multiple cell types. Conversely, iPSCs may also be advantageous for the examination of loci in cases where the cell type(s) affected are unknown. This is therefore beneficial for investigation of the CAD-associated 9p21 locus for example; it is unclear which cell type(s) 9p21 variants act through in order to contribute to CAD. Genome editing of iPSCs followed by differentiation into multiple lineages may help to reveal which cell types are affected by 9p21 variants.

However, whilst potentially suitable for a number of genome editing approaches, there are some inherent difficulties with the culture and differentiation of iPSCs that may need to be considered for some projects. iPSC culture requires expensive culture media, and differentiation of these cells can be challenging, with protocols only available for a subset of cell types. Further, often only a proportion of cells fully differentiate, meaning that isolation of differentiated cell pools is required.

The effect of long-term culture during the rAAV targeting procedure should also be considered. Due to the high number of cell divisions necessary, cells may begin to display an “aged” phenotype induced by the prolonged culture. Due to this, any difference in phenotype exhibited after editing may simply be due to the aged phenotype of the cells. As such, in certain experimental systems, such differences may result in misleading data. Because of this possibility, the development of aged controls should be performed alongside the edited cells by culturing the cells in such a way so as to mimic the ageing process experienced by the targeted cells as closely as possible.

Another prohibitive aspect of rAAV technology is the cost and time needed to generate targeted cell lines. The cheapest genome editing approach is likely CRISPR, whereby a single, short gRNA requires cloning into a vector, with subsequent transfection and screening. ZFN and TALEN methods require plasmid assembly; this can be expensive and time-consuming. However, multiple attempts may be needed (which can increase the time and cost of cell line production) with different gRNA/ZFN/TALEN designs.

With rAAV, plasmid synthesis, virus production and purification, titre determination and subsequent infection and screening are needed. This is a time-consuming and costly procedure, which is also iterative, as it requires optimisation of multiple aspects of the technique. These successive attempts at rAAV genome editing increase the cost and time needed to generate edited cell lines.
rAAV, like all technologies, provides benefits and limitations that must be addressed prior to use. It is therefore important that genome editing work utilises the most suitable technique available, balancing the benefits each approach provides against its limitations.

1.4.3.3 Example Uses of rAAV Technology

rAAV technology for alteration of the genome has been used by many different groups for a variety of purposes. Stable alterations to the genome are possible with this technology, which can be broadly classified based upon the specificity of the integration; “random” and “specific”. In the former case, the rAAV vector integrates at random into the genome by non-homologous methods due to the absence of homology arms; the vector only carries, typically, a promoter and a gene between the ITRs. Such a method is used when the location of vector insertion is not considered to be important, and that the expression of the transgene is all that is desired.

“Specific” use of the rAAV methodology relies upon HR into a target locus due to homology arms (as described above). These methods are often used to make more specific changes to the genome, where off-target effects are generally avoided.

In addition to the stable, integration-requiring experimental methodologies, rAAV is also used for transient transgene expression. Growth and division coupled with degradation of the vectors results in loss of transgene expression over time in a population of cells; hence, the effect of the transgene is only transient.

In all three cases of rAAV use, they have their merits with respect to certain experimental designs. Below, examples are given for the two stable integration approaches.

Random integration of rAAV vectors was utilised by Moore et al. (2005) in novel work that showed, for the first time, that rAAV delivery of RNAi (RNA interference) molecules can be performed for successful treatment of hepatitis B infection. Hepatitis B virus (HBV) is a pararetrovirus, and therefore has an RNA-based genome. As RNA is degraded by RNAi, Moore et al. (2005) investigated whether RNAi could be used to reduce HBV load. Indeed, seven days after transduction (efficiency 95%) with rAAV vectors (500-1000 virus particles/cell), the level of HBV produced by stably-infected HEK293 and HepG2 cells was reduced by 89% and 98%, respectively. Expression of the RNAi from the random integration site(s) in the genome continued to reduce HBV titre after 5 months, with reductions of 80-90% for both cell lines compared to non-
transduced controls. This work demonstrates that rAAV vectors could be used for therapeutic delivery of RNAi; however, the development of vectors with a specific integration site would be better suited for gene therapy uses.

In similar work by Cruz et al. (2007), integration of siRNA against the α-1 antitrypsin (AAT) gene in in vitro (HepG2) and in vivo (mouse) models was performed. Mutations in this gene can result in AAT deficiency and liver disease through hepatocellular accumulation of the protein and subsequent polymer formation. Random integration of rAAV vectors into the genome resulted in siRNA expression and a reduction in AAT expression by ~50% in HepG2 cells. In mouse, a 2-3-fold decrease in serum AAT was observed after 14 days, with reduced accumulation of AAT monomers but no change in polymeric AAT (due to the long half-life of this protein). Cruz et al. (2007) therefore showed that stable expression of siRNA can reduce synthesis of nascent AAT molecules, and postulate that long-term reduction in AAT production may enable clearance of liver disease-inducing AAT polymers.

A novel use for rAAV has been developed by Cunningham et al. (2015). rAAV vectors were used to insert DNA into random locations in the genome. The insert DNA was flanked by PiggyBac ITRs (a transposon system identified in a species of moth) in the vector. Co-infection of cells with the vector and a PiggyBac transposase-encoding rAAV induced transposition of the transgene out of the rAAV and into the genome. A 20-fold increase in efficiency compared to standard rAAV methods was observed. This was used to correct urea metabolism defects in mouse liver when injected in utero prior to birth. Further discussion of the PiggyBac system and its potential benefits to other rAAV editing applications is discussed in section 3.3.

Early work with specific integration of rAAV vectors into a target locus was performed by Hirata et al. (2002), who used selection cassettes for the first time that encode G418 resistance. Targeted integration of the vectors into the HPRT and hTERT loci with acceptable efficiency (1%) was observed. Targeted cells could be chemically-selected and identified by PCR and sequencing. This work was the first to show on-target integration with a selection cassette in the rAAV vector.

rAAV technology can be used to develop in vitro models of disease. Shimada et al. (2009) introduced a known behaviour-altering mutation into the CK1ε gene (this mutation is known as tau mutation, which affects the circadian rhythm of mice and hamsters) of Rat1-R12 cells in order to study circadian rhythms. Selection, PCR screening and sequencing identified a single clone that harboured the tau mutation.
Successful integration of the vector yielded cells that displayed shortened circadian rhythms as assessed by luciferase activity (measured over 120 hours).

In addition to making small base changes at target loci, larger insertions can be made to, for example, monitor gene expression. Fernandez et al. (2007) used rAAV to insert an enhanced GFP (EGFP)-luciferase fusion protein gene adjacent to, and downstream of, the c-Myc oncogene in primary human foreskin fibroblasts, placing the EGFP-luciferase fusion under the control of the c-Myc promoter, thereby allowing both selection of targeted cells and the measurement of c-Myc promoter activity. Fernandez et al. (2007) demonstrated that such reporters would be useful for monitoring gene regulation – treatment of the cells with serum (known to induce c-Myc) increased EGFP-luciferase expression.

Large insertions can also be used to integrate entire gene sequences at specific loci. Barzel et al. (2015) treated haemophilia B in mice using rAAV genome editing. This disease is caused by deficiency in the human gene F9, encoding a coagulation factor. In this work, mice were injected with a rAAV8 targeting the albumin gene that contained the human F9 gene. Upon integration of the vector, the human F9 gene was fused to the albumin gene. The two genes were separated by the sequence for a 2A peptide (derived from plus-strand RNA viruses) that allows multiple proteins to be encoded by a single mRNA through exon skipping. Thus, the F9 gene would be expressed to the same level as albumin, whilst not affecting albumin expression. Mice were treated at 2 days or 8 weeks of age by injection of the vector. Subsequent HR into the albumin gene led to expression of F9 up to 20% above the normal level of F9 (compared to non-infected controls). This effect was maintained for several months in both 2-day and 8-week old mice (at the time of treatment).

rAAV genome editing of specific loci has also been performed in other model organisms. An example of this, by Hickey et al. (2011), involved the targeting of porcine fibroblasts for somatic cell nuclear transfer-based generation of genetically altered pigs. In this work, rAAV targeting of the fumarylacetoacetate hydrolase gene generated heterozygous knock-out cells that were used for the generation of heterozygous knock-out pigs. Deficiency of this gene in humans causes hereditary tyrosinaemia, a disease causing childhood cirrhosis and liver cancer. The heterozygous knock-out pigs displayed normal tyrosine levels but reduced enzyme levels and activity, and may be a useful model for future studies of this disease.
Prior to this, porcine models of cystic fibrosis were also developed using a similar approach by Rogers et al. (2008). Here, the authors generated pigs with either null CFTR alleles or a mutation that recapitulates the most common mutation observed in human patients (deletion of amino acid 508, phenylalanine).

rAAV has not only been used to create and study models of human disease, but also animal disease. Richt et al. (2007) produced cows lacking the normal cellular prion protein that, when misfolded, causes bovine spongiform encephalopathy (BSE) in cattle and Creutzfeldt-Jakob disease (CJD) in humans. Cattle lacking this protein were assessed for a number of parameters (e.g. behavioural, reproductive, haematological etc.) and were identified to be no different to normal cattle, but are more resistant to prion protein propagation. Such animals may provide future insight into prion disease, and, as suggested by the authors, may be industrially important for the manufacture of bovine-derived products that are devoid of prion protein.

The work described above is a small selection of the works that have utilised rAAV genome editing for a number of purposes. As can be seen from the diversity in these projects, the rAAV technique provides a flexible method of genome editing.

1.5: Experimental Approaches Taken in this Work
Genome editing is becoming an important technology in the research of multiple diseases with known genetic components (as well as non-disease-related genetic research). Coronary artery disease is one of the biggest sources of mortality globally and has a strong genetic component. GWAS have identified nearly 50 loci associated with this disease; identifying the causative variants and mechanisms that are acting at these loci is important in an attempt to better understand the aetiology of CAD.

In this work, two loci with putative causative SNPs/mechanisms were investigated by genome editing. These loci are the 1p13 locus (putative causal SNP rs12740374; Musunuru et al. 2010) and the 9p21 locus (putative causal SNPs rs10811656 and rs10757278; Harismendy et al. 2011). More details regarding the 1p13 and 9p21 loci can be found in chapters 4 and 5 respectively, together with the data obtained from the genome editing and functional experiments performed to investigate these loci.

These two loci were selected in order to validate previous publications that proposed the identification of causal SNPs in these regions. Thus, these two projects were chosen as proof-of-principle investigations, in order to demonstrate the utility and applicability
of rAAV genome editing for complex disease genetic research as a means by which specific variants can be examined directly.

Genome editing using the rAAV method was chosen as a method by which to investigate the SNPs at these loci. At the time of commencement of this work, the genome editing methods available were limited to the ZFN and TALEN approaches (CRISPRs had not yet been developed for this purpose). Further, TALENs were in their infancy and had not been widely used. ZFN technology was not chosen because of limitations with the technique, namely the expense of plasmid construction, the potential need for multiple designs, and the risk of off-target effects.

rAAV technology was therefore chosen due to the low off-target rates observed and the reliance of the technique upon HR to mediate the integration of the vector. rAAV technology is licensed by Horizon Discovery Group plc (Cambridge, UK). The Department of Cardiovascular Sciences (University of Leicester, UK) used this technology under license from Horizon Discovery Group plc as part of a Centre of Excellence scheme, which provides access to the technology for academic institutions.

The works discussed in this thesis were the first projects to utilise this technique at the University of Leicester. Because of this, optimisation of the method was required with the cell lines chosen for the individual projects, prior to functional analysis of the cell lines generated.

1.6 Aims and Objectives
A total of four main aims for this work were established prior to commencement of the projects. These were parental cell line characterisation (chapter 3), optimisation and generation of protocols for use of the technique with the chosen cell lines (chapters 4 and 5) and the functional investigation of CAD-associated, putative-causal SNPs at the 1p13 (chapter 4) and 9p21 (chapter 5) loci. As an aside, the use of mesenchymal stem cells (MSCs) with the rAAV technique was also attempted. MSCs derived from the Wharton’s jelly of the umbilical cord from multiple donors were used for the 1p13 and 9p21 projects (chapters 4 and 5, respectively). Previously, MSCs derived from adipose tissue (Li et al. 2011) and bone marrow (Ito et al. 2004) have been shown to be transducible by rAAV. In this work, Wharton’s jelly-derived MSCs (WJ-MSCs) were selected due to their availability and to determine if they were suitable for rAAV genome editing, potentially providing disease-relevant cell types through differentiation post-targeting.
To investigate the purportedly-causal 1p13 and 9p21 SNPs, rAAV genome editing was used to alter the genotype of these SNPs at these loci, generating isogenic cell lines that differ at just the target variants. The discernment of genotype-specific effects upon transcription factor binding and gene expression (the proposed mechanisms through which these SNPs are thought to act) can therefore be determined directly whilst controlling for background variation.

Functional analysis of genome-edited cell lines consisted of assessing the effect of genotype upon transcription factor binding at the sites of the SNPs. Following this, the expression of local genes was measured and compared between genotype groups in order to determine the effect of SNP genotype upon local gene regulation.

Recently, another locus associated with CAD (6p24) has been investigated using a similar approach with the CRISPR/Cas9 system. Beaudoin et al. (2015) used the CRISPR/Cas9 system to investigate the effect of the deletion of a transcription factor binding site (for further details, please refer to section 6.1.1). The work presented herein is, to the best of our knowledge, the first to describe the analysis of CAD-associated regulatory SNPs using rAAV genome editing.
Chapter 2: Materials and Methods

In this section, details of cell line culture and all experimental procedures are given. For protocols requiring the handling of cells, the materials required are given only when they are unique to that protocol, or are not described in section 2.1: Cell Lines and Culture Techniques. Further, common laboratory consumables (such as 1.5ml microcentrifuge tubes) and reagents (e.g. molecular biology-grade water) are not included in the materials lists but their use is described in the protocols as appropriate. All PCR primers were purchased from Fisher Scientific (UK) or Eurofins Genomics (Germany).

2.1: Cell Lines and Culture Techniques
For both the 9p21 and 1p13 genome editing projects, several cell lines have been used. In the following sub-sections, details for each cell line are given, including the source of the cell line and the media and techniques required for culture. All cell lines were incubated at 37°C in a humidified incubator with 5% carbon dioxide. All culture media was sterile filtered with a 0.22µm syringe filter after addition of supplements and before use. Media was stored at 4°C and warmed in a 37°C incubator prior to use. Media supplements were stored at -20°C prior to addition to culture media. Non-heat inactivated foetal calf serum (FCS) was purchased from Gibco (Life Technologies, UK) and 1% penicillin/streptomycin (10,000U/ml stock solution) was purchased from Hyclone (GE Healthcare Life Sciences, UK). All work with virus-infected cell lines was carried out in a tissue culture laboratory in a biosafety level II cabinet reserved for this purpose. Standard or virus-containing waste was disposed of in accordance with local guidelines. All procedures detailed in subsequent sections utilise these media and culture conditions unless otherwise specified.

2.1.1: EA.hy926
EA.hy926 cells were kindly provided by Dr Dave Lodwick (University of Leicester, UK). These cells were cultured in Dulbecco’s Modified Eagle Medium (DMEM; high glucose with L-glutamine, Sigma, UK) supplemented with 10% foetal calf serum (FCS) and 1% penicillin/streptomycin.
2.1.2: HCT116
HCT116 cells were generously provided by Horizon Discovery Group plc (UK) and were cultured in DMEM/F12K media, which comprises a 1:1 mix of DMEM (Sigma, UK) and F12 Kaighn’s (Hyclone, UK) media, both supplemented with 10% non-heat inactivated FCS and 1% penicillin/streptomycin prior to preparation of complete media.

2.1.3: HEK293T
HEK293T cells were kindly provided by Horizon Discovery Group plc (UK), and were cultured in high glucose DMEM with L-glutamine media (Sigma, UK), supplemented with 10% FCS and 1% penicillin/streptomycin.

2.1.4: HeLa-S3
HeLa-S3 cells were purchased from Cell Lines Service (Germany), and were cultured in DMEM/F12K media (see section 2.1.2) supplemented with 10% FCS and 1% penicillin/streptomycin.

2.1.5: HepG2
HepG2 cells were purchased from the ATCC, USA (catalogue number ATCC HB-8065). These cells were cultured with Eagle’s Minimum Essential Medium with L-glutamine (EMEM, Corning, UK), supplemented with 1% penicillin/streptomycin and 10% non-heat inactivated FCS.

2.1.6: Mesenchymal Stem Cells (MSCs)
Mesenchymal stem cells, isolated from umbilical cord, were a generous gift from Mrs. Sukhvir Rai. These cells were cultured in media containing 45% Hams F12 (PAA, UK) and 45% α-MEM (Sigma, UK) supplemented with 10% FCS, 1% antibiotic/antimycotic (Invitrogen, UK) and 10ng/ml Basic Fibroblast Growth Factor (Peprotech, USA).

2.1.7: THLE-3
THLE-3 cells were obtained from the ATCC, USA (catalogue number ATCC CRL-11233). These cells were cultured in collagen-coated T75 flasks (collagen was derived from calf skin; 0.1% in 0.1M ethanoic acid, Sigma, UK) using BEGM media (Lonza, UK) supplemented with BEGM Bullet Kit (Lonza, UK), 10% FCS and 1% penicillin/streptomycin. Coated flasks were prepared by combining 2.55ml of collagen
solution with 47.45ml PBS. 8.8ml of the diluted collagen was transferred to a T75 flask and spread out to cover the flask completely. 5ml additional PBS was then added to the flask, which was then incubated in a safety cabinet for 2 hours. After incubation, the coating solution was removed and discarded. The flasks were stored for up to 2 weeks at 4°C.

### 2.1.8: Passaging Procedures

Passaging of all cell lines (for either general culture or for the specified protocols in the following sections) utilised a phosphate buffered saline (PBS, Oxoid, UK) wash after removal of culture media and was followed by a trypsin/EDTA treatment. MSC cells were treated with 1x trypsin/EDTA, whereas all other cell lines were treated with 2x trypsin/EDTA (10x trypsin/EDTA comprises 5g/l trypsin and 2g/l EDTA and was purchased from VWR, UK). 1x and 2x working stocks were made from 10x stock by dilution with PBS. Cells were incubated in trypsin/EDTA at 37°C/5% CO₂ until complete dissociation of cells was observed microscopically. Trypsin neutralisation was performed by adding 2.5 volumes of complete media to the trypsin/cell suspension and a single cell suspension made by trituration. Dilution of the cells in appropriate complete culture media was performed in a new flask or plate.

### 2.1.9: Cell Counting

When required, plating of cells at specific densities utilised this procedure and reagents. 8μl of tritutated trypsinised cell suspension (neutralised with media) was mixed with 8μl 0.4% trypan blue in PBS (Hyclone, UK). 8μl of this was applied to a haemocytometer (cleaned with 70% IMS and dried thoroughly) replete with a coverslip. The number of cells in two of the large outer squares was counted, and this value was multiplied by 1x10⁴ to give the number of cells per ml of suspension. This value was used to calculate volumes of cell suspension required to provide a desired final cell number.

### 2.1.10: Cell Line Cryopreservation

All cells were preserved using the same procedure. A freeze mix comprising 50% FCS, 40% complete media and 10% DMSO (Sigma, UK) was prepared, using complete media appropriate for the cell line. Cells were trypsinised and neutralised as described in section 2.1.8, and transferred to a 50ml Falcon tube. This was centrifuged at 1000rpm
for 5 minutes. The supernatant was discarded and the cell pellet was resuspended with freeze mix. This was aliquotted into labelled cryovials and frozen at -80°C in a polystyrene tube box for at least 24 hours before being transferred to liquid nitrogen.

2.1.11: Cell Line Ageing

Generation of clonal cell lines by rAAV-mediated homologous recombination requires long-term culture of cells. Therefore, during the procedure the cells undergo a large number of cell divisions and therefore may display an “aged” phenotype. To control for this, parental cells that have not been targeted with a rAAV can be cultured in such a way as to closely mimic the conditions experienced by targeted cells; the protocol for this technique is described below.

It should be noted that this protocol only produces aged controls that account for the prolonged cell culture of the edited cell lines. As a result, the cell lines produced from this procedure have not been exposed to G418 selection (section 2.4.2), nor have they undergone transfection with the pML-Cre plasmid for Cre-mediated selection cassette removal (section 2.4.8). This was the case for the aged HepG2 and HeLa-S3 controls.

Aged controls for the heterozygous HCT116 edited cells were obtained from parental cells infected with rAAV that recombined the selection cassette but not the SNPs of interest into the target 9p21 locus. These cells underwent G418 selection and Cre-mediated removal of the selection cassette. The resulting controls had the same SNP genotype as the parental HCT116 cells, but were suitable controls for age, G418 exposure, and pML-Cre transfection of the heterozygous edited cell lines. The following ageing protocol was then utilised using these controls to produce aged lines for homozygous (double knock-in) edited HCT116 cells. These cells were only age-matched for these cells, and as such had not been exposed to two rounds of G418 selection and pML-Cre transfection.

Protocol:

30 cells/well were plated in a 96-well plate (200μl media per well). Plates were cultured for 2 weeks prior to counting the number of colonies in each well. 2 wells with <10 colonies were expanded to a T75 culture and passaged once using a splitting ratio equal to the number of colonies in the original well (e.g. if 3 colonies, the flask was split 1:3). If a large number of colonies were in the well, the flasks were passaged multiple times, equal to the number of colonies (e.g. if 10 colonies, split 1:5 then 1:2). These
procedures were repeated twice more to mimic the cell divisions the cells would undergo during the G418 selection, pre-Cre transfection single cell dilution and post-Cre transfection single cell dilution. Cell lines were named after the wells from which the colonies were taken at each stage of the procedure. As two wells were taken forward each time, a total of 8 individual aged lines was generated.

2.2: Parental Cell Line Characterisation
2.2.1: Proliferation Analysis Using SRB Assay

Materials Required:
- 0.4% Sulphorhodamine B (SRB, Sigma, UK) in 1% acetic acid
- 10% Trichloroacetic acid (TCA, Sigma, UK)
- 10mM Tris-base (Sigma, UK)
- 1% Acetic acid (Fisher, UK)
- BioTek ELx808 or ELx800 plate readers (BioTek, USA)

Protocol:
For each cell line and treatment condition, 10,000 cells were plated into 8 wells of 96-well plates, with one plate for each day of analysis (please see individual experiment for further details). 8 wells containing just culture media were included for background control per plate. After 16-18 hours, the day 0 plate was fixed by aspirating the media and adding 100μl 10% TCA to each well and incubating at room temperature for 20 minutes. The TCA was then aspirated and the wells washed three times with 200μl distilled water. After the final wash, 100μl 0.4% SRB reagent was added to each well and incubated for 15 minutes. The SRB was removed and the plate rinsed 3 times with 100μl 1% acetic acid. The plate was then left to dry at room temperature in the dark. On subsequent days, the remaining plates were fixed as described. Data for all plates was obtained following fixation. 100μl 10mM Tris-Base was added to each well and incubated for 10 minutes at room temperature prior to measuring absorbance on a BioTek ELx808/ELx800 plate reader with a 490nm filter. Data were analysed by calculating the mean absorbance value for each cell line and subtracting the mean background reading. Data were then normalised to day 0 results to account for starting cell number and plotted graphically. Standard deviations are reported on the graphs as error bars. After normalisation, the log₁₀ of the data is calculated and plotted against...
time. The linear part of the growth curve was identified, and the gradient of this line identified. The doubling time of each line was calculated using the equation below:

\[
\text{Doubling time} = 24(\log_{10}(2)/\text{gradient})
\]

### 2.2.2: G418 Death Curve

**Materials Required:**
- 100mg/ml G418 (Invitrogen, UK) in ultra-pure water, 0.22μm sterile-filtered

**Protocol:**
A 2x dilution series was prepared according to table 2.1. 100μl of each G418 concentration was added to 3 wells of a 96-well plate for each cell line. 1000 cells, in 100μl media, were plated in each well and incubated for 2 weeks. After culturing, all wells were assessed to determine the lowest concentration at which all cells are killed in all 3 wells – this is the optimal concentration of G418 to be used for selection. Optimal concentrations for each cell line are given in section 2.4.2.

<table>
<thead>
<tr>
<th>Final Concentration G418 (mg/ml)</th>
<th>Volume of G418 100mg/ml Stock Solution (μl)</th>
<th>Volume of Media (μl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>10.5</td>
<td>339.5</td>
</tr>
<tr>
<td>1.4</td>
<td>9.8</td>
<td>340.2</td>
</tr>
<tr>
<td>1.3</td>
<td>9.1</td>
<td>340.9</td>
</tr>
<tr>
<td>1.2</td>
<td>8.4</td>
<td>341.6</td>
</tr>
<tr>
<td>1.1</td>
<td>7.7</td>
<td>342.3</td>
</tr>
<tr>
<td>0.9</td>
<td>6.3</td>
<td>343.7</td>
</tr>
<tr>
<td>0.8</td>
<td>5.6</td>
<td>344.4</td>
</tr>
<tr>
<td>0.7</td>
<td>4.9</td>
<td>345.1</td>
</tr>
<tr>
<td>0.5</td>
<td>3.5</td>
<td>346.5</td>
</tr>
<tr>
<td>0.3</td>
<td>2.1</td>
<td>347.9</td>
</tr>
<tr>
<td>0.1</td>
<td>0.7</td>
<td>349.3</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>350.0</td>
</tr>
</tbody>
</table>

**Table 2.1: G418 Dilution Series used for G418 Death Curve.** Given value of final concentration refers to final experimental concentration; the concentration after preparing these dilutions is 2x the final concentration as a 1:1 dilution with complete media is performed.
2.2.3: rAAV2-GFP Infection

Materials Required:
- rAAV2-GFP virus (Vector Biolabs, USA) at known titre.
- Evos FL Cell Imaging System (Life Technologies, UK)

Protocol:
50,000 cells were plated into 3 wells of a 12-well plate for each cell line to be examined. After culturing for 24 hours, rAAV2-GFP virus was added to give a range of multiplicities of infection (MOI; up to 100,000 genome copies/cell). The plates were incubated for 2-3 days and cells analysed by fluorescence microscopy using an Evos FL Cell Imaging System. The proportion of GFP-expressing cells for each MOI tested was assessed.

2.2.4: Optimisation of Diagnostic PCR Assays for 1p13 and 9p21 Projects

Reaction conditions for the post-selection PCR assays require optimisation to determine the suitability of each potential diagnostic assay primer pair, for inclusion of primers in the vector sequence. A combination of touchdown PCR and gradient PCR was utilised using genomic DNA from one cell line from each project. Genomic DNA was isolated using the GenElute Mammalian DNA miniprep kit (Sigma, UK) using the manufacturer’s protocol. All PCR was performed using GoTaq Hot Start DNA Polymerase from Promega (UK) with dNTPs obtained from Sigma (UK) (see section 2.2.4.1 for reaction mix).

Initially, touchdown PCR using the recommended cycling conditions from Horizon Discovery Group plc were tested to identify if a single band corresponding to the parental (“wild-type”, WT) product could be obtained (section 2.2.4.2). If not, gradient PCR using standard cycling conditions (section 2.2.4.3) was used to test the assays. If an annealing temperature was observed to produce the target product, a modified touchdown PCR was tested that incorporated this annealing temperature in the final cycling step.

During optimisation, standard gel electrophoresis was used (1% agarose gel, variable voltage applied between experiments to optimise separation of the ladder marker used – see section 2.4.5 for further details regarding the optimised diagnostic PCR protocols). A general protocol for gel electrophoresis is given in section 2.2.4.4. Please refer to other sections for experiment-specific gel electrophoresis conditions.
2.2.4.1 GoTaq PCR Reaction Mix

The following mix was used for all PCR reactions using GoTaq Hot Start DNA polymerase (Promega, UK). Further, experiment-specific, details are provided in the relevant sections.

- 4µl GoTaq Colourless Flexi buffer (5x)
- xµl forward primer (10µM)
- xµl reverse primer (10µM)
- 0.2µl dNTPs (25mM each)
- 1.2µl MgCl₂ (50mM)
- 0.2µl GoTaq Hot Start polymerase (5Uµl⁻¹)
- xµl template DNA (concentration variable)
- Water to 20µl final volume

2.2.4.2: Standard Touchdown PCR Protocol

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>94</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>64</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>60/kb</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>61</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>60/kb</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>58</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>60/kb</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>57</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>60/kb</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Final Extension</td>
<td>70</td>
<td>300</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>Final Extension</td>
<td>70</td>
<td>300</td>
</tr>
</tbody>
</table>

Table 2.2: Standard Touchdown PCR Protocol for use with GoTaq Hot Start DNA Polymerase
2.2.4.3: Standard Gradient PCR Protocol

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>94</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>55-65</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>60/kb</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Final Extension</td>
<td>70</td>
<td>300</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>Hold</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.3: Standard Gradient PCR Protocol for use with GoTaq Hot Start DNA Polymerase

2.2.4.4: Standard Gel Electrophoresis Protocol

Materials Required:
- Agarose powder (Melford, UK)
- 25x TAE buffer (Life Technologies, UK), diluted in water to 1x
- GelRed Nucleic Acid Gel Stain, 10,000x in water (Biotium, USA)
- Glassware (Duran bottle, conical flask etc)
- Balance
- Microwave
- 6x blue loading dye (30% glycerol [Sigma, UK], 0.25% bromophenol blue [Fisher, UK], 69.75% ultra-pure water)
- Gel casting tray, combs, tank and autoclave tape
- Ladder, as appropriate for samples (see individual sections for experiment-specific ladders)
- Power pack and leads
- Gel documentation instrument (Gene Genius Bioimaging System, Syngene, UK) and software (GeneSnap v.6.03, Syngene, UK)

Protocol:
Gel casting trays were prepared by sealing the two open ends with autoclave tape and placing the comb into the slots. Agarose solutions were prepared by weighing agarose into suitably-sized glassware. The amount of agarose required is calculated using the formula:
Mass agarose \( (g) = \frac{\text{Tray volume (ml)}}{100} \times \text{Final % agarose} \)

1x TAE was added to the agarose (equal to the volume of the tray, as used in the equation above). 1μl of GelRed Nucleic Acid Stain was added for every 10ml of 1x TAE. Glassware was covered (but not sealed) and microwaved until the agarose had dissolved completely. Molten agarose was poured into the gel tray and allowed to set completely (at least 20 minutes) prior to comb removal. Samples were prepared by adding 6x loading dye to a final concentration of 1x. Once set, the autoclave tape at either end of the gel tray was removed and the tray placed into the gel tank filled with 1x TAE. Samples and ladders were pipetted into the wells and the lid of the tank closed. Electrophoresis was performed at an experiment-specific voltage for as long as required for the samples to migrate a suitable distance for visualisation. UV transillumination was performed using a Gene Genius Bioimaging System with GeneSnap v.6.03 software (Syngene, UK).

2.2.4.5: EA.hy926 9p21 Genomic Instability Determination

Materials Required:
- GoTaq Hot-Star DNA polymerase kit (Promega, UK)
- dNTPs (25mM each, Sigma, UK)
- DNA samples (conc. ~100-150ng/μl)
- Primers:
  - 9p21 Vector 1 Assay_F: 5’-TCACAGCCCTGAAGGACAAT-3’
  - 9p21 Vector 1 Assay_R: 5’-GGGCAGCAGATATCTTTCCA-3’
  - Chr1_F: 5’-ATACTTGGTACCCCTTATTTCTCAAGGCAAAC-3’
  - Chr1_R: 5’-TGATTTCTCGAGTTGGAATCCAGGCAAGTGTTTTC-3’
- Thermal cycler (G-Storm GS4, G-Storm, UK)
- Gel electrophoresis equipment (section 2.2.4.4)

Protocol:
The following PCR mix was prepared for each sample and cycled according to the conditions given in table 2.4. Gel electrophoresis was performed using a 1% agarose gel ran at 125V. 30μl of each sample (comprising 25μl sample and 5μl 6x loading dye) was loaded per well, with 5μl of Hyperladder 1kb ladder (Bioline, UK) per row of samples.
- 5μl GoTaq Flexi Colourless buffer (5x)
- 1μl forward primer (10μM)
- 1μl reverse primer (10μM)
- 0.25μl dNTPs (25mM each)
- 1.5μl MgCl\(_2\) (50mM)
- 1μl template DNA
- 0.25μl GoTaq Hot-Start DNA polymerase (5U/μl)
- 15μl water

![Table 2.4: Cycling Conditions for PCR Assays to Assess EA.hy926 9p21 Genomic Instability](chart)

2.2.5: SNP Genotyping by Sanger Sequencing

In order to identify the genotype of the SNPs of interest in the cell lines to be used for genome editing, and to verify the sequence of the region, Sanger sequencing was used. It was also used for genotyping the genome-edited cell lines. PCR amplification of the region containing the SNP(s) followed by gel electrophoresis to confirm successful PCR preceded Sanger sequencing. All DNA samples were pure genomic DNA isolated using the GenElute Mammalian DNA miniprep kit (Sigma, UK) using the manufacturer’s recommended protocol.

In the following sections, oligonucleotide primer sequences and PCR conditions are given for the two regions of interest, as well as a general Sanger sequencing protocol.

2.2.5.1: SNP Genotyping PCRs

Materials Required:
- MyTaq Hot Start DNA polymerase (Bioline, UK)
- Primers:
  - SEQ SNP rs12740374 F = 5’-GAACACATTTTCAGGGGAGC-3’
- SEQ SNP rs12740374 R = 5'-GGATGGGTTCGTGTCCAGT-3'
- SEQ 9p21 SNPs F = 5'-GTAGGTACAGAATACAAAG-3'
- SEQ 9p21 SNPs R = 5'-GCCCTGACATGGCTAGACTC-3'

Protocol:
A PCR mix comprising the following was prepared. Samples were cycled according to the parameters given in tables 2.5 and 2.6 for 1p13 and 9p21 loci, respectively.

PCR Mix (per sample):
- 5µl MyTaq buffer (5x)
- 1µl forward primer (10µM)
- 1µl reverse primer (10µM)
- 0.5µl MyTaq Hot Start DNA Polymerase enzyme (5Uµl⁻¹)
- 50-100ng template DNA
- Water to 25µl final volume

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>95</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>Denaturation</td>
<td>95</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>64.3</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>72</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Final Extension</td>
<td>72</td>
<td>600</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>Hold</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.5: rs12740374 Sequencing PCR Cycling Conditions

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>95</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>Denaturation</td>
<td>95</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>58</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>72</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Final Extension</td>
<td>72</td>
<td>600</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>Hold</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.6: rs10811656/rs10757278 Sequencing PCR Cycling Conditions
Samples were analysed using gel electrophoresis. A 1% gel ran at 100V was used for all samples. 5μl of each sample was run with 1μl 6x loading dye. 10μl PCRSizer 100bp ladder (Norgen, Canada) was included.

2.2.5.2: Sanger Sequencing Protocol

Materials Required:
- ExoSAP-IT (Affymetrix, UK)
- BigDye Terminator v3.1 (Applied Biosystems, UK)
- BigDye Terminator v1.1/3.1 Sequencing Buffer, 5x (Applied Biosystems, UK)
- DyeEx 2.0 Spin kit (Qiagen, UK)

Protocol:
5μl of PCR product was combined with 2μl of ExoSAP-IT and incubated at 37°C for 15 minutes, followed by 15 minutes at 80°C. For each sequencing reaction, the following mix was prepared and cycled according to table 2.7. Samples were purified using Qiagen’s DyeEx 2.0 Spin kit (following the manufacturer’s protocol). Samples were submitted to the Protein and Nucleic Acid Chemistry Laboratory (PNACL) at the University of Leicester, UK. Sequence traces were analysed using FinchTV (Geospiza Inc., USA) software.

Sequencing Mix (per sample):
- 2μl BigDye Terminator v1.1/3.1 Sequencing Buffer (5x)
- 2μl ExoSAP-IT treated PCR product
- 0.5μl BigDye v3.1 Terminator
- 0.4μl primer (10μM)
- 5.1μl water

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>Denaturation</td>
<td>96</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>50</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>60</td>
<td>240</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Hold</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.7: Sanger Sequencing Cycling Protocol
2.2.6 SNP Genotyping by Microarray

DNA samples for parental and a selection of targeted cell lines were submitted to Affymetrix (UK) for genotyping as part of a larger, unrelated cohort of samples, on the Axiom™ Genome-Wide UKB WCSG Genotyping Array (comprising 845,487 probesets covering 825,928 SNPs). DNA was isolated from cells (using the recommended protocol) with the Sigma (UK) GenElute Mammalian Genomic DNA miniprep kit. DNA quantification was performed using a NanoDrop 8000 (Thermo Scientific, UK). DNA samples were diluted to 15ng/μl in 10mM Tris buffer, 0.1mM EDTA, pH 8.0. After dilution, spectrophotometric analysis was performed to confirm the concentration, and that the 260/280 ratio was between 1.8 and 2.0, and the 260/230 ratio is greater than 1.5, as recommended by Affymetrix (UK). After preparing samples with these specifications, at least 50μl of each sample was submitted for analysis. SNP genotype calling was performed by a statistician in the Department of Cardiovascular Sciences, University of Leicester, UK.

2.3: rAAV2 Production and Titre Estimation

2.3.1: Virus Production

Materials Required:

- HEK293T cells (5x T175 flasks at 70% confluency, low passage [<20])
- Optim-MEM media, no phenol red (Life Technologies, UK)
- Water bath at 37°C
- Dry ice/ethanol bath
- pAAV plasmid (supplied after DNA synthesis by Horizon Discovery Group plc, UK)
- pDG-Helper plasmid (Plasmid Factory, Germany)
- Lipofectamine LTX kit (Invitrogen, UK)
- ViraKit AAV Purification kit (Virapur, USA)
- Benzonase Nuclease (≥250U/μl; Sigma, UK)
- Lysis buffer (0.15M NaCl, 50mM Tris-HCl, pH 8.5; Sigma, UK)
- Vacuum pump (Integra Vacusafe, SLS, UK)
Protocol:
Ten T75 flasks were each seeded with $4 \times 10^6$ HEK293T cells, with cells taken from 5 T175 flasks. Seeded flasks were incubated for 24 hours. The following day, the media in the flasks (with cells at ~70% confluency) was removed and replaced with 9ml complete media before returning to the incubator for at least 30 minutes. Meanwhile, the transfection mix was prepared by combining 187.5µg of each plasmid (pAAV vector and pDG-helper plasmid) and 375µl of Plus reagent (from the Lipofectamine LTX kit) in a final volume of 10ml of serum-free media. This was incubated at room temperature for 5 minutes prior to addition of 560µl Lipofectamine LTX. The resulting mix was incubated at room temperature for 30 minutes. 1ml of the transfection mix was added to each T75 flask. Flasks were rocked gently to mix and returned to the incubator. 24 hours later, cell viability was assessed visually. After 3 days, virus purification was performed using Virapur’s ViraKit AAV Purification kit as follows.

Media was aspirated from all flasks and retained in Falcon tubes (labelled “Media”). 3ml PBS was added to each flask and was used to wash cells. PBS was removed and stored in a Falcon tube (“PBS”). Cells were trypsinised completely by incubation at 37°C with 2ml 2x trypsin and gentle agitation. Cell suspensions from all flasks were combined together into a Falcon tube (“Cells”) and 20ml complete media added to neutralise the trypsin. The “Media”, “PBS” and “Cells” tubes were centrifuged at 1000xg for 5 minutes. The supernatant from the “PBS” tube was discarded. The supernatant of the “Media” tube was transferred to new tubes (“Clarified Media”). The media was discarded from the “Cells” tube. All cell pellets were retained and resuspended in 10ml complete media and combined into one 50ml Falcon (labelled “All Cells”). The “All Cells” tube was centrifuged at 1000xg for 5 minutes and the supernatant removed. The “Clarified Media” and “All Cells” tubes were frozen at -80°C until purification was performed.

The “Clarified Media” and “All Cells” tubes were thawed in a 37°C water bath. After 10 minutes, the “All Cells” tube was removed and 3ml of lysis buffer was added. The tube was vortexed briefly and returned to the water bath. After a total of 30-40 minutes of thawing, both tubes were removed from the water bath and frozen again in a dry ice/ethanol bath for 10 minutes. Once frozen for 10 minutes, the tubes were thawed for 10 minutes in the water bath, vortexing the “All Cells” tube occasionally. This freeze/thaw procedure was performed twice more. The contents of the two tubes was combined together and aliquotted between 50ml Falcon tubes. The tubes were
centrifuged at 2900rpm for 30 minutes. The supernatant from all tubes was collected together into a T75 flask. For every 10ml media in the flask, 1μl of Benzonase nuclease was added. Using a stripette, the media was mixed gently to distribute the nuclease. The flask was then incubated upright at 37°C in an incubator for 30 minutes.

Meanwhile, the kit components were prepared in the tissue culture cabinet according to the manufacturer’s protocol. After incubation, the contents of the T75 flask was applied to the filter unit and the virus particles purified according to the manufacturer’s instructions. Particles were eluted with 1.5ml Elution buffer 3 and were stored at -80°C in 500μl aliquots. A 20μl aliquot (stored at 4°C) was retained for titre determination (section 2.3.2).

2.3.2 Virus Titre Determination
Materials Required:
- Amplification-grade DNaseI with 10x buffer (Sigma, UK)
- Sodium Dodecyl Sulphate, 20% (SDS, Sigma, UK)
- Sodium Chloride (Sigma, UK)
- Proteinase K (Sigma, UK)
- SensiMix SYBR No Rox 2x master mix (Bioline, UK)
- G418_F primer: 5’-ACCTTGCTCCTGCCGAGAAAGTAT-3’
- G418_R primer: 5’-CGATGTTTCGCTTGGTGGTCGAAT-3’
- Rotor-Gene Q instrument (Qiagen, UK)
- Rotor-Gene strip tubes and caps (Qiagen, UK)

Protocol:
Virus samples were DNase treated to eliminate any residual genomic DNA by combining the following and incubating at 37°C for 30 minutes and then at 95°C for 2 minutes.

- 5μl purified virus
- 5μl DNase buffer (10x)
- 10U Amplification-grade DNaseI
- Water to a final volume of 50μl
Then, proteinase K treatment of the sample to release the viral DNA from the capsid was performed by mixing the following and incubating at 56°C for 1 hour, followed by 5 minutes at 95°C:

- 50µl DNaseI-treated virus sample
- 75µl proteinase K buffer (1% SDS, 200µM NaCl in water)
- 1µl proteinase K (20mg/ml)
- 24µl water

After proteinase K treatment, the sample was diluted 1:10 to dilute any PCR inhibitors and subjected to qPCR, using the following mix per sample (prepared as a master mix):

- 12.5µl SensiMix SYBR No Rox master mix (2x)
- 1µl G418_F primer (10µM)
- 1µl G418_R primer (10µM)
- 9.5µl water

Triplicate samples of 24µl were prepared, and 1µl of diluted virus sample, standard curve dilution (table 2.8) or water (for no template control, NTC) was added. Samples were cycled in a Rotor-Gene Q according to table 2.9.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Template DNA (µg/ml)</th>
<th>Number of Copies per ml of G418R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000X10⁰</td>
<td>2.821x10¹¹</td>
</tr>
<tr>
<td>2</td>
<td>2.000x10¹</td>
<td>5.642x10¹⁰</td>
</tr>
<tr>
<td>3</td>
<td>4.000x10²</td>
<td>1.128x10¹⁰</td>
</tr>
<tr>
<td>4</td>
<td>8.000x10³</td>
<td>2.257x10⁹</td>
</tr>
<tr>
<td>5</td>
<td>1.600x10⁴</td>
<td>4.514x10⁸</td>
</tr>
<tr>
<td>6</td>
<td>3.200x10⁴</td>
<td>9.027x10⁷</td>
</tr>
<tr>
<td>7</td>
<td>6.400x10⁴</td>
<td>1.805x10⁷</td>
</tr>
<tr>
<td>8</td>
<td>1.280x10⁵</td>
<td>3.611x10⁶</td>
</tr>
<tr>
<td>9</td>
<td>2.560x10⁶</td>
<td>7.222x10⁵</td>
</tr>
<tr>
<td>10</td>
<td>5.120x10⁷</td>
<td>1.444x10⁵</td>
</tr>
<tr>
<td>11</td>
<td>1.024x10⁷</td>
<td>2.889x10⁴</td>
</tr>
<tr>
<td>12</td>
<td>2.048x10⁸</td>
<td>5.777x10³</td>
</tr>
</tbody>
</table>

Table 2.8: G418R Virus Titre Assay Standard Curve
Table 2.9: G418 Virus Titre Assay Cycling Conditions. For use with SensiMix SYBR No Rox master mix (2x) and Rotor-Gene instrument.

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>95</td>
<td>600</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denaturation</td>
<td>95</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>Annealing/Extension</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

Upon completion of the run, the “Quantitation” analysis program within the Rotor-Gene Q software (Qiagen, UK) was used to analyse the data and obtain a titre value from the unknown samples, based upon the known values of the standard curve samples. The standard curve was assessed for linearity according to the following parameters:

- Correlation coefficient of determination ($r^2$) >0.99
- Gradient of the line (m) = -3.300+/-0.3
- Efficiency = 1.00+/-0.1

Additionally, variation in replicate C_T values was assessed; replicates were accepted if the difference between samples was no greater than 0.5 cycles. The titre of the virus was calculated by multiplying the titre value obtained by 300 (to account for dilution of the virus sample during the viral DNA isolation procedure).

2.4: Generation of Isogenic Cell Lines by Genome Editing: Virus Infection, Antibiotic Selection and Cre Recombination of the Selection Cassette

2.4.1 Cell Line Infection

Protocol:
5x10^5 cells were plated into two T75 flasks in 10-12ml fresh media. These were incubated overnight at 37°C/5% CO₂. The following day, the virus aliquots were defrosted on ice. Meanwhile, media was removed from the flasks and replaced with 5ml fresh media. One flask was labelled “Infection” and one “Control”. When thawed, 150μl of virus was added to the “Infection” flask only. Both flasks were incubated for 3 days.
2.4.2: Infected Cell Selection

Materials Required:
- Sterile-filtered G418 in water (100mg/ml, Invitrogen, UK)

Protocol:
After 3 days of infection, several 96-well plates were seeded with a variety of cell numbers. For each seeding density, 1 plate was seeded with cells from the “Control” flask. Media was supplemented with an appropriate concentration of G418 (based upon the cell line). The optimal plating densities and G418 concentrations used are given in table 2.10. Media, cells and G418 were combined in sterile Duran bottles or Falcon tubes prior to plating. The calculations below show how the volumes of each were calculated:

\[
Volume \ of \ G418 \ Stock \ (V_{G418})(\mu l) = 1000 \left(\frac{20nG}{S}\right)
\]
\[
Volume \ of \ media \ (ml) = 20n - \frac{100An}{B} - V_{G418}
\]
\[
Volume \ of \ cell \ suspension \ (\mu l) = 1000 \left(\frac{100An}{B}\right)
\]

Where:
- \( n = Total \ number \ of \ plates \ required \)
- \( A = Number \ of \ cells \ per \ well \)
- \( B = Cell \ suspension \ concentration \ (cells/ml) \)
- \( G = Final \ G418 \ concentration \ (mg/ml) \)
- \( S = Stock \ G418 \ concentration \ (mg/ml) \)

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Optimal Seeding Densities (cells/well)</th>
<th>G418 Selection Concentration (mg/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HepG2</td>
<td>2000</td>
<td>0.8</td>
</tr>
<tr>
<td>Mesenchymal stem cells</td>
<td>2000-3000</td>
<td>0.3</td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>1000</td>
<td>1.1</td>
</tr>
<tr>
<td>HCT116</td>
<td>1000</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Table 2.10: Cell Line Seeding Densities and G418 Selection Concentrations

After seeding, the plates were incubated for 2 weeks, or until there was complete cell death in the control plates. Colony counting was then performed (section 2.4.3).
2.4.3 Selected Cell Colony Counting
The number of colonies in each well was counted manually using a microscope, with the number of colonies recorded. During counting, in cases where it appeared as if multiple colonies had merged into one large colony of cells, such wells were classified as being derived from >1 colony. Wells containing colonies were marked on the plate lid with a permanent marker. Typically, wells with <10 colonies were considered acceptable to take forward for further analysis.

2.4.4 Cell Consolidation and DNA Isolation
Materials Required:
- 96-well PCR plates and adhesive film seals (Geneflow, UK)
- DirectPCR Lysis buffer (Viagen, USA)
- Proteinase K from *Tritirachium album* (800U/ml; Sigma, UK)
- Thermal cycler (G-Storm GS4, G-Storm, UK)
- For freezing plates:
  - Plate freeze mix (14% DMSO, 50% FCS, 36% complete media)
  - Parafilm
  - Paper towel and zip-lock plastic bags

Protocol:
After counting, wells containing colonies were consolidated into new 96-well plates in cases where there were few colony-containing wells on each plate. This step was not performed in cases where the majority of wells contained colonies. To perform the consolidation, wells were trypsinised with 30μl 2x trypsin until the cells had detached, then they were neutralised with fresh media (170μl) and transferred to a new plate. The new location of each well was recorded for traceability of the samples. Consolidated plates were incubated for at least 48 hours prior to DNA isolation.

DNA was isolated from plates using Viagen’s DirectPCR lysis buffer. PCR plates containing 19.09μl of DirectPCR reagent and 0.91μl of Proteinase K per well were prepared. Cells were washed with 200μl PBS and trypsinised with 30μl 2x trypsin. Upon dissociation, cells were triturated and 6μl of cell suspension was added to the corresponding well in the PCR plate. These were sealed with an adhesive film and
incubated in a PCR cycler for 15 minutes at 55°C followed by 45 minutes at 85°C. DNA samples were stored at -20°C.

The remaining cells in the 96-well tissue culture plate were either neutralised with 175μl complete media and returned to the incubator, or frozen at -80°C. For this, 80μl of plate-freeze mix was added to each well. Plates were sealed with parafilm, wrapped in tissue paper, placed in a plastic bag and stored in a -80°C freezer. Plates were thawed, when required, by removing the plastic bag and paper towel, and placing the parafilm-wrapped plates in a zip-lock bag. This was submerged in a 37°C water bath until almost completely defrosted. The plates were removed from the bag and 100μl of complete media added. Plates were incubated for 24 hours before media was replaced.

2.4.5 Identification of Successfully Targeted Cell Pools by Diagnostic PCR Screen
It should be noted that the below primer sequences and cycling conditions for the 9p21 vector refer to the 2nd generation virus only, as this is the vector that was used to make all targeted lines. A discussion of this is given in sections 5.2 and 5.3.

Materials Required:
- GoTaq Hot Start DNA Polymerase kit (Promega, UK)
- dNTPs (25mM each, Sigma, UK)
- 96-well PCR plates and seals (Geneflow, UK) or strip caps (Fisher, UK)
- Thermal cycler (G-Storm GS4, G-Storm, UK)
- Primers:
  - 1p13_Left_F = 5’-CCCTTTCCAAAGAGGATAG-3’
  - 1p13_Left_R = 5’-CCCTGCTTCTTGAATTCTGC-3’
  - 9p21_Left_F = 5’-TGTTTGCACCAAAAGAGGTG-3’
  - 9p21_Left_R = 5’-GCGCAATACCACAGTGAACA-3’
- Gel electrophoresis equipment, including 6x loading dye (section 2.2.4.4)
- Hyperladder 1kb (Bioline, UK), diluted 1:4 with water and 6x loading dye (3μl Hyperladder 1kb, 2μl 6x loading dye and 7μl water per lane).
Protocol:
All reagents and DNA samples were defrosted on ice. The following was prepared, as a master mix, for each sample.

- 4µl GoTaq Flexi Colourless buffer (5x)
- 0.15µl forward primer (10µM)
- 0.15µl reverse primer (10µM)
- 0.2µl dNTPs (25mM each)
- 1.2µl MgCl$_2$ (25mM)
- 0.2 µl GoTaq Hot Start DNA polymerase (5U/µl)
- 12.1µl water

18µl of master mix was added to each well of a 96-well plate. 2µl of DNA was added to this and the plates sealed securely with adhesive film or strip caps. Samples were cycled according to table 2.11. Gel electrophoresis following PCR was performed. 4µl of 6x loading dye was added to each sample, and 8µl was loaded onto a 104-well (300ml) gel. A 0.8% agarose gel was used for the 1p13 diagnostic PCRs, and a 1% gel was used for the 9p21 samples. 30µl of GelRed Nucleic Acid Stain was used per gel. 12µl of a 1:4 dilution of Hyperladder 1kb ladder was used as a marker. Gels were electrophoresed at 65V and visualised using UV transillumination. Table 2.11 shows the expected wild-type (WT) and knock-in (KI) products, derived from non-targeted and targeted chromosome DNA, respectively.
<table>
<thead>
<tr>
<th>Locus</th>
<th>Homology Arm Screened</th>
<th>Cycling Conditions</th>
<th>Number of Cycles</th>
<th>WT band (bp)</th>
<th>KI Band (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p13</td>
<td>Left</td>
<td>Initial Denaturation 95°C for 180 seconds</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denaturation 95°C for 15 seconds</td>
<td>3</td>
<td>1730</td>
<td>1509</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing 64°C for 30 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension 70°C for 120 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denaturation 95°C for 15 seconds</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing 61°C for 30 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension 70°C for 120 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denaturation 95°C for 15 seconds</td>
<td>45</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing 59.3°C for 30 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension 70°C for 120 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Final Extension 70°C for 300 seconds</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hold 10°C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9p21</td>
<td>Left</td>
<td>Initial Denaturation 95°C for 180 seconds</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Denaturation 95°C for 15 seconds</td>
<td>40</td>
<td>2250</td>
<td>1563</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing 58°C for 30 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension 70°C for 155 seconds</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Final Extension 70°C for 300 seconds</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hold 10°C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.11: Cycling Conditions and Expected Bands for 1p13 and 9p21 Diagnostic PCR Assays.** “WT” bands size refers to parental chromosome PCR product, whilst “KI” refers to a band derived from PCR amplification of targeted chromosomes.
2.4.6 Expansion of Diagnostic PCR Positive Pools

Protocol:
Following diagnostic PCR screening of DNA samples, expansion of diagnostic positive samples is required. Where necessary, frozen plates were revived according to the procedure described in section 2.4.4 and cultured for at least 48 hours prior to expansion.

When wells reached ~70% confluency, they were trypsinised with 30μl 2x trypsin after a PBS wash. When detached, the cells were tritutated and 170μl complete media added. The entire contents of the wells were transferred to a 24-well plate containing 300μl media. Plates were cultured until the wells reached ~70% confluency, at which point they were PBS washed and trypsinised with 150μl 2x trypsin, triturated, and neutralised with 350μl complete media. The 500μl cell suspension was transferred to a 6-well plate and 1500μl complete media added. The cells were incubated until ~70% confluent. The wells were then washed with PBS, trypsinised with 500μl of 2x trypsin and neutralised with 500μl complete media after trituration. The resulting 1ml of cell suspension was transferred to a T75 flask, and allowed to culture prior to expansion into several flasks for banking, with one flask kept for either single cell dilution or pML-Cre transfection according to the following:

- Single cell dilution (section 2.4.7) – performed if there were more than 3 colonies in the expanded well and/or the transfection efficiency of the cell line is low
- pML-Cre transfection (section 2.4.8) – performed if there were 1-3 colonies in the expanded well (this protocol requires a single cell dilution as part of the procedure, but a prior single cell dilution was preferred for >3 colony pools when using a cell line with low transfection efficiency).

2.4.7 Single Cell Dilution

This protocol was used to isolate clonal, isogenic cell lines that originate from a single parental cell, producing a pure cell line. The technique employed to achieve this is the single cell dilution, whereby a known number of cells is diluted so as to obtain very low cell numbers when plated in 96-well plates. These individual cells are then cultured, allowing expansion to form large populations of clonal, isogenic cells.

Cell survivability is reduced at low cell numbers. To circumvent this issue, there are two solutions:
• Increase the number of plates seeded with single cells to increase the probability of colony growth in some of the wells. This method is not preferred as the large number of plates needed requires significant handling time and large volumes of media.

• Increase the number of cells plated per well so as to result in the growth of a single colony more frequently. 2 or more cells are plated into each well, which increases the survivability slightly, but only to a level where one of the cells proliferates and the remainder fail to do so.
  
  o The following are optimised cell numbers (per well) for the cell lines used in these projects to obtain a suitable number of single colonies:
    • HepG2: 2-5
    • HeLa-S3: 1
    • HCT116: 1

Materials Required:

• 10cm Petri dishes, sterile
• Swinging bucket centrifuge, capable of holding Falcon tubes
• Duran bottle(s)

Protocol:

A cell suspension at 10,000 cells/ml was prepared after a PBS wash and trypsinisation of a T75 flask of cells at ~70% confluency. The volume of media required per seeding density was aliquotted into sterile Duran bottles (20ml used per 96-well plate). To the bottle, an appropriate volume of cell suspension (at 10,000 cells/ml) was added, calculated using the formula below:

\[
Volume\ of\ cell\ suspension\ (\mu l) = 10cn
\]

where:

\[
c = \text{number of cells per well}
\]

\[
n = \text{number of plates}
\]

The contents of each Duran bottle were swirled gently to generate a homogenous cell suspension. This was transferred to a 10cm dish and a multi-channel pipette was used to aliquot 200μl into each well of the 96-well plates. Plates were incubated for 2-3 weeks, with media replaced as necessary.
When colonies could be observed microscopically, the number of colonies in each well were counted and recorded. All wells containing a single colony were marked on the lid of the plate. Once colonies were large enough for DNA isolation, DNA samples were obtained using the DirectPCR buffer (Viagen, USA) according to section 2.4.4. DNA samples were analysed by diagnostic PCR according to section 2.4.5. Any diagnostic PCR-positive colonies were expanded according to section 2.4.6 and transfected with pML-Cre to remove the selection cassette (section 2.4.8.2).

2.4.8 pML-Cre Plasmid Preparation and Selection Cassette Removal by pML-Cre Transfection

2.4.8.1 pML-Cre Plasmid Transformation and Purification

Materials Required:
- pML-Cre plasmid (~200ng; provided by Horizon Discovery Group plc, UK)
- DH5α competent *Escherichia coli* cells (Bioline, UK)
- Sterile 2.5% Luria broth (LB, Sigma, UK) media in water
- Sterile 2.5% Luria broth (Sigma, UK) media in water with 0.1mg/ml ampicillin (Sigma, UK)
- Sterile LB agar with 0.1mg/ml ampicillin (2.5% LB, 1.5% agar from Sigma, UK and Difco, UK respectively; ampicillin from Sigma, UK).
- Heat block set at 42°C
- Shaking 37°C incubator
- 10cm Petri dishes, sterile
- 37°C incubator
- Bunsen burner
- 100% ethanol
- Metal spreader
- QIAprep spin miniprep kit (Qiagen, UK)
- NanoDrop 8000 (Thermo Scientific, UK)

Protocol:

pML-Cre plasmid was transformed into DH5α *E. coli* cells by combining 50μl of DH5α cells with ~200ng of plasmid DNA. The sample was mixed gently, and incubated on ice for 30 minutes. Then, the sample was heated at 42°C for 45 seconds on a heat block,
followed by 2 minutes on ice. 100μl of LB media was added and the sample incubated at 37°C in a shaking incubator (200rpm) for 1 hour. 100μl of transformed cells were plated on LB agar plates with ampicillin and incubated at 37°C overnight. Following this, a single colony was picked and cultured in 5ml LB media with ampicillin in universal tubes overnight. A glycerol stock was taken from one of the tubes for long-term storage of the transformed bacteria by combining 600μl 100% glycerol with 400μl of culture and storing at -80°C. The remaining culture was used for plasmid preparation using the Qiaprep spin miniprep kit (Qiagen, UK), using the recommended protocol. DNA concentration was determined using a NanoDrop 8000 spectrophotometer. Eluted plasmid was stored at -20°C.

2.4.8.2: Selection Cassette Removal by pML-Cre Transfection

Materials Required:
- pML-Cre plasmid (5μg per transfection)
- Lipofectamine LTX kit (Invitrogen, UK)
- Optim-MEM media, no phenol red (Life Technologies, UK)

Protocol:
For each cell line to be transfected, 1.0x10^6, 1.5x10^6, and 2x10^6 cells were plated into a 6-well plate and incubated overnight. After 16-18 hours, the wells were inspected to identify which seeding density produced 70-80% confluency. The media was removed from the wells that were not 70-80% confluent, and the plate was returned to the incubator whilst the transfection mix was prepared. 5μg of pML-Cre plasmid (per transfection) was diluted in Opti-MEM to a final volume of 125μl. 5μl of Plus reagent was added, mixed, and incubated at room temperature for 7 minutes. In a separate tube, 15μl of Lipofectamine LTX and 110μl of Opti-MEM were combined. The 125μl of diluted Lipofectamine LTX reagent was added to the diluted plasmid (total volume 250μl) and mixed gently. The transfection mix was incubated at room temperature for 15 minutes before 750μl of Opti-MEM was added. The media from the selected well in the 6-well plate was removed and replaced with 1ml of fresh, complete media. The 1ml transfection mix was added, drop-wise, to the well and agitated gently to mix. The cells were incubated for 3 hours before the media in the well was removed and replaced with 2ml complete media. 16-24 hours later, 96-well plates were seeded with single cells (detailed in section 2.4.7). Plates were grown for ~2 weeks prior to colony counting of

60
all wells. DNA was harvested from wells containing single colonies using DirectPCR lysis buffer as described in section 2.4.4. Plates were screened using the Cre PCR assay (section 2.4.9).

2.4.9 PCR Screening of pML-Cre Transfected Single-Cell Derived DNA Samples

Materials Required:
- 96-well PCR plates and adhesive seals (Geneflow, UK) or strip caps (Fisher, UK)
- GoTaq Hot Start DNA polymerase kit (Promega, UK)
- 10mM dNTPs in water (Sigma, UK)
- Cre screening primers (see table 2.12 for primer sequences)
- Swinging bucket centrifuge with plate holders.
- Thermal cycler (G-Storm GS4, G-Storm, UK)
- Gel electrophoresis equipment, including 6x loading dye (section 2.2.4.4)
- PCRSizer 100bp ladder (Norgen, Canada)

Protocol:
A PCR master mix was prepared with the following per sample:
- 4µl GoTaq Flexi Colourless buffer (5x)
- 0.15µl Cre primer X (10µM)
- 0.15µl Cre primer Y (10µM)
- 0.3µl Cre primer Z (10µM)
- 0.2µl dNTPs (10mM each)
- 1.2µl MgCl₂ (25mM)
- 0.2µl GoTaq Hot Start DNA polymerase (5Uµl⁻¹)
- 11.8µl water

18µl of master mix was aliquotted into the wells of 96-well PCR plates. 2µl of template DNA was added to each well and the plate was sealed with adhesive film or strip caps. The plate was inverted several times to mix and centrifuged at 3000rpm for 1 minute to collect the sample. Plates were cycled according to the conditions given in tables 2.13 (1p13 project) or 2.14 (9p21 project). Gel electrophoresis was performed using a 2% 104-well (300ml) gel. 4µl of 6x loading dye was added to each sample, before loading 8µl of sample onto the gel with 10µl PCRSizer 100bp ladder, with 1 ladder well per 12
samples. Gels were electrophoresed at 100V for ~1 hour 45 minutes, until the blue dye had migrated within 1-2cm of the end of the gel. Gel images were taken with the Syngene Gene Genius Bioimaging system and assessed to determine which bands were present.

The following approximate band sizes are expected for all Cre assays (exact sizes are given in tables 2.12):

- ~400bp – corresponds to targeted chromosome from which the selection cassette has not been removed.
- ~234bp – only amplified after successful removal of the selection cassette by Cre recombination.
- ~200bp – produced from non-targeted chromosome

There are several possible combinations of bands that can be observed:

- ~200bp only – non-targeted cells have proliferated alongside the targeted cells during the selection process and have not been diluted out. This allows for growth of non-targeted colonies after the Cre transfection single cell dilution.
- ~234bp only – successful Cre recombination of a DKI line; no WT chromosomes, and hence ~200bp product, are present.
- ~200bp and ~400bp – the Cre transfection has failed to remove the selection cassette from successfully targeted cells.
- ~200bp and ~234bp – successful Cre recombination has occurred, removing the selection cassette. The additional 34bp present in the 234bp product arises from a loxP site that remains behind after Cre recombination. Amplified from a single KI sample.
- ~200bp, ~234bp and ~400bp – successful Cre recombination has occurred, but there is contamination with cells that have not undergone Cre recombination. These arise from samples containing more than one colony despite appearing as a single colony during counting.
Table 2.12: Expected Band Sizes (with Primer Combinations) and Primer Sequences for Cre PCR Screening

<table>
<thead>
<tr>
<th>Locus</th>
<th>Expected Possible Band Sizes (bp)</th>
<th>Primer Sequences (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p13</td>
<td>Y+Z = 190</td>
<td>X = GCTACTTCCATTTGTCACGTCC</td>
</tr>
<tr>
<td></td>
<td>Y+Z = 224</td>
<td>Y = TCCAGCTATTTGGAGCAGTG</td>
</tr>
<tr>
<td></td>
<td>X+Z = 408</td>
<td>Z = TACAGCACTGTCACTGTCCG</td>
</tr>
<tr>
<td>9p21</td>
<td>Y+Z = 202</td>
<td>X = TGGTGGAATGGCAGG</td>
</tr>
<tr>
<td></td>
<td>Y+Z = 236</td>
<td>Y = GACACTCCACAAATGGGC</td>
</tr>
<tr>
<td></td>
<td>X+Z = 400</td>
<td>Z = GGTCGAACCTAAAAGCCAAG</td>
</tr>
</tbody>
</table>

Table 2.13: Cycling Conditions for 1p13 Cre PCR Assay

<table>
<thead>
<tr>
<th>Step</th>
<th>Number of Cycles</th>
<th>Cycle Stage</th>
<th>Temperature (°C)</th>
<th>Incubation Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>94</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>64</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>61</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>59.3</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Final Extension</td>
<td>70</td>
<td>300</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>Hold</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>Step</td>
<td>Number of Cycles</td>
<td>Cycle Stage</td>
<td>Temperature (°C)</td>
<td>Incubation Time (seconds)</td>
</tr>
<tr>
<td>------</td>
<td>-----------------</td>
<td>-------------</td>
<td>------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Initial Denaturation</td>
<td>94</td>
<td>180</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>64</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>61</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>59</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>Denaturation</td>
<td>94</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annealing</td>
<td>58</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Extension</td>
<td>70</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Final Extension</td>
<td>70</td>
<td>300</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>Hold</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.14: Cycling Conditions for 9p21 Cre PCR Assay

Cre positive cell samples were expanded and banked following procedures 2.4.6 and 2.1.10 respectively. Genotype determination was performed using restriction endonuclease digestion (section 2.4.10) and/or Sanger sequencing (section 2.2.5). DNA was obtained either from DirectPCR lysis buffer (Viagen) or Sigma’s GenElute Mammalian DNA miniprep kit using the recommended protocol.

2.4.10 SNP Genotyping by Restriction Endonuclease Digestion

Restriction digests were performed according to the procedures described in table 2.15. Samples were incubated in 0.2ml tubes in a thermal cycler (G-Storm GS4, G-Storm, UK) using the indicated conditions, followed by gel electrophoresis as described.
<table>
<thead>
<tr>
<th></th>
<th>1p13</th>
<th>9p21</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP(s)</td>
<td>rs12740374</td>
<td>rs10811656/rs10757278</td>
</tr>
<tr>
<td>Enzyme (concentration, supplier)</td>
<td><em>BsmHII</em> (10U/μl, New England Biolabs, UK)</td>
<td><em>HpaII</em> (10U/μl, Thermo Scientific, UK)</td>
</tr>
<tr>
<td>Supplied Buffer Name (conc.)</td>
<td>CutSmart buffer (10x)</td>
<td>Fast Digest buffer (10x)</td>
</tr>
<tr>
<td>Volume of Supplied Buffer per Digest Sample (μl)</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>Volume of Enzyme (μl)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Volume of PCR Product (μl)</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Volume of Water (μl)</td>
<td>39.5</td>
<td>8</td>
</tr>
<tr>
<td>Total Digest Volume (μl)</td>
<td>50</td>
<td>15</td>
</tr>
<tr>
<td>Incubation Temperature (°C)</td>
<td>65</td>
<td>37</td>
</tr>
<tr>
<td>Incubation Time (minutes)</td>
<td>240</td>
<td>180</td>
</tr>
<tr>
<td>Volume of Digest Mix Analysed (μl)</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Volume of 6x Loading Dye (μl)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total Electrophoresis Sample Volume (μl)</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>Agarose Concentration (%)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Voltage Applied (V)</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>Banding if Homozygous Non-Risk (bp)</td>
<td>519</td>
<td>163 + 263</td>
</tr>
<tr>
<td>Banding if Heterozygous (bp)</td>
<td>248 + 272 + 519</td>
<td>163 + 263 + 426</td>
</tr>
<tr>
<td>Banding if Homozygous Risk (bp)</td>
<td>248 + 272</td>
<td>426</td>
</tr>
</tbody>
</table>

Table 2.15: SNP Genotyping by Restriction Digestion Procedure

2.5 RNA Isolation Using Qiagen RNeasy Mini Kit

Materials Required:

- Swinging-bucket rotor centrifuge
- Microfuge
- RNeasy mini kit (Qiagen, UK)
- RNase-free DNaseI set (Qiagen, UK)
- Amplification-grade DNaseI (Sigma, UK)
- NanoDrop 8000 UV-Vis spectrophotometer (Thermo Scientific, UK)
Protocol:
RNA was isolated from cells by centrifuging a suspension of cells at 1000rpm for 5 minutes. The supernatant was discarded and the pellet resuspended in 1ml of PBS. The cell suspension was transferred to a 1.5ml Eppendorf tube and centrifuged at 1000rpm for 5 minutes. The PBS was removed and the cell pellet subjected to RNA isolation using the Qiagen RNeasy mini kit, following the manufacturer’s protocol, eluting in 50μl of RNAse-free water. DNaseI treatment was performed either with the on-column DNAseI digestion kit (Qiagen) or using amplification-grade DNaseI (Sigma), following the recommended protocols for either treatment method. RNA yield was determined by NanoDrop 8000. Samples were stored at -80°C.

2.6 SuperScript III Reverse Transcription
Materials Required:
- RNA (defrosted and stored on ice at all times)
- SuperScript III kit components (Invitrogen, UK)
- Thermal cycler (G-Storm GS4, G-Storm, UK)

The SuperScript III system (Invitrogen) was used for the majority of reverse transcription reactions. This system was used for the successful optimisation of all qPCR assays except for the IRF-1 and ANRIL transcripts. Optimisation of the ANRIL assays was not possible with this system. An alternative, the SensiFAST cDNA synthesis kit, was successfully used for the optimisation of the ANRIL qPCR assays (see section 2.7). Further, due to the availability of the SensiFAST reverse transcribed cDNA, the IRF-1 assay was also optimised with this kit.

Protocol:
2100ng of RNA was reverse transcribed using the SuperScript III system using oligo dT priming, using the recommended protocol, except RNAse OUT was not used and was replaced with water. The recommended RNAse H step was followed (final volume 21μl). “No RT” reactions, containing all components except the reverse transcriptase, were also prepared for all RNA samples. cDNA was diluted to a final concentration of 10ng/μl with water prior to storage at -20°C.
2.7 SensiFAST cDNA Synthesis Reverse Transcription

Materials Required:

- RNA (on ice)
- SensiFAST cDNA synthesis kit (Bioline, UK)
- Thermal cycler (G-Storm GS4, G-Storm, UK)

This kit was used as an alternative to the SuperScript III system for the ANRIL transcript assays, and, due to cDNA availability, the IRF-1 assay. With the SensiFAST cDNA synthesis kit, all ANRIL assays could be optimised.

When performing qPCR experiments, the RNA samples of interest were reverse transcribed, by default, using SuperScript III. Where applicable to the genes of interest (i.e. ANRIL or IRF-1), SensiFAST cDNA synthesis was also performed. The reference gene 36B4 was tested using both cDNAs; normalisation to 36B4 was performed in a like-for-like manner (e.g. gene data from SensiFAST cDNA was normalised to 36B4 data obtained from the same SensiFAST cDNA samples).

Protocol:
1000ng of RNA was reverse transcribed using the SensiFAST cDNA synthesis kit according to the manufacturer’s protocol, with inclusion of the optional incubation step (48°C for 15 minutes). “No RT” reactions were also prepared for all samples. Final cDNA concentration is 50ngμl⁻¹; this was diluted to 10ng/μl working concentration with water before storage at -20°C.

2.8 Design and Optimisation of qPCR Assays

qPCR assays using the SYBR Green chemistry were utilised for all qPCR approaches in this work. Assays were designed to produce amplicons of 100-300bp, and, for gene expression analyses (where possible), to span large introns to prevent the possibility of amplifying contaminating genomic DNA. Primer pairs were designed either by eye or using the Primer-BLAST program, hosted by NCBI (Ye et al. 2012), or were obtained from previous publications; MTAP, DMRTA1 and ANRIL exon 18-19 primers were taken from Harismendy et al. (2011). The remaining ANRIL primer sequences were taken from Holdt et al. (2010). Assays were analysed using Primer-BLAST to assess the number of potential off target products as well as the complementarity scores (ideally low) for the primers and the predicted melting temperature (T_m) of the primers.
Once designed, all assays were tested and optimised to produce a single PCR product with no primer dimerisation. To achieve this, cycling conditions and primer concentrations were adjusted and optimised prior to testing of the assay with a standard curve.

Initially, gradient PCR was performed (using SensiMix SYBR No Rox 2x master mix) using an annealing temperature of 55-65°C. Samples were analysed by gel electrophoresis (section 2.2.4.4) using a 100ml, 2% agarose gel with an applied voltage of 100V. Based upon the gel electrophoresis banding observed, further iterations of optimisation were performed as necessary, altering parameters such as annealing time and temperature, cycle number, use of a combined annealing/extension step, addition of extra magnesium chloride to that already present in the master mix, final primer concentration, and (for the ANRIL transcripts analysed in the 9p21 project, see section 2.10), the reverse transcription kit used. Each optimisation step was analysed using gel electrophoresis. An assay was determined to be optimised when a single product of the correct size was produced to an acceptable level. Once optimised, all assays were subjected to standard curve analyses using three replicates per cDNA concentration to determine if the assay amplifies in a linear manner across a range of template concentrations. Standard curves can also be used to determine the template concentrations at which the assay fails to detect changes in concentration (the intervening concentrations at which the assay works as expected is termed the linear dynamic range of the assay). Further, standard curves can be used to determine various parameters that determine an assay’s suitability for use (discussed below).

Two-fold dilutions of template DNA were used for each standard curve. For gene expression assays, typically 100ng-6.25ng/reaction was tested. Generally, this range of template was sufficient for most samples and assays. Therefore, the linear range of the assay could not be determined from the curve (as these concentrations of template produced an acceptable curve with all data points); however, it does mean that the assay is acceptable at the concentrations tested. It should be noted that if the take-off values (see below) for samples of interest fell outside the previously tested, known linear range of an assay, the standard curve was repeated with a wider range of template concentrations. This was performed to ensure the sample take-off values observed remained within the detection limits.linear range of the assay. An alternative method would have been to adjust the volume of the template in order to bring the sample into the known linear range of the assay. However, repeating the standard curve was the
preferred method because expansion of the linear range not only confirms linearity for the current samples of interest, but also for future samples.

All standard curves were analysed using two separate programs from within the Rotor-Gene Q software; the “Quantitation” and “Comparative Quantification” options. For the Quantitation analysis method, cycle threshold (C_T) values were calculated automatically by the software for each sample, as well as various parameters about the assay. These include the efficiency of the assay, the gradient of the line produced by log transformation of the given concentration, and the r^2 value (the degree to which the equation of the line calculates concentration based upon the given concentrations).

Standard curves were accepted if they met the following criteria:

- Efficiency values of between 0.90 and 1.10, with 1.00 as the desired value.
- Slope gradient values of between 3.00 and 3.60, with 3.30 as the target value.
- r^2 values of above 0.99
- Triplicate C_T values that do not differ by more than 0.5 cycles

Further to the above method, standard curves were also analysed using the Comparative Quantification method. This method does not calculate C_T values; instead “take-off” values are calculated from the raw data of each sample independently of each other. For this, the 2^{nd} derivative of each data point of the amplification plot data is used. The take-off value is the cycle number that is 20% of the maximum value for the sample.

Additionally, amplification efficiency is also calculated from the 2^{nd} derivative data for each sample. Therefore, the comparative quantification method calculates efficiency and take-off values for each sample independently, which are then used by the program to determine the ratio of one sample to a calibrator sample (which is set to 1.00).

When analysing a standard curve using this software, an amplification value of 1.6 or above is acceptable, with 2.0 the target value. Furthermore, the take-off values between triplicates should not differ by more than 0.5 cycles. To calculate the r^2 value for the assay, as well as the gradient of the straight line, the log_{10} of the known concentration of each sample was calculated. The take-off values were plotted against the log-transformed concentrations (as a scatter plot) in Microsoft Excel. By drawing a linear trendline (line of best fit) on the resulting graph, the gradient and r^2 values are
determined. The same values are accepted for the comparative quantification analysis method as for the quantitation algorithm.

2.9 1p13 Gene Expression Analysis

Materials Required:

- Qiagen RNEasy RNA kit (section 2.5)
- SuperScript III Reverse Transcription kit (section 2.6)
- SensiMix SYBR No Rox master mix, 2x (Bioline, UK)
- Primers, 10μM – see table 2.16
- Rotor-Gene Q instrument (Qiagen, UK)
- Corbett CAS1200 liquid handling robot (Corbett, UK)
- Bench Top Ioniser (Charles Water, UK)
- 100-well Rotor-Discs (Qiagen, UK)
- Rotor-Disc adhesive seals (Qiagen, UK)
- Conducting, disposable tips (Qiagen, UK)
- Rotor-Disc Heat Sealer (Qiagen, UK)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Forward Primer Sequence (5'-3')</th>
<th>Reverse Primer Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>36B4</td>
<td>TCGACAATGGCAGCATCTAC</td>
<td>GCCTTGACCTTTTCAGCAAG</td>
</tr>
<tr>
<td>SORT1</td>
<td>TCTTGGGGTTGGAGATAGCAGC</td>
<td>TTCCCTCCAGACACCTCTTGCT</td>
</tr>
<tr>
<td>CELSR2</td>
<td>GTCGACCTCAAAACCCGCTTCCT</td>
<td>CTCACAGTATGGCCCAAGGT</td>
</tr>
<tr>
<td>PSRC1</td>
<td>TGCCCACTGTGAACCTCTTTTG</td>
<td>GAGAAGATGCTGGGCTTTTG</td>
</tr>
<tr>
<td>SARS</td>
<td>TATGCAAGCAAGACAAATCGGAG</td>
<td>TCATCAATGAGGAGTCCGG</td>
</tr>
<tr>
<td>MYBP4L</td>
<td>ACATCCTGTGATTGAGAGGC</td>
<td>AGCCCTTCGACCCGTGATC</td>
</tr>
<tr>
<td>PSMA5</td>
<td>GTACGACAGGGGGCCGTGAATA</td>
<td>TTCTCCACAGCTAGGCCACAC</td>
</tr>
<tr>
<td>WDR47</td>
<td>TCTCCCTGGAGCTCCTACTC</td>
<td>GAAAGATCCTGGGTGTGCA</td>
</tr>
<tr>
<td>TMEM167B</td>
<td>TTTGACACTGTGCCTACTTC</td>
<td>CACGCTGTTCCAATCACAG</td>
</tr>
</tbody>
</table>

Table 2.16: Primer Sequences of 1p13 qPCR Assays with 36B4 Reference Assay

Protocol:

Three independent RNA samples were isolated from each cell line at three different passages (1 week apart). 2100ng of each RNA sample was reverse transcribed using the SuperScript III kit (section 2.6) and diluted to a working concentration of 10ng/μl in water.
A master mix was prepared, composed of the following for each sample (tested in triplicate). Primer concentrations used are given in table 2.17. Three NTC samples were included per run.

- 12.5µl SensiMix SYBR No Rox (2x)
- XnM forward primer (10µM)
- XnM reverse primer (10µM)
- Water to a final volume of 22µl

Robotic setup of each reaction was performed; 22µl of master mix and 3µl of cDNA was added to each qPCR reaction. 100-well Rotor-Discs were used which were sealed with adhesive seals. Reactions were cycled according to the conditions given in table 2.17. All runs required an initial denaturation of 95°C for 10 minutes and, upon completion of each run, a 60-99°C melt curve was performed.

Data from each run was analysed firstly by assessing the melt curve and accepted if a single peak was present for all cDNA samples. Further, the NTC and “No RT” samples were accepted if they showed no amplification, or showed a melt peak that is not present in any cDNA samples (indicative of primer dimers). Absence of such peaks from cDNA samples was acceptable, as such products were not affecting quantification of the target product. Additionally, “no RT” samples were permitted to show amplification of the target amplicon when the take-off value was greater than 5 cycles (32-fold smaller) than the cDNA samples.

Once run data was assessed for acceptability, the “comparative quantification” algorithm was used (part of the Rotor-Gene Q program). As three independent batches of RNA were obtained, these were treated as independent experiments when analysed. Thus, the three WT samples were used as the calibrator for their respective experiments. This provided a ratio of all samples within the experiments compared to WT cells. The algorithm calculates this ratio by firstly calculating the mean amplification efficiency for all samples in the run. Then, the mean take-off value for the sample of interest is subtracted from the mean take-off value of the calibrator. The resulting value is used as the exponent of the amplification efficiency to give a ratio between the two samples.

Outlier replicates were removed from the data. An outlier was defined as any replicate that showed no amplification or had a take-off value that is >0.5 cycles different from the remaining two replicates. Removal of only a single replicate per triplicate was permitted.
After outlier removal, data were exported to Microsoft Excel for analysis; the ratio values obtained from the comparative quantification were used. For each sample, the gene of interest (GOI) value was divided by the reference gene value, to normalise for RNA loading etc. Thus, the resulting values for all experiments/GOIs were fold changes in expression, normalised to WT cells and a reference gene.

By performing this calculation for all samples, 3 values are obtained per cell line. To determine if there is a genotype-specific effect upon expression of the genes of interest, statistical analysis was performed comparing the homozygous risk (GG) with the heterozygous (GT) lines. For these comparisons, the Mann-Whitney test was performed using Stata/IC v.13.1 software (StatCorp LP, USA). *p*-values were obtained for all group combinations of interest.

### Table 2.17: 1p13 qPCR Assay Final Primer Concentrations and Cycling Conditions

<table>
<thead>
<tr>
<th>Gene</th>
<th>Final Primer Concentration (mM)</th>
<th>Denaturation Temperature and Time</th>
<th>Annealing Temperature and Time</th>
<th>Extension Temperature and Time</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>36B4</td>
<td>300</td>
<td>95°C for 15 seconds</td>
<td>58°C for 15 seconds</td>
<td>72°C for 20 seconds</td>
<td>35</td>
</tr>
<tr>
<td>SORT1</td>
<td>500</td>
<td>95°C for 15 seconds</td>
<td>58°C for 45 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>CELSR2</td>
<td>300</td>
<td>95°C for 15 seconds</td>
<td>58°C for 45 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>PSRC1</td>
<td>500</td>
<td>95°C for 15 seconds</td>
<td>58°C for 45 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>SARS</td>
<td>500</td>
<td>95°C for 15 seconds</td>
<td>58°C for 45 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>MYBPHL</td>
<td>400</td>
<td>95°C for 15 seconds</td>
<td>61°C for 30 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>PSMA5</td>
<td>200</td>
<td>95°C for 15 seconds</td>
<td>60°C for 15 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>WDR47</td>
<td>200</td>
<td>95°C for 15 seconds</td>
<td>65.1°C for 15 seconds</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>TMEM167B</td>
<td>400</td>
<td>95°C for 15 seconds</td>
<td>58.1°C for 10 seconds</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

2.10 9p21 Gene Expression Analysis

The subsequent procedures were each repeated on three separate occasions to reduce the likelihood of external variables influencing the data obtained from the experiment. These experiments are referred to “experiment 1, 2 or 3” as required.

2.10.1 Interferon-γ Treatment

Materials Required:

- Recombinant Human Interferon-γ (PeproTech, USA) diluted to a final concentration of 20mg/µl in PBS containing 0.1% BSA (Sigma, UK)
Protocol:
250,000 cells were plated into the wells of a 6-well plate. For each treatment condition, two wells were plated (e.g. 2 untreated, 2 IFN-γ treated). These wells were identified by the following: “U-A” and “U-B” for the two untreated wells, and “T-A” and “T-B” for the two treated wells. Cells were incubated overnight. Following incubation, the cells were washed once with 1ml PBS after removing the culture media. 2ml of fresh culture media was added to each well. 10μl of 0.1% BSA in PBS was added to the “U” wells (untreated vehicle control) and 10μl of 20ng/ml IFN-γ in 0.1% BSA/PBS. Cells were incubated for 24 hours.

2.10.2 RNA Isolation of Interferon-γ Treated Cells With Zymo Direct-Zol RNA Miniprep kit
Due to the large number of samples required for this experiment, the Zymo Direct-Zol RNA Miniprep kit was used for isolation of RNA instead of the Qiagen RNEasy kit.

Materials Required:
- Zymo Direct-Zol RNA Miniprep kit with TRI-Reagent (Zymo Research, USA)

Protocol:
24 hours after IFN-γ stimulation, cells were washed with 2ml PBS. 950μl of TRI-Reagent was added to each well and incubated for 5 minutes. The lysed cells were triturated thoroughly and transferred to a 1.5ml tube. RNA was isolated from each sample using the Zymo Direct-Zol RNA Miniprep kit following the recommended protocol. RNA was eluted in 50μl RNAse-free water, and stored at -80°C.

2.10.3 Reverse Transcription of RNA Derived from Interferon-γ Treated Cells
For all samples, two reverse transcription reactions were performed. 2100ng of each RNA was reverse transcribed using the SuperScript III system (see section 2.6), and 1000ng of RNA was reverse transcribed using the SensiFAST system (section 2.7).

2.10.4 9p21 qPCR Assays and Analysis
Materials Required:
- SensiMix SYBR No Rox master mix, 2x (Bioline, UK)
- Primers, 10μM – see table 2.18 for sequences
• Rotor-Gene Q instrument (Qiagen, UK)
• Corbett CAS1200 liquid handling robot (Corbett, UK)
• Bench Top Ioniser (Charles Water, UK)
• 100-well Rotor-Discs (Qiagen, UK)
• Rotor-Disc adhesive seals (Qiagen, UK)
• Conducting, disposable tips (Qiagen, UK)
• Rotor-Gene Heat Sealer (Qiagen, UK)
• Strip tubes and caps (Qiagen, UK)

2.10.4.1 qPCR Reaction Setup
Each cDNA sample was analysed in triplicate, with three NTC samples per run. No RT samples were also analysed in triplicate; these were run independently of the cDNA samples with a positive (pooled cDNA) control. A master mix sufficient for all samples was prepared containing 12.5μl of SensiMix SYBR No Rox 2x mix per qPCR sample. Details regarding primer concentrations and the requirement for additional MgCl₂ are given in table 2.19. Water was added to a final volume of 22μl per sample.

All runs using HCT116 cDNA were setup robotically due to the number of samples, whereas HeLa-S3 cDNA qPCR reactions were setup manually. 22μl of master mix was added to each reaction tube, and 3μl of cDNA (10ng/μl) was added. Samples were cycled according to the conditions given in table 2.19, with a 10 minute initial denaturation at 95°C and a melt curve of 60-99°C. All runs contained all samples from a particular experiment, thus generating three data sets per cell type per gene.

2.10.4.2 qPCR Data Analysis
Comparative quantification was used for the 9p21 data. Each run was initially analysed individually (to assess the quality of the data obtained from the run), followed by combined analysis of all data.

Firstly, for each run, the melt curve was assessed to confirm that a single product had been amplified. Then, the output of the comparative quantification algorithm (discussed in sections 2.8 and 2.9) was used to obtain data that could be used for subsequent analyses. For any given triplicate, a replicate was determined to be an outlier if its take-off value differed by >0.5 cycles from the other two values or if no amplification was observed - in either case, the replicate was removed from all subsequent analyses. No
more than one replicate per triplicate was excluded. If more than one required exclusion (i.e. due to inconsistent take-off values), all 4 cDNA samples for that cell line were repeated (this action was taken due to the analysis method utilised). Normalisation to untreated cell expression levels and the reference gene was performed as follows:

1. The mean of the U-A and U-B samples for all genes was determined ($U_{\mu}$).
2. The mean of the T-A and T-B samples for all genes was calculated ($T_{\mu}$).
3. The $U_{\mu}$ value for the GOI was divided by the reference gene $U_{\mu}$ value ($U_{\text{norm}}$).
4. The $T_{\mu}$ value for the GOI was divided by the reference gene $T_{\mu}$ value ($T_{\text{norm}}$).
5. To calculate the fold change of a cell line’s expression of the gene of interest, $T_{\text{norm}}$ was divided by $U_{\text{norm}}$.

Data from cell lines with identical genotype and age were combined together and plotted as a column graph with error bars representing the standard error of the mean. Hypothesis testing was assessed by Mann-Whitney test using Stata/IC v.13.1 (StatCorp LP, USA). Where applicable, normalisation to IRF-1 was performed following 36B4 normalisation. This was carried out only when comparing genotypes.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Forward Primer Sequence (5'-3')</th>
<th>Reverse Primer Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>36B4</td>
<td>TCGACAATGGGCAGCATCTAC</td>
<td>GCCTTGACCTTTTCAGCAAG</td>
</tr>
<tr>
<td>CDKN2A (p16)</td>
<td>CATGGAGCCTTCGGCTGACT</td>
<td>CCATCATCATGACCTGGGATCG</td>
</tr>
<tr>
<td>CDKN2A (p14/ARF)</td>
<td>GGGTTTCGTGGTTACATCCC</td>
<td>CCCATCATCATGACCTGGTCTT</td>
</tr>
<tr>
<td>CDKN2B (p15)</td>
<td>GAATGCGCCAGGAGAACAAG</td>
<td>CCATCATCATGACCTGGGATCG</td>
</tr>
<tr>
<td>MTAP</td>
<td>GGGTCAATGGTTCCCTTGACCTT</td>
<td>TGTGGATGCCCAGGACAGGCTCAC</td>
</tr>
<tr>
<td>DMRTA1</td>
<td>GGGATGCAGAGCAGTTGGTT</td>
<td>TTTTGTTCCCCAATGACACCA</td>
</tr>
<tr>
<td>IRF-1</td>
<td>TGGCTGGGACATCAACAAAGG</td>
<td>TCCGCTCTGGTCTTTCAACC</td>
</tr>
<tr>
<td>ANRIL exons 1-5</td>
<td>TGCCGGAGCTGTCGACCC</td>
<td>CTTTGTATCTCCTGCTGGAATCAGAATG</td>
</tr>
<tr>
<td>ANRIL exons 4-6</td>
<td>TGTACTAACCACCTGGACTACCTGC</td>
<td>TCCACCAACCTAACAGTGATGCTTG</td>
</tr>
<tr>
<td>ANRIL exons 17-18</td>
<td>CAGAGCAATTCAGTGCAAG</td>
<td>GATTGGCAAAACAGCTG</td>
</tr>
<tr>
<td>ANRIL exons 18-19</td>
<td>GGGATGCAATGAGCTATTGAGGCC</td>
<td>GGTCCAGTCCTGGTCTGACCCAC</td>
</tr>
<tr>
<td>Total IFN-α</td>
<td>AGAATCCTCCTTTCTCTCG</td>
<td>TCTGCAAACCTCCAGGCA</td>
</tr>
<tr>
<td>IFN-β</td>
<td>TGCTCTCCTGTTGCTTCC</td>
<td>CATCTCATAGATGGTCATGCGG</td>
</tr>
</tbody>
</table>

Table 2.18: 9p21 qPCR Assay Primer Sequences
<table>
<thead>
<tr>
<th>Gene</th>
<th>Reverse Transcription System Used</th>
<th>Final Concentration Each Primer (nM)</th>
<th>Additional MgCl₂ Required (μl; 50mM)</th>
<th>Annealing Temperature and Time</th>
<th>Extension Temperature and Time</th>
<th>Number of Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>36B4</td>
<td>SSIII + SF</td>
<td>300</td>
<td>0</td>
<td>58°C for 15 seconds</td>
<td>72°C for 20 seconds</td>
<td>35</td>
</tr>
<tr>
<td>CDKN2A (p16)</td>
<td>SSIII</td>
<td>400</td>
<td>0</td>
<td>65°C for 15 seconds</td>
<td>72°C for 30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>CDKN2A (p14/ARF)</td>
<td>SSIII</td>
<td>400</td>
<td>0</td>
<td>65°C for 15 seconds</td>
<td>72°C for 30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>CDKN2B (p15)</td>
<td>SSIII</td>
<td>400</td>
<td>0</td>
<td>65°C for 15 seconds</td>
<td>72°C for 30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>MTAP</td>
<td>SSIII</td>
<td>400</td>
<td>0</td>
<td>65°C for 15 seconds</td>
<td>72°C for 30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>DMRTA1</td>
<td>SSIII</td>
<td>400</td>
<td>0</td>
<td>65°C for 15 seconds</td>
<td>72°C for 30 seconds</td>
<td>40</td>
</tr>
<tr>
<td>IRF-1</td>
<td>SF</td>
<td>400</td>
<td>0</td>
<td>64°C for 20 seconds</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>ANRIL - exons 1-5</td>
<td>SF</td>
<td>100</td>
<td>0.5</td>
<td>61°C for 45 seconds</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>ANRIL - exons 4-6</td>
<td>SF</td>
<td>100</td>
<td>0.5</td>
<td>61°C for 60 seconds</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>ANRIL - exons 17-18</td>
<td>SF</td>
<td>200</td>
<td>0.5</td>
<td>61°C for 60 seconds</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>ANRIL - exons 18-19</td>
<td>SF</td>
<td>200</td>
<td>0.5</td>
<td>61°C for 14 seconds</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Total IFN-α</td>
<td>SF</td>
<td>400</td>
<td>0</td>
<td>60°C for 45 seconds</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>IFN-β</td>
<td>SF</td>
<td>400</td>
<td>0</td>
<td>60°C for 45 seconds</td>
<td></td>
<td>40</td>
</tr>
</tbody>
</table>

Table 2.19: qPCR Reaction Mixes and Cycling Conditions for 9p21 Assay.

The reverse transcription kit used is also given; SSIII = SuperScript® III (Life Technologies, UK); SF = SensiFast cDNA Synthesis kit (Bioline, UK). Combined annealing and extension conditions indicate that a single step was utilised for these cycling steps.
In addition to the genes analysed for expression at the 9p21 locus (detailed in tables 2.18 and 2.19), optimisation of a qPCR assay for a circular splice form of *ANRIL* was attempted. This splice variant was first identified by Burd et al. (2010), and is amplified using two forward primers in exons 14 and 5 of *ANRIL*. Unfortunately, this assay could not be optimised for use, due to inconsistent amplification between runs with the same conditions and reaction mix, and the inability to generate a standard curve for confirmation of assay linearity. Modification of cycling conditions and reaction mix failed to rectify this issue.

An alternative approach that could have been taken in order to quantify circular forms of *ANRIL* include RNA-sequencing (RNA-seq). With this method, quantification of RNA transcripts is achieved by using next-generation sequencing to generate millions of short sequence reads. The number of reads mapping to a particular sequence can be used for quantification. To enrich for circular transcripts, which are expressed at low levels, it may be necessary to adopt a modified protocol to remove linear RNA prior to RNA-seq, such as that described by Jeck *et al.* (2013). However, it was beyond the scope of this work to utilise this methodology.

### 2.11 1p13 Chromatin Immunoprecipitation (ChIP)
Chromatin immunoprecipitation (ChIP) experiments were performed according to the protocol described below. The recommended protocol provided with the Imprint® chromatin immunoprecipitation kit was used, but the volume of media and PBS used at some steps was modified to ease handling of the samples. The entire protocol is described below with all changes made to the protocol given in place of those in the manufacturer’s procedure.

**Materials Required:**
- Imprint® Chromatin Immunoprecipitation kit (Sigma, UK)
- 37% formaldeyde (Sigma, UK)
- Sanyo SoniPrep 150 sonicator (Fisher, UK)
- Ice-cold PBS (Oxoid, UK)
- 1.25M glycine in water (Sigma, UK)
- 5M NaCl in water (Sigma, UK)
- 100% ethanol
- Rocking platform
- Orbital shaker
- Incubator set to 65°C
- Anti-C/EBPα antibody (cat. no. sc-9314, Santa Cruz Biotechnology, USA)
- Normal goat IgG antibody (cat. no. sc2028, Santa Cruz Biotechnology, USA)
- 1x Tris-EDTA buffer (Sigma, UK)
- Microfuge
- GenElute PCR Clean-up kit (Sigma, UK)

Procedure:
250,000 cells were plated into 2 or 3 wells of a 6-well plate in 2ml media. WT cells were seeded into 3 wells; all other lines were plated into two wells. The cells were incubated for 16 hours. Then, the media was replaced with fresh culture media and the cells incubated for a further 24 hours. The cells were washed with 1ml PBS and trypsinised with 1ml 2x trypsin. Cells were tritutrated after addition of 1ml media and all cells of the same line were pooled together in a Falcon tube. The cells were counted and 250,000 cells were transferred to a 1.5ml Eppendorf tube. For all samples, one tube containing 250,000 cells was taken, except for the WT line (2 additional cell aliquots for positive and negative controls were also taken and processed alongside the experimental samples). The volume of media in each tube was adjusted to be equal (with fresh media) prior to centrifugation for 5 minutes at 200xg. The media was removed and the cell pellet was washed with 1ml of PBS by centrifugation (5 minutes at 200xg). The PBS was removed, and the cell pellets were resuspended in 1ml fresh culture media. 30μl of 37% formaldehyde was added to each tube and the tubes inverted 5 times to mix. The samples were incubated at room temperature for 10 minutes on a rocking platform. Fixation was quenched by adding 111.1μl 1.25M glycine and inverting several times to mix. The samples were centrifuged to pellet the fixed cells (5 minutes at 200xg) and the supernatant removed. Samples were washed 3 times, by centrifugation, with 1ml ice-cold PBS and then stored at -80°C until the ChIP procedure was carried out.

Frozen cell pellets were defrosted on ice. 50μl of Nuclei Preparation buffer was added, and the tubes incubated for 10 minutes at room temperature, after which they were vortexed for 10 seconds. The isolated nuclei were pelleted by centrifugation (5
minutes at 200xg). The supernatant was removed and the pellet resuspended in 50μl Shearing buffer containing Protease Inhibitor Cocktail (PIC; 10μl of PIC per 1ml Shearing buffer). Each sample was incubated on ice for 10 minutes, with brief vortexing every 2-3 minutes. The samples were then sonicated, on ice, using a SoniPrep 150 sonicator using an amplitude of 11 microns. Each sample was sonicated with 5 pulses for 30 seconds each, with 30 seconds rest (on ice) in-between each pulse. The samples were then centrifuged at 17,000xg for 10 minutes. The supernatant was removed and transferred to new 1.5ml tubes. These were either used immediately for ChIP or stored at -80°C.

To begin the ChIP procedure, a suitable number of assay wells were washed once with 150μl of Antibody buffer. 100μl fresh Antibody buffer was then added to the wells. Antibodies were added to the wells as appropriate (2μg C/EBPα antibody, 2μg normal goat IgG antibody or 1μg RNA polymerase). All wells were sealed and incubated on an orbital shaker at 100rpm for 90 minutes at room temperature. Meanwhile, the DNA samples were prepared. Frozen samples were defrosted and stored on ice. The volume in each tube was determined and, if required, adjusted to 50μl with Shearing buffer (to account for small losses during sonication). 50μl of Dilution buffer was added, and each sample was mixed gently but thoroughly. 5μl was removed (input samples) and transferred to a new 1.5ml tube and stored on ice. The remaining 95μl (output sample) was also stored on ice until required.

After 90 minutes of antibody binding, the strip wells were washed with 150μl Antibody buffer 3 times, triturating gently in-between to ensure removal of unbound antibody. The output samples (95μl) were added to the respective wells and incubated at room temperature on an orbital shaker (at 100rpm) for 90 minutes.

Meanwhile, the cross links in the input samples were reversed by adding 20μl of water and 1μl of 5M NaCl. Samples were incubated at 99-100°C for 15 minutes and allowed to cool to room temperature. The input chromatin was purified using the GenElute PCR Clean-Up kit, following the recommended protocol. Input samples were eluted in 30μl water and stored at -20°C.

Once the 90 minute immunoprecipitation step was completed, the incubated DNA solution was removed from the wells, which were then washed 6 times (letting the wells stand for 2 minutes during each wash) with 150μl IP Wash buffer. Washes were removed by inverting the plate over an absorbent towel. After the final wash, each well was washed once with 150μl 1x Tris-EDTA buffer. 39μl of DNA Release buffer and
1µl Proteinase K were added to the wells, which were then sealed and incubated at 65°C for 15 minutes. 40µl of Reversing solution was then added, and mixed by pipetting. Wells were re-covered and incubated for 90 minutes at 65°C. Towards the end of the incubation, the output sample purification columns were prepared by applying 500µl of Column Preparation solution to them and centrifuging at 12,000xg for 1 minute.

After the 90 minute incubation, the contents of each sample well were transferred to 1.5ml tubes and 400µl DNA Binding buffer was added. Each tube was vortexed before transferring the contents to the prepared purification column. Samples were centrifuged for 1 minute at 14,000xg and the flow-through discarded. Columns were washed with 500µl Wash buffer by centrifuging at 14,000xg for 1 minute. After removing the flow-through, the columns were centrifuged again for 2 minutes at 14,000xg. The columns were transferred to new 1.5ml tubes and 30µl of Elution solution applied. Samples were centrifuged for 1 minute at 14,000xg. The eluate was re-applied to the column and re-centrifuged. The eluted output DNA was stored at -20°C prior to qPCR analysis (section 2.13).

2.12 9p21 Chromatin Immunoprecipitation (ChIP)

ChIP was performed on IFN-γ stimulated and non-stimulated cells. Given below are the steps that precede the ChIP procedure (the IFN-γ stimulation). Please refer to section 2.11 where indicated in the protocol below for the ChIP procedure.

Materials Required:

- 20µg/µl IFN-γ (PeproTech, USA) in PBS containing 0.1% BSA (Oxoid, UK and Sigma, UK, respectively)
- PBS containing 0.1% BSA (Oxoid, UK and Sigma, UK, respectively)
- Imprint® Chromatin Immunoprecipitation kit (Sigma, UK)
- 37% formaldehyde (Sigma, UK)
- Sanyo SoniPrep 150 sonicator (Fisher, UK)
- Ice-cold PBS (Oxoid, UK)
- 1.25M glycine in water (Sigma, UK)
- 5M NaCl in water (Sigma, UK)
- 100% ethanol
- Rocking platform
- Orbital shaker
- Incubator set to 65°C
- Anti-STAT1 antibody (cat. no. sc-346, Santa Cruz Biotechnology, USA)
- Normal rabbit IgG antibody (cat. no. sc-2027, Santa Cruz Biotechnology, USA)
- 1x Tris-EDTA buffer (Sigma, UK)
- Microfuge
- GenElute PCR Clean-up kit (Sigma, UK)

Procedure:
250,000 cells were seeded into the wells of 6-well plates. For the WT line, 5 wells were plated, and for all other lines 4 wells were plated. Cells were treated with IFN-γ as described (section 2.10.1). 2 wells for each line were treated with IFN-γ, with the remaining wells treated with vehicle. After 24 hours of treatment, cells were pooled together and counted. 250,000 cells of each line and treatment condition were used for the ChIP procedure. For WT cells, two additional samples were included (containing 250,000 cells) for IgG negative and RNA polymerase (positive) controls. ChIP was performed as described in section 2.11, substituting anti-STAT1 antibody for the anti-C/EBPα and normal rabbit IgG for the normal goat IgG antibody.

2.13 ChIP-qPCR Analysis of Transcription Factor Binding at the 1p13 and 9p21 Loci
Following the isolation of DNA by ChIP, targeting the transcription factors of interest, qPCR was used to quantify the level of transcription factor binding. The following protocol was used for both the 1p13 and 9p21 ChIP samples.

Materials Required:
- SensiMix SYBR No Rox master mix, 2x (Bioline, UK)
- Rotor-Gene Q cycler (Qiagen, UK)
- Rotor-Gene plasticware (4-well strips and caps) (Qiagen UK)
- Primers (table 2.20)
<table>
<thead>
<tr>
<th>Primer Name</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p13_ChIP_F</td>
<td>GTTAAGTTGCGCTTCTTTGC</td>
</tr>
<tr>
<td>1p13_ChIP_R</td>
<td>GACACCAGAACCAGACTTG</td>
</tr>
<tr>
<td>9p21_ChIP_F</td>
<td>GTTAGTTGGAACTGAAGCTG</td>
</tr>
<tr>
<td>9p21_ChIP_R</td>
<td>AAAGCGTACAAATTTAAAG</td>
</tr>
</tbody>
</table>

**Table 2.20: Primer Sequences for 1p13 and 9p21 ChIP-qPCR Assays**

Procedure:

Input and output chromatin samples were diluted 1:3 with water. Triplicate qPCR reactions were setup for each sample (as well as 3 NTCs per run), each comprising:

- 12.5μl SensiMix SYBR No Rox (2x) mix
- 400nM forward primer (10μM)
- 400nM reverse primer (10μM)
- 3μl diluted DNA
- Water to a final volume of 25μl

Samples for both 1p13 and 9p21 ChIP samples were cycled in a Rotor-Gene Q machine with an initial denaturation at 95°C for 10 minutes followed by 45 cycles of 95°C for 15 seconds (denaturation) and 58°C for 25 seconds (combined annealing and extension).

Upon completion of each run, melt curve analysis was used to confirm the amplification of a single product with no detectable primer dimerisation. Outlier replicates were removed if they showed an absence of amplification or a take-off greater than 0.5 cycles different to the remaining two replicates. NTCs were acceptable if no amplification occurred, or displayed take-off values at least 7 cycles later than the DNA-containing samples. IgG negative control assays were deemed negative if no amplification was observed, or if the take-off value was 10 cycles or more greater than the DNA samples. IgG negative control samples were determined to be positive if the take-off values were less than 10 cycles after the DNA samples (in this case, input and output samples were adjusted, as described below).

Average take-off values were used for calculation of the amount of DNA in each sample using the equation of the standard curve for the assay. These values were multiplied by the dilution values (1200 for input samples, 30 for output samples). The % input value was calculated with these values using one of two calculations (determined by the amplification of the IgG control):
No IgG negative control amplification:

\[
\% \text{ input} = \frac{\text{output}}{\text{input}} \times 100
\]

where:

\[
\text{output} = \text{calculated output DNA (ng)}
\]

\[
\text{input} = \text{calculated input DNA (ng)}
\]

IgG negative control amplification:

\[
\text{Adjusted } \% \text{ input} = \frac{\text{output}_{\text{corrected}}}{\text{input}_{\text{corrected}}} \times 100
\]

where:

\[
\text{output}_{\text{corrected}} = \text{calculated output DNA} - \text{calculated IgG DNA}
\]

\[
\text{input}_{\text{corrected}} = \text{calculated input DNA} - \text{calculated IgG DNA}
\]

Calculated % input or adjusted % input values were combined by genotype and normalised to parental genotype values. The data were analysed with the Mann-Whitney test using Stata/IC v.13.1 (StatCorp LP, USA).
Chapter 3: Parental Cell Line Characterisation and Optimisation of rAAV Genome Editing Protocols

3.1 Introduction

3.1.1 Overview of rAAV Genome Editing
Recombinant adeno-associated virus (rAAV)-mediated genome editing is a technique used to make alterations to the genome in a targeted manner. Homologous recombination of the single-stranded DNA vector into the locus of interest enables the replacement of native DNA with that of the vector, incorporating the desired modifications into the target locus alongside a selection cassette. The selection cassette provides a means by which the cell population can be enriched for cells that have integrated the virus, and typically utilises antibiotic selection. Following selection, cells targeted at the correct locus can be identified by a diagnostic PCR screen using nested PCR primers that are included within the selection cassette.

Once targeted cells have been found and expanded, the selection cassette can be removed by Cre recombination. rAAV vectors incorporate loxP sites that flank the selection cassette, permitting its removal. A single loxP site (34bp) remains after the removal of the selection cassette.

Vector design considerations are discussed in section 3.1.2, as is the need for parental cell line characterisation to both inform the vector design process, and also identify cell lines that are suitable for the procedure. The results of the parental cell line characterisation are given in section 3.2.

A detailed discussion of the technical procedures required for rAAV genome editing are given in section 3.1.3. Figure 3.1 provides an overview of the procedure, whilst figure 3.2 shows the diagnostic PCR and Cre PCR strategies used in this study.

3.1.2 rAAV Vector Design Considerations and the Importance of Parental Cell Line Characterisation
rAAV vector design begins with a bioinformatic analysis of the region of interest. This provides valuable information regarding the site to be targeted and also the surrounding region. Careful consideration of homology arm placement is required so as to ensure that the selection cassette, and the resulting cloning scar (loxP site) after Cre recombination, have minimal effects that may influence downstream experiments. To this end, wherever possible, selection cassettes are designed such that they do not
recombine into annotated coding or regulatory sequences. In this work, vectors were designed during September 2011 by examining data presented on the UCSC genome browser for multiple genomic features, including coding and regulatory regions. When designing the vectors, the location of the selection cassette avoided these regions to minimise unintended effects mediated through selection cassette/loxP site insertion. Gene annotations were assessed using GENCODE data (Harrow et al. 2012). Regulatory regions were identified using ENCODE (Dunham et al. 2012) ChIP-seq data for multiple chromatin modifications associated with regulatory regions, such as enhancers and insulators. The modifications considered included mono-, di-, and trimethylation of residue 4 (lysine) of histone 3 proteins (abbreviated H3K4me1, H3K4me2 and H3K4me3, respectively), and acetylation of histone 3 residue 27 (lysine; H3K27ac). H3K4me1 is associated with active/poised enhancers, H3K4me2 and H3K4me3 are associated with transcription factor binding sites and promoters, and H3K27ac is found in regions with active enhancers. Further, DNAse hypersensitivity marks and transcription factor binding sites (identified through ChIP-seq analyses), were also assessed and avoided during vector design.

Whilst ChIP-seq data produced by the ENCODE project has been considered in this work as an aid to rAAV vector design, it may also be used to assess chromatin marks and transcription factor binding sites bioinformatically in a number of cell types as a means to identify additional regulatory regions. This may therefore provide a basis for the identification and examination of novel, putatively regulatory variants that lie within regulatory elements identified by ENCODE data analysis.

A further consideration for the vector design is the location of the modification(s) in relation to the selection cassette. The selection cassette is in the centre of the vector DNA that is used as the template for the homologous recombination event, and is therefore incorporated into the genome with the highest rate of integration. Because homologous recombination is often incomplete, editing of a target site increases in likelihood the closer the site is to the selection cassette. Thus, the position of the modification in relation to the selection cassette should be considered, in addition to the placement of the selection cassette in the target genomic DNA. A compromise between these two factors may need to be made.

Once a vector design has been determined, sequencing of the region of interest in the cell lines to be edited is required. The resulting sequencing data are used to ensure that the vector sequence is identical to the target locus in the cell line, so that additional,
unintentional alterations are not made to the genome upon vector recombination. In this work, individual SNPs at the target loci were to be edited. Sequencing of the region allowed for the determination of SNP genotype, ensuring that the correct SNP allele was incorporated into the vector.

Antibiotic selection is frequently used with rAAV genome editing; determining which antibiotic to use is required. Parental cell line characterisation experiments utilise a dilution series of antibiotic concentrations in order to assess the sensitivity of a cell line to the antibiotic. This experiment is also able to identify the concentration of antibiotic to use during selection; an optimal amount is required. This is because insufficient antibiotic may permit growth of non-targeted cells, whilst excessive concentrations may inhibit growth of targeted cells. When characterising parental cells, the optimal concentration is the lowest concentration that leads to complete cell death.

Sensitivity testing not only aids with the optimisation of the targeting protocols, but also with the vector design, as it both provides information as to which antibiotic resistance gene to use and determines the final homology arm length. The selection cassette includes the resistance gene, a promoter (in this work, the PGK promoter was used), two nested primer sequences and two loxP sites. Thus, the homology arms can be designed to be of an optimal length for the project when taking into account the total length of the selection cassette, which varies with the antibiotic resistance gene used.

An additional vector feature that may be considered is the inclusion of a shRNA directed against the antibiotic resistance gene within the selection cassette. Located at the 5’ and 3’ ends of the vector (adjacent to both ITRs), the use of a shRNA may be beneficial in a number of targeting strategies, and are used to enrich cell populations with cells that have undergone homologous recombination, rather than off-target NHEJ of the vector. During NHEJ insertion, the entire vector sequence, including the ITRs, is incorporated into the genome. This permits the expression of the antibiotic resistance gene in cells that have not undergone homologous recombination at the target site. These cells are thus able to proliferate during the selection procedure, reducing the proportion of correctly targeted cells in the population. However, when a vector containing the shRNA integrates by NHEJ, the shRNA is expressed, thereby reducing expression of the antibiotic resistance protein; the cells are therefore susceptible to antibiotic selection. The shRNA therefore acts to enrich cell populations with cells that have undergone homologous recombination, as these cells will not integrate the shRNA and remain resistant to the antibiotic. This enrichment enables easier identification,
isolation and expansion of the desired cells by increasing the probability that any given cell that survives the selection process has undergone successful integration of the vector at the target locus. More information regarding the use of this enrichment technique for the 9p21 project can be found in chapter 5.

A final aspect of vector design that must be considered is the sequence of the diagnostic PCR primers to be included in the vector. A simple diagnostic PCR screening strategy was utilised to identify KI of the selection cassette at the locus of interest (discussed in section 3.1.3 and figure 3.2). Briefly, two nested PCR primer sequences are included in the selection cassette, one adjacent to each of the homology arms. These primer sequences are derived from the homology arm sequence; the primer sequence adjacent to the left homology arm is derived from the right homology arm, and vice versa. PCR amplification using one of the nested primers and a primer from outside of the vector sequence (external primer) generates two potential products that differ in size. In the strategy implemented in this work, the larger of the two products is amplified from non-targeted chromosomes, whilst a shorter product is made from chromosomes in which recombination of the selection cassette has occurred.

The PCR assay needs to extend into the genomic DNA (from the selection cassette and into the native, non-homology arm DNA) to enable identification of targeted cells through co-amplification of KI-specific and WT-specific sequences with the same primer pair. Design considerations for these assays include the length of the two expected PCR products (derived from non-targeted and targeted chromosomes). The length of the two products should be different enough so as to enable distinction between them, without causing preferential amplification of one product over the other due to large differences in amplicon length.

Putative primer pairs can be designed using *in silico* methods of primer design, such as Primer-BLAST (NCBI; Ye et al. 2012). Following this, assays are tested experimentally using genomic DNA from the cell line of interest to determine if the non-targeted product can be amplified with good efficiency, allowing assay optimisation. Two assays, one per homology arm, are included in the vector. Once suitable assays have been established and all of the previously-described processes have been performed, the vector design is complete.

Parental cell line characterisation is not only important for vector design but also optimisation of multiple aspects of the procedure. The first of these that may be considered is the serotype of rAAV to use. There are multiple different serotypes of the
rAAV capsid, each with different tropisms for different cell types. For example, Keiser et al. (2011) tested three serotypes with HeLa cells (serotype 1, 2 and 5), and identified serotype 2 (also termed rAAV2) from the three tested to be the most efficient at transducing HeLa cells. The entry of the virus into the cell, intracellular trafficking and nuclear import was slow for serotype 2, but it yielded the highest level of transduction compared to serotypes 1 and 5. Serotype 1 showed inefficient capsid removal, whereas the serotype 5 virus was degraded in the nucleus. This work highlights the variability between serotypes, and shows that determining which serotype is suitable for the cell line(s) of interest is an important consideration.

The most widely used serotype for rAAV genome editing is rAAV2, and is suitable for a variety of cell types (Choi et al. 2005). In order to identify which serotype is appropriate, infection of the cell line with a rAAV encoding a reporter, such as green fluorescent protein (GFP), can be used. Successful transduction of the cells can be determined by fluorescence microscopy at various time points after infection, as transduced cells will fluoresce green. By testing a range of multiplicities of infection (MOI; the number of virus particles per cell), the efficiency of transduction can be optimised.

Determination of cell line proliferative ability is also an important part of parental cell line characterisation. Long-term cell culture and clonal expansion are required for rAAV genome editing. As such, the cells used need to be able to proliferate for extended periods, without exhibiting replicative senescence or genomic instability. In order to assess this, cells can be cultured and their proliferative ability evaluated qualitatively, based upon their growth characteristics in vitro. However, a quantitative method may be preferable, and cell growth rate can be determined experimentally using several methods, including measurement of the rate of incorporation of [3H]-thymidine, flow cytometry and several colourimetric assays. Such colourimetric methods include tetrazolium reduction assays (such as the MTT assay), ATP content assays and protein-binding assays (for example, the sulphorhodamine B [SRB] assay, Skehan et al. 1990).

Another requirement of the cells used for rAAV editing is the ability of the cell to grow clonally from a single cell. This is needed for generation of isogenic cell lines. Clonal lines allow for post-editing experimental investigation of the alterations made to the genome, enabling the establishment of direct associations between modification and phenotype.
During characterisation, cells can be plated at low seeding densities, e.g. 1 cell/well of a 96-well plate. The cells can be cultured until colonies appear, and the efficiency of obtaining single colonies can be determined for each seeding density. To confirm that the cells can proliferate further, they can be expanded to a T75 culture.

The proliferative ability of cells is a key factor determining the cell lines that can be used. Primary cells would be unsuitable because of their inability to proliferate for extended durations. Cancer lines and stem cells are alternatives. However, cancer cells may not fully recapitulate the functions of their primary cell counterparts. Stem cells, such as embryonic, induced pluripotent and mesenchymal stem cells may be suitable for the technique, but these cells can be difficult to culture and differentiate. It is therefore important to characterise the cells considered for rAAV genome editing prior to use with the technique.

### 3.1.3 Technical Procedures for rAAV Genome Editing

rAAV genome editing was used in the projects discussed in sections 4 and 5. The methodology utilised was developed by Horizon Discovery Group plc (Cambridge, UK); the following is an overview of the approach established by this company for rAAV genome editing, describing virus synthesis, transduction, targeted cell selection and identification, selection cassette removal, and, ultimately, generation of isogenic cell lines. Further information can be found in chapter 2, as well as in the protocols of Khan et al. (2011). Figure 3.1 provides an overview of rAAV genome editing.

The rAAV vector plasmid (pAAV) is generated by DNA synthesis. Following purification of the plasmid after culture in *E. coli* cells, the pAAV is transfected into HEK293T cells alongside a helper plasmid. The helper plasmid encodes the rep and cap genes, permitting assembly of the virus particles.

Once produced, the virus particles can be isolated and purified using commercial kits (such as Virapur’s AAV Purification Kit) or by centrifugation in CsCl or iodixonol density gradients (Khan et al. 2011). The virus titre can be estimated using qPCR with primers targeting the selection cassette resistance gene, or by Southern/dot blot assays (Khan et al. 2011).

After purification, the cell line of interest is infected with the virus particles over a period of 2-3 days. During this time, uptake of the virus occurs, which involves clathrin-dependent endocytosis (Bartlett et al. 2000; Uhrig et al. 2012). The virus particles within the internalised endosome require the acidification of the compartment in order
to allow transduction (Bartlett et al. 2000), suggesting that low endosomal pH may help to stabilise the vector or help it to escape from the endosome. rAAV particles begin to accumulate at the nucleus within 30 minutes of endocytosis (Bartlett et al. 2000). Entry of the particles into the nucleus has been shown to require importin-β and shuttling of the rAAV through the nuclear pore complex (NPC) (Nicolson & Samulski, 2014). At this point, the rAAV is still encapsidated; it is currently unclear how uncoating occurs. Once uncoating has occurred, homologous recombination (Vasileva et al. 2006) of the vector into the target locus can occur. Whilst it is unclear why this occurs (Deyle & Russell, 2009), it is known that induction of DSBs increases recombination rates (Miller et al. 2003), implicating this mechanism in normal rAAV HR. Further, single strand nicks have also been shown to facilitate rAAV HR (Metzger et al. 2011). It is therefore possible that the rAAV molecule acts as a repair template for HR following DNA damage at the target locus.

Following infection, cells are plated (typically in 96-well plates) under selection with the antibiotic to which resistance is provided by the selection cassette; the duration of selection is dependent upon the antibiotic used.

After selection, wells containing colonies are subjected to crude DNA isolation (using, for example DirectPCR reagent from Viagen) following visual assessment of each well to determine presence/absence of colonies and colony number. DNA is isolated in this way to facilitate the screening of 100s-1000s of samples. A diagnostic PCR assay (figure 3.2A) that produces differential PCR products from targeted and non-targeted DNA samples is performed. Agarose gel electrophoresis of diagnostic PCR products is carried out to visualise the fragments; successful targeting is indicated by production of a shorter PCR product, which can only be produced when the selection cassette has integrated at the target locus.

Once populations containing targeted cells are identified, genotyping is performed to confirm the integration of the desired alteration(s). Following this, the selection cassette can be removed by Cre-mediated recombination. Cre recombinase is provided by transfection of a Cre-expressing plasmid into the cells. The Cre recombinase enzyme promotes site-specific recombination at the loxP sites; the DNA between the two loxP sites is removed and a single loxP site remains following the recombination, forming a “cloning scar”. Careful design of the vector with respect to the selection cassette integration site is required in order to prevent loxP-mediated effects.
Single cell-derived clones are then produced from Cre plasmid-transfected cells, generating isogenic populations. This is carried out by plating a single cell per well in multiple 96-well plates and culturing cells until colonies develop. At this stage, another PCR assay is performed in order to identify colonies that have undergone successful Cre recombination. The Cre PCR assays (figure 3.2B) utilise 3 PCR primers, specific for different parts of the vector/target locus. Primer “X” is located within the selection cassette, whilst primers “Y” and “Z” are sequences from either homology arm. Primer sequences are designed such that 3 possible, differently-sized products can be made that correspond to chromosomes that are parental (no loxP or selection cassette), Cre-recombined (a single, 34bp loxP site remains) and non-Cre recombined chromosomes (where the selection cassette has not been removed). Please refer to section 2.4.9 for more details.

Cre PCR assay samples can be analysed by gel electrophoresis and banding assessed in order to identify Cre-positive samples. These can be expanded from a single colony to a large population and banked. At this point, the clonal lines can be used for functional experiments, or, if required, re-targeted and the procedure repeated to yield homozygous cells.
Figure 3.1: Overview of rAAV Genome Editing Procedure. A. The parental cell line genome is homozygous at the site of interest (red bar). B. rAAV vectors are designed with two homology arms; the sequence of these is identical to the target locus, except at the site of interest (purple bar). The selection cassette and loxP sites are located in the centre of the vector. A selectable marker (e.g. antibiotic resistance) is encoded by the selection cassette. Inverted terminal repeats (ITRs), essential for viral packaging, are located at the 5’ and 3’ ends and are not integrated into the genome. C. Homologous recombination of the vector into the parental cell genome results in incorporation of the selection cassette and the alteration at the site of interest, generating heterozygous cells. D. Cre recombination, provided by transfection of a plasmid encoding Cre recombinase, removes the selection cassette and one loxP site, leaving a single loxP site behind (34bp). The cells remain heterozygous at the target site.
Figure 3.2: Diagnostic PCR and Cre PCR Screening Strategies. A. The selection cassette contains PCR primer sequences taken from adjacent genomic DNA. PCR amplification with the diagnostic primers produces two differently-sized products. If the template DNA has not been targeted (top), a long PCR product is made. If targeted (bottom), a shorter product is amplified because of the additional primer site, closer to the external (non-vector) primer. Heterozygous cells produce both products, whilst non-targeted cells produce the long product; gel electrophoresis (right) allows for negative and positive samples to be distinguished. B. The Cre PCR strategy uses 3 primers to amplify a total of 3 possible products, depending upon whether the DNA sample is derived from parental cells, targeted cells that have not undergone Cre recombination, and targeted cells that have had the selection cassette removed successfully. The 3 PCR products are all different sizes, as shown in the figure, and so gel electrophoresis (right) can discriminate between unsuccessful (“negative”) and successful (“positive”) selection cassette removal.
3.2 Results

As discussed above, parental cell line characterisation is an important aspect of rAAV genome editing. In this work, characterisation of a number of cell lines was performed. For the 1p13 locus project (chapter 4), HepG2 and THLE-3 cells were assessed for suitability. Both of these cell lines are derived from liver cells; HepG2 cells are hepatocellular carcinoma cells, whilst THLE-3 cells are SV40 T antigen-immortalised primary hepatocytes (Pfeifer et al. 1993). These two cell lines were selected for characterisation because a liver-expressed transcription factor and pathway were to be investigated in edited cells.

For the 9p21 project (chapter 5), EA.hy926 and HeLa-S3 cells were characterised. HCT116 cells were also chosen for use in this project, but characterisation was not performed due to the extensive use of these cells for rAAV genome editing by other groups (Kohli et al., 2004; Matsumoto et al. 2009; Ericson et al. 2010).

EA.hy926 cells are a hybridoma of HUVEC and A549 (lung carcinoma) cells, and were chosen so as to provide an endothelial-like cell for examination of the 9p21 SNPs. HeLa-S3 cells are a clonal derivative of the HeLa cell line. HeLa cells were utilised by Harismendy et al. (2011) to investigate putative functional SNPs at 9p21 (chapter 5), and so HeLa-S3 cells were selected in order to be able to replicate the experiments performed by this group in a genetically-homogenous cell population. HCT116 cells were chosen for use in this project so as to provide an additional cell type in which to examine the effect of the 9p21 SNPs (as the mechanisms affected by these SNPs is unknown) and as a well-characterised, high-efficiency control for the genome editing procedure.

Mesenchymal stem cells, derived from Wharton’s jelly of the umbilical cord, were also selected for use with both the 1p13 and 9p21 projects. These cells were chosen so as to provide disease-relevant cell types for examination of genotype-specific effects through differentiation of the cells after editing.

This work is, to the best of our knowledge, the first use of rAAV genome editing to investigate the functional effects of SNPs. The ability to target individual nucleotides and generate clonal cell lines permits the analysis of regulatory elements at, in these projects, single base pair resolution.
3.2.1 Parental Cell Line Characterisation: Observation of Cultured Cells and Optimisation of Culture Techniques

All cell lines were cultured and passaged using standard media and techniques to establish a stably growing population of cells. During this time, proliferation rate was assessed qualitatively in order to determine suitability of the cells for the technique.

HeLa-S3, HepG2 and EA.hy926 cells were observed during culture to be able to grow quickly and robustly for long periods. MSCs appeared to be slightly slower than the cell lines, but otherwise seemed to grow at a rate conducive with rAAV genome editing. THLE-3 cells, however, grew very slowly initially and maintained the same growth rate after several passages. Because of this, THLE-3 cells were not carried forward for the remaining parental characterisation experiments and were discounted as a potential cell line for genome editing.

During optimisation of the culture technique, it was observed that HepG2 cells do not form a monolayer of cells, but rather isolated colonies of cells. Colony development decreased cell proliferation noticeably, but failed to prevent growth completely. To circumvent this, trypsinisation and cell trituration could be used to separate the cells of the colonies into a single cell suspension. Addition of culture media and continued culture permitted subsequent proliferation of the cells. Because proliferation is not fully inhibited by colony growth, and can be remedied by colony dispersal, this feature of HepG2 cells was not used to exclude them from further characterisation experiments.

3.2.2 Parental Cell Line Characterisation: Sulphorhodamine B (SRB) Cell Proliferation Assay

Cell proliferation was determined quantitatively using the SRB assay, according to the protocol described in section 2.2.1. This assay was performed for HeLa-S3, HepG2 and EA.hy926 cells, and was not performed for HCT116 cells due to their previous use in rAAV genome editing experiments. Further, MSCs were not included in this experiment as they were unavailable at the time the experiment was performed. Results from the assay are given in figures 3.3.
Figure 3.3: Sulphorhodamine B (SRB) Cell Proliferation Assay for EA.hy926, HeLa-S3 and HepG2 cells over 6 days Post-Seeding. (A) Absorbance values were normalised to the starting number of cells (time_0), and therefore represent fold change in cell number relative to the time_0 value. Data is derived from a total of 8 wells per cell line. Measurements were taken twice for each well and the average result taken. Error bars represent standard deviations. (B) Data presented in figure 3.3A were log_{10} transformed in order to identify the linear growth phase of each cell line and for calculation of doubling time.

From the SRB assay data, the doubling time of each cell line was determined according to the calculation discussed in section 2.2.1. The doubling time of EA.hy926 cells was found to be 13.1 hours, for HepG2 cells 15.0 hours and for HeLa-S3 cells 15.2 hours. These values are calculated from the log-linear part of the curve, and therefore represent doubling times under favourable conditions, including optimal nutrient concentration, low cellular waste product concentration and suitable cell density. These conditions are
met at the start of the experiment, but as each cell line proliferates, one or more of the aforementioned factors becomes limiting to cell growth and the number of cells for all cell lines begins to plateau as proliferation is reduced, and, in the case of EA.hy926 cells, cell death starts to occur after 100-120 hours of culture. Conversely, HeLa-S3 cells are capable of proliferating to a greater cell number before proliferation reduces (indicated by greater absorbance values), compared to HepG2 and EA.hy926 cells.

From the SRB proliferation assay data, it was concluded that EA.hy926, HeLa-S3 and HepG2 cells should continue to be considered as suitable for genome editing. These lines were taken forward for further characterisation.

3.2.3 Parental Cell Line Characterisation: Single Cell Dilution

Single cell dilution experiments were performed for HeLa-S3, EA.hy926 and HepG2 cells according to the protocol given in section 2.4.7 with a variety of seeding densities, ranging from 1-5 cells/well. This experiment was only used qualitatively to determine if the cells were able to proliferate from a single cell and expand to a large population (a confluent T75 of cells). All cell lines were capable of being expanded from a single cell to a T75 culture over a period of 3-4 weeks after initial plating.

EA.hy926 and HeLa-S3 cells grew well throughout the procedure, and were capable of being expanded from a single cell. The HepG2 cells required colony dispersal to promote growth (see section 3.2 for more information). Further, it was observed that for HepG2 cells, the number of single colonies (per well) that grew from a single cell was lower than the number of single colonies (per well) derived from plating multiple cells (2-5 cells/well). Therefore, it was determined that multiple cells should be plated per well for this line in order to maximise the number of wells containing a single colony. Because all cell lines could be expanded from a single colony to a T75 culture, it was determined that the capability for long-term culture and the proliferative potential of these lines is sufficient for genome editing.

3.2.4 Parental Cell Line Characterisation: Transducibility with Serotype 2 rAAV-GFP

Transducibility with rAAV particles is an integral part of the genome editing procedure that has been chosen for this work. As such, ensuring the cell lines of interest are capable of being transduced by rAAV is essential. The most widely used serotype of recombinant AAV is serotype 2 (rAAV2), and it is this serotype that was tested for use
with the cell lines (see section 2.2.3 for protocol). The main objective of this experiment was to determine if the cells are readily transduced by rAAV2 through use of a recombinant virus encoding green fluorescent protein (GFP) which can easily be visualised by fluorescence microscopy. A second reason to perform this experiment is to identify the multiplicity of infection (MOI, the number of virus particles per cell) required for complete transduction of a cell population.

For this experiment, HeLa-S3, HepG2 and mesenchymal stem cells were tested. Prior to performing this experiment, EA.hy926 were deemed unsuitable for genome editing due to genomic instability at the 9p21 locus (section 3.2.7).

For all cell lines tested, transduction with rAAV2 was successful. The efficiency of transduction approached 100% at an MOI of 100,000 after 2 days of infection for all cell lines. 25,000 cells were seeded in a 24-well plate and cultured in 500μl of media. Figure 3.4 shows fluorescence microscopy images of HeLa-S3 (figure 3.4A), HepG2 (figure 3.4B) and MSCs (figure 3.4C), all exhibiting high transduction efficiency.

![Figure 3.4: Fluorescence Microscopy of rAAV2-GFP Transduced Cells.](image)

Transduction was most efficient, approaching 100%, with an MOI of 100,000 for all cell lines. (A) HeLa-S3 cells; (B) HepG2 cells; (C) MSCs. Note that due to the large, flat morphology of MSCs, only a small number of cells could be visualised in any one field of view. The above is representative of all cells in the well. Scale shown is 400μm.

From these experiments, it was concluded that rAAV2 is suitable for all cell lines tested. It should be noted that whilst a high proportion of GFP-expressing cells have been observed, this is not indicative of the efficiency that can be expected to be observed with the rAAV vectors, as homologous recombination occurs at a low frequency.
A high virus titre is required for infection of cells during rAAV genome editing. Titres were measured by qPCR after production and 150μl used per round of infection (as recommended by Horizon Discovery Group plc). A specific MOI was not utilised in this work due to inconsistent titre results obtained from qPCR analysis, despite acceptable standard curves for all runs. Rago et al. 2007 state that inhibitors can interfere with PCR-based rAAV titre quantification assays and recommends the use of a fixed volume of virus stock. In this work, qPCR-based titre quantification was utilised to confirm successful virus production and not for MOI calculation. Thus, the MOI’s reported here are only used as indicators of rAAV2 suitability for the cell lines characterised.

3.2.5 Parental Cell Line Characterisation: G418 Death Curve
Selection of the rAAV infected cells with the correct concentration of antibiotic is crucial in successfully screening targeted cells. Use of low antibiotic concentrations may allow non-targeted cells to proliferate, thereby reducing the likelihood of successful diagnostic PCR screening (due to significantly more non-targeted DNA than recombinant DNA). Conversely, use of excessive concentrations of antibiotic may promote death of not only non-transduced cells but also targeted cells. It is for these reasons that the optimal concentration of antibiotic is determined experimentally. The optimised concentration is the lowest concentration at which all non-targeted cells are killed; it is therefore expected that the resistance gene in the selection cassette will be able to provide enough resistance to the antibiotic so as to allow proliferation of transduced cells.

The protocol used for this experiment is given in section 2.2.2. HeLa-S3, HepG2 and mesenchymal stem cells were all tested. As shown in figure 3.5, the optimal concentrations of G418 have been identified as 1.1mg/ml for HeLa-S3 cells and 0.8mg/ml for HepG2 cells. 0.3mg/ml for MSCs was determined to be the ideal concentration to use (data not shown). These data indicate sensitivity of the cells to G418, and have determined which concentration of G418 should be used during the selection procedure.
Figure 3.5: G418 Death Curve Phase Contrast Microscopy. Images for optimal G418 concentrations for HeLa-S3 (A; 1.1mg/ml) and HepG2 (C; 0.8mg/ml), showing complete cell death after 2 weeks. However, the next lowest concentration of G418 tested for HeLa-S3 (B; 0.9mg/ml) and HepG2 (D; 0.7mg/ml) show some cells survive the treatment (indicated by white arrows).

3.2.6 Parental Cell Line Characterisation: Genotyping

Prior to vector synthesis and virus production, it is important to be aware of the genotype of the cell lines of interest so that the correct vector(s) are designed and used. Genotyping of the parental cell lines was performed using Sanger sequencing of PCR amplification products (see section 2.2.5 for full details).

In addition, genotyping by microarray (Axiom™ Genome-Wide UKB WCSG Genotyping Array at Affymetrix, section 2.2.2.6) was also performed to independently validate the sequencing data of the parental cells used for the targeting experiments (all genotypes assigned by sequencing were confirmed by the array), as well as the genotype of some of the genome edited lines. Genotype data for the cell lines of interest are given in table 3.1. Sequence traces showing the genotype of the locus of interest of the principle cell lines used for the two projects (HCT116, HeLa-S3 and HepG2) are shown in figure 3.6.
<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Locus</th>
<th>SNP of Interest</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA.hy926</td>
<td>9p21</td>
<td>rs10811656</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>GG</td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>9p21</td>
<td>rs10811656</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>AA</td>
</tr>
<tr>
<td>HCT116</td>
<td>9p21</td>
<td>rs10811656</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>AA</td>
</tr>
<tr>
<td>HepG2</td>
<td>1p13</td>
<td>rs12740374</td>
<td>GG</td>
</tr>
<tr>
<td>MSC - donor 2</td>
<td>1p13</td>
<td>rs12740374</td>
<td>TG</td>
</tr>
<tr>
<td>MSC - donor 18</td>
<td>9p21</td>
<td>rs10811656</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td>1p13</td>
<td>rs12740374</td>
<td>GG</td>
</tr>
<tr>
<td>MSC - donor 19</td>
<td>9p21</td>
<td>rs10811656</td>
<td>CT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>AG</td>
</tr>
<tr>
<td></td>
<td>1p13</td>
<td>rs12740374</td>
<td>GG</td>
</tr>
<tr>
<td>MSC - donor 21</td>
<td>9p21</td>
<td>rs10811656</td>
<td>TT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>GG</td>
</tr>
<tr>
<td></td>
<td>1p13</td>
<td>rs12740374</td>
<td>TG</td>
</tr>
<tr>
<td>MSC - donor 25</td>
<td>9p21</td>
<td>rs10811656</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td>1p13</td>
<td>rs12740374</td>
<td>GG</td>
</tr>
<tr>
<td>MSC - donor 27</td>
<td>9p21</td>
<td>rs10811656</td>
<td>CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rs10757278</td>
<td>AA</td>
</tr>
<tr>
<td></td>
<td>1p13</td>
<td>rs12740374</td>
<td>TT</td>
</tr>
</tbody>
</table>

Table 3.1: SNP Genotype of Cell Lines and Primary MSCs Extracted from Multiple Umbilical Cord Donors, Determined by Sanger Sequencing
Figure 3.6: Sequence Trace Data for HCT116, HeLa-S3 and HepG2 DNA for the Loci of Interest. HCT116 and HeLa-S3 are both homozygous for the C allele of rs10811656 (*), and homozygous for rs10757278 (*) allele A. HeLa-S3 cells are heterozygous for rs1333047 (*), whereas HCT116 cells are homozygous A. HepG2 cells are homozygous for the G allele of rs12740374 (*).

It is interesting to note that despite HeLa-S3 cells being homozygous for the C and A alleles of rs10811656 and rs10757278 respectively, they are heterozygous for the SNP rs1333047, which is located 27bp away from rs10757278. This SNP is within the same CEU linkage block as rs10811656 and rs10757278. Heterozygosity at this locus highlights the importance of considering ethnicity of cell lines. HeLa-S3 cells are derived from an American female of African ancestry. Comparing linkage of the CEU and YRI populations using the online tool SNAP (hosted by the Broad Institute, Johnson et al. 2008) reveals that there is a distinct difference in linkage disequilibrium between the two populations.

Using rs10757278 as the tag SNP, rs1333047 has an $r^2$ value of 0.967 in the CEU population, but linkage between these two SNPs in the YRI population is below an $r^2$ value of 0.1 (determined using 1000Genomes data on SNAP; Broad Institute, Johnson et al. 2008), but with a D’ score of 1.000. This indicates that the population frequency of these SNPs is different within the YRI population, affecting the $r^2$ value.

Successful homologous recombination of targeting vectors into the HeLa-S3 line could potentially alter the genotype of rs1333047, depending upon which chromosome the vector recombines into. Therefore, upon generation of heterozygous cells at rs10811656/rs10757278, the genotype of rs1333047 may remain heterozygous or
become homozygous. Sequencing of the region in the cells produced by genome editing is important to establish the genotype of rs1333047, as any differences observed between parental cells and edited cells could be attributed to the alteration in rs1333047 genotype, and not the alteration of rs10811656 and rs10757278 alleles.

3.2.7 Diagnostic PCR Assay Testing and Optimisation & Identification of 9p21 Genomic Instability in EA.hy926 Cells

Experimental determination of suitability for each diagnostic PCR assay is required to establish whether a putative assay can be optimised for use. This step must be performed prior to the synthesis of the vector plasmid so that the primer sequences can be incorporated into the selection cassette sequence.

The protocol for these experiments is discussed in section 2.2.4. Multiple assays were tested for both homology arms for each vector (the 1p13 vector and the 1st and 2nd generation 9p21 vectors); one assay per homology arm that gave a distinct band of the correct size was included in the vector sequence. Primer pair sequences and cycling conditions for optimised assays for the 1p13 and second 9p21 (containing the shRNA and altered homology arm location) vectors are given in section 2.4.5. Representative gel images showing the use of the optimised assays for diagnostic PCR screening are given in figures 4.2 (1p13), 5.2 (9p21 HeLa-S3), 5.4 (9p21 HCT116 single KI) and 5.6 (9p21 HCT116 DKI).

During the optimisation of the diagnostic assay for the first 9p21 vector, EA.hy926 DNA isolated from cells that had been passaged during the course of normal growth over several weeks was used as a template. Surprisingly, no 9p21 PCR product from these DNA samples could be obtained, even with additional optimisation. However, a product could be obtained from HepG2 and HeLa-S3 DNA samples, indicating that this was not due to an assay issue. Further, the EA.hy926 DNA samples showed successful amplification of a chromosome 1 PCR product (as did the included controls), so the DNA samples were suitable for PCR. Figure 3.7 shows the agarose gel image from this experiment. The protocol utilised for this experiment is given in section 2.2.4.5.

These findings indicate that at higher passage numbers, EA.hy926 cells appear to show genomic instability at the 9p21 locus. This conclusion was reached because lower-passage EA.hy926 cells were used for DNA isolation for genotyping; EA.hy926 cells were successfully genotyped by sequencing the region. Because of this potential for loss/alteration of 9p21, EA.hy926 cells were not taken forward for any further work.
**Figure 3.7: PCR Analysis Indicates Aged EA.hy926 Cells Display Genomic Instability at 9p21.** 3 EA.hy926 DNA samples isolated after long-term culture (E1-3), as well as HepG2 (HP) and HeLa-S3 (HL) samples were analysed for the presence of 9p21 and 1p32 using PCR. Target products are as indicated (A) and (B) for 9p21 and 1p13 products, respectively. As can be seen, all 3 EA.hy926 samples tested show reduced 9p21 target band amount and additional, non-specific products compared to HepG2 and HeLa-S3 DNA samples. This is not an assay issue (HepG2 and HeLa-S3 positive control samples are as expected) and cannot be due to poor EA.hy926 DNA sample quality (the 1p32 control assay worked with similar efficiency for all samples). NTC = no template control. Ladder used is Hyperladder 1kb (Bioline, UK).

**3.3 Discussion**

This work has enabled the development of rAAV genome editing as a tool for functional genetic studies. The preceding data indicates that determination of cell line suitability for the procedure is important prior to commencement of vector design and synthesis, as well as ensuring cell lines selected are suitable for the technique. This is exemplified by THLE-3 (slow growth) and EA.hy926 (likely genomic instability at the 9p21 locus of interest) cell lines. Additional characterisation is also beneficial with respect to cell-line specific requirements for optimal targeting.

Whilst rAAV genome editing has been used successfully for a wide variety of purposes (see 1.4.3.3 and section 6.2.1), further uses of the technique may be developed to help investigate a particular research question. Key developments in this field in the future would be the removal of the need to use loxP sites and Cre recombination. Currently, vector designs must ensure that the cloning scar that remains after targeting
(a single loxP site) is not left in a location where it could have a functional effect (e.g. within exons or regulatory elements). However, whilst every care is taken to ensure this, unintentional/unavoidable alteration of functional elements can occur. Thus, for some projects, rAAV genome editing is unsuitable because of this factor. The development of vectors that do not leave extraneous DNA behind after removal of the selection cassette would be advantageous for both easier vector design and greater applicability for investigation of a wider variety of genomic loci.

One modification that may enable such a development in this technology is the PiggyBac transposase system. This system, first discovered by Fraser et al. (1983) in the moth Trichoplusia ni, utilises two inverted terminal repeats (ITRs) that flank the DNA to be transposed. A transposase enzyme moves the DNA and ITRs to a new location; the site of integration typically requires a TTAA consensus sequence, but low frequencies of integration have been observed at other sequences (Li et al. 2013). Most importantly for genome editing applications is the property of the transposase to remove ITRs and intervening DNA without altering the genomic DNA – no cloning scar remains. Further, removal of the transposon to a subsequent location after removal by the transposase can be prevented by using an “excision only” PiggyBac transposase (available from System Biosciences Inc., USA).

This system has been successfully used by Cunningham et al. (2015) in rAAV vectors for random integration of their genes of interest (eGFP, mOtc and mASS). This group showed that minimal and full length ITRs both gave similar integration efficiencies. The full length ITRs are 313bp (5’) and 230bp (3’), whereas the minimal ITRs are 67bp and 40bp long respectively. This is important, as long ITRs would limit the length of the homology arms that could be used in the rAAV vector. By using the minimal ITRs, there would be a total loss of 39bp from the vector’s homology arms, which is unlikely to reduce efficiency by any appreciable amount.

However, full-length ITRs could be used if a smaller antibiotic resistance gene is used. One such antibiotic resistance gene confers resistance to Zeocin®; this gene is 374bp in length (Pfeifer et al. 1997). G418R, as used in the vectors discussed in this work, is 794bp long. By replacing G418R with ZeocinR, one would save 424nt in the ssDNA vector. The combined length of both PiggyBac ITRs is 543nt; subtraction of this from the ZeoR and G418R size difference and accounting for the two loxP sites (34nt each) means that using the PiggyBac ITRs and ZeocinR would require a loss of 51nt from the homology arms compared to a G418R and loxP vector. By combining the
minimal PiggyBac ITRs with the Zeocin® resistance gene, one would gain an extra 385bp, thereby increasing the length of the homology arms when using the same promoter as that found within the G418 selection cassette (the PGK promoter).

An alternative use for the PiggyBac system could be the deletion of specific parts of the genome by placing the minimal ITRs at specific locations within the homology arms. Provision of the excision-only PiggyBac transposase plasmid will remove the selection cassette and also the genomic DNA between the ITRs. This could be used to create gene knock-outs or remove regulatory elements in order to examine their function.

In summary, future work with rAAV genome editing will focus on developing the technology further in order to increase the number of ways in which it may be used. This technology is anticipated to be used extensively for functional analysis of disease-associated loci identified by GWAS. For all of these projects, cell line characterisation is essential to ensure that suitable cell lines are selected for editing. The experiments discussed here provide a comprehensive assessment of multiple characteristics of each cell line. By performing these analyses, two cell lines were identified as unsuitable for rAAV editing (for different reasons), highlighting the importance of parental cell line characterisation.
Chapter 4: Functional Analysis of the 1p13 Locus Using rAAV-Mediated Genome Editing

4.1 Introduction

4.1.1 The Role of the 1p13 Locus in Coronary Artery Disease

The 1p13.3 locus has been associated with CAD and LDL cholesterol levels by multiple groups using GWAS (tables 4.1 and 4.2). This, in conjunction with the fact that LDL cholesterol is a known risk factor for CAD, suggests that the 1p13 locus may contribute to CAD through alteration of LDL metabolism. As demonstrated in table 4.2, this association is replicated by multiple groups across a number of populations. Further, an association with rs599839 and abdominal aortic aneurysm (Jones et al. 2013) has also been identified.

A total of 6 lead SNPs have been identified by GWAS as associated with CAD or LDL cholesterol phenotypes. Table 4.1 gives these SNPs, along with additional data, including SNP alleles, risk-associated alleles, minor allele frequency and linkage disequilibrium data ($r^2$) for each SNP with rs12740374. rs12740374 was chosen as the reference for the $r^2$ values as this is the first SNP given in the table (which are ordered according to rs accession number).

<table>
<thead>
<tr>
<th>SNP</th>
<th>SNP Alleles</th>
<th>Risk Allele</th>
<th>Minor Allele Frequency (Allele)</th>
<th>CEU Linkage Disequilibrium with rs12740374 ($r^2/D'$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12740374</td>
<td>G/T</td>
<td>G</td>
<td>0.20 (T)</td>
<td>N/A</td>
</tr>
<tr>
<td>rs599839</td>
<td>G/A</td>
<td>A</td>
<td>0.36 (G)</td>
<td>0.895/1.000</td>
</tr>
<tr>
<td>rs602633</td>
<td>T/G</td>
<td>T</td>
<td>0.35 (T)</td>
<td>0.895/1.000</td>
</tr>
<tr>
<td>rs629301</td>
<td>G/T</td>
<td>T</td>
<td>0.24 (G)</td>
<td>1.000/1.000</td>
</tr>
<tr>
<td>rs646776</td>
<td>C/T</td>
<td>T</td>
<td>0.24 (C)</td>
<td>1.000/1.000</td>
</tr>
<tr>
<td>rs660240</td>
<td>T/C</td>
<td>C</td>
<td>0.23 (T)</td>
<td>0.927/1.000</td>
</tr>
</tbody>
</table>

Table 4.1: 1p13 SNPs Associated with CAD and LDL Cholesterol Levels. A search of the NIH Catalog of Published Genome-Wide Association Studies (Welter et al. 2014; Hindorff et al. 2015) using search terms “coronary artery disease”, “coronary heart disease”, “LDL cholesterol”, “myocardial infarction” and “myocardial infarction (early onset)” found several studies that report an association between 1p13 SNPs and these phenotypes. Additional information regarding SNP alleles, risk-associated alleles, minor allele frequencies (from Ensembl), and linkage disequilibrium with rs12740374 is given for the CEU HapMap population (obtained from SNAP data, Johnson et al. 2008).
### Table 4.2: Summary of Published 1p13 CAD/LDL GWAS Associations

Data were obtained from the NIH GWAS Catalog as described for table 4.1 and given as provided by this database. *p*-values refer to the association identified for the given SNP for the phenotype of interest in each study given. “CAD” = coronary artery disease, “CHD” = coronary heart disease, “LDL-C” = LDL cholesterol, “LAS” = large artery stroke, “IS” = ischemic stroke, “MI-Early” = premature MI.

<table>
<thead>
<tr>
<th>SNP</th>
<th>GWAS Phenotype</th>
<th>p-Value</th>
<th>Publication</th>
<th>Population(s)</th>
<th>Initial</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Number of Cases</td>
<td>Number of Controls</td>
<td>Number of Cases</td>
<td>Number of Controls</td>
</tr>
<tr>
<td>rs12740374</td>
<td>LDL-C</td>
<td>2x10^-42</td>
<td>Kathiresan et al. (2009)</td>
<td>European (meta-analysis)</td>
<td>19,840</td>
<td>up to 20,623</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>9x10^-29</td>
<td>Lettre et al. (2011)</td>
<td>African-American + African-Carribean</td>
<td>260</td>
<td>5053</td>
</tr>
<tr>
<td>rs599839</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAD or LAS</td>
<td>8x10^-17</td>
<td>Dichgans et al. (2014)</td>
<td>European + South Asian (meta-analysis)</td>
<td>22,233</td>
<td>75,921</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CAD, 2167 LAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAD or IS</td>
<td>1x10^-8</td>
<td>Dichgans et al. (2014)</td>
<td>European + South Asian (meta-analysis)</td>
<td>22,233</td>
<td>88,766</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CAD, 12,389 IS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHD</td>
<td>4x10^-9</td>
<td>Samani et al. (2007)</td>
<td>European</td>
<td>1926</td>
<td>2938</td>
</tr>
<tr>
<td></td>
<td>CHD</td>
<td>3x10^-10</td>
<td>Schunkert et al. (2011)</td>
<td>European</td>
<td>22,233</td>
<td>64,762</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>6x10^-33</td>
<td>Willer et al. (2008)</td>
<td>European</td>
<td>8589</td>
<td>12981</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>1x10^-15</td>
<td>Sandhu et al. (2008)</td>
<td>European</td>
<td>22,233</td>
<td>64,762</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>1x10^-7</td>
<td>Wallace et al. (2008)</td>
<td>European</td>
<td>1955 (hypertensive)</td>
<td>3494 (individuals + twins)</td>
</tr>
<tr>
<td>rs602633</td>
<td>CAD</td>
<td>1x10^-8</td>
<td>Dichgans et al. (2014)</td>
<td>European + South Asian (meta-analysis)</td>
<td>33,398</td>
<td>75,726</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>1x10^-10</td>
<td>Teslovich et al. (2010)</td>
<td>European</td>
<td>95,454</td>
<td>-</td>
</tr>
<tr>
<td>rs646776</td>
<td>CAD</td>
<td>8x10^-11</td>
<td>Reilly et al. (2011)</td>
<td>European</td>
<td>1808</td>
<td>915</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>2x10^-12</td>
<td>Sabatti et al. (2009)</td>
<td>Finnish Founder</td>
<td>4763</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>3x10^-29</td>
<td>Kathiresan et al. (2008)</td>
<td>European</td>
<td>2758</td>
<td>18,544</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>8x10^-23</td>
<td>Aulchenko et al. (2009)</td>
<td>European + Orkney Islanders</td>
<td>17,083 European, 714 Orkney Islanders</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CHD</td>
<td>6x10^-10</td>
<td>Coronary Artery Disease (CAD) Genetics Consortium, (2011)</td>
<td>European + South Asian</td>
<td>8424 European, 6996 South Asian</td>
<td>7268 European, 7794 South Asian</td>
</tr>
<tr>
<td>rs660240</td>
<td>LDL-C</td>
<td>1x10^-26</td>
<td>Waterworth et al. (2010)</td>
<td>European + South Asian</td>
<td>17,723 European</td>
<td>37,774 European, 9665 South Asian</td>
</tr>
</tbody>
</table>

109
Figure 4.1: The Genes and SNPs of Interest in the 1p13 Region. Genes of interest within the 1p13 region are shown, as is the lead variant identified by GWAS (rs599839, highlighted green). Also shown is the SNP examined in this work (rs12740374, highlighted red) and SNPs with an $r^2$ of 0.8 or greater with rs599839, identified using SNAP (hosted by the Broad Institute, USA; Johnson et al. 2008). A custom track was generated containing the high LD SNPs.
The 1p13.3 locus contains three genes, *SORT1*, *CELSR2*, and *PSRC1*. Of these genes, *SORT1* is the best characterised, and encodes the protein sortilin (Petersen et al. 1999). The analysis of the amino acid sequence of sortilin by Petersen et al. (1999) revealed the presence of protein motifs that resemble those found in known lysosomal sorting receptors, suggesting that sortilin may also be a sorting receptor. Further work by this group showed that sortilin localises to the endoplasmic reticulum and Golgi complex using expression and immunocytochemistry experiments. Petersen et al. (1999) concluded that sortilin is a sorting receptor for endosomal trafficking. Sortilin is capable of binding and sorting multiple ligands, and localises with the trans-Golgi network and the cell membrane.

The *CELSR2* gene encodes a protein that is homologous to the *Drosophila melanogaster flamingo* sub-family of proteins. The CELSR (Cadherin, EGF-Like Seven-pass-Receptor) proteins are a group of structurally-atypical cadherins, which are calcium-dependent transmembrane proteins that provide adhesion between cells (Halbleib & Nelson, 2006). Most studies investigating CELSR2 have focussed on developmental roles of the protein, which functions in both neurological (Qu et al. 2014) and pancreatic (Cortijo et al. 2012) development. CELSR2 has also been shown to be involved in cell adhesion between rat testis Sertoli cells (Beall et al. 2005), and down-regulated in some breast cancers (Huang et al. 2005).

*PSRC1* encodes the proline/serine-rich coiled coil 1 protein (Lo & Wang, 2002). This protein is down-regulated by the cell cycle regulator p53 by direct binding of p53 to the *PSRC1* promoter in response to DNA damage (Hsieh et al. 2008). PSRC1 has been shown to associate with tubulin (Hsieh et al. 2008) and also the mitotic spindle during cell division (Jang et al. 2008). Here, PSRC1 has been reported to bind multiple proteins that are involved in mitosis or cell cycle regulation whilst bound to the spindle. These include KIF2A, a microtubule depolymerase; this interaction promotes spindle turnover and produces tension along the microtubules (Jang et al. 2008). PSRC1 also interacts with MCAK, which controls chromosome congression during metaphase (Jang & Fang, 2011). PSRC1 also interacts with End-Binding (EB) proteins 1 and 3 on the microtubules. The interaction with EB1 negatively regulates the growth of the plus-end of the microtubule (the end growing towards the periphery of the cell) (Zhang et al. 2013). EB3 and PSRC1 interact in a tubulin-dependent manner, which activates the β-catenin pathway and increases cyclin D1, promoting proliferation.
4.1.2 The Functional Role of \textit{SORT1} in Coronary Artery Disease

Since the identification of the \textit{SORT1} gene and sortilin protein, functional investigation of sortilin has revealed a diverse range of functions for this protein, including in metabolism, neuronal function (reviewed by Nykjaer & Willnow, 2012), and cancer. Sortilin has been shown to promote breast cancer cell migration, adhesion and invasion (Roselli \textit{et al.} 2015), whereas in prostate cancer, sortilin negatively regulates progranulin, reducing motility, invasion and proliferation (Tanimoto \textit{et al.} 2015). Sortilin has also been implicated in colorectal cancer, where its over-expression is likely promoting uptake of neurotrophic factor, a protein that promotes cell survival and growth of neurons (Akil \textit{et al.} 2011).

Sortilin has been shown by multiple groups to be involved in lipid metabolism pathways, indicating that this may be the causative gene at the 1p13 locus. One of the earliest lipid-related studies involving sortilin reported that it binds lipoprotein lipase (LPL) at the cell surface and facilitates its internalisation and subsequent catabolism in the lysosome (Nielson \textit{et al.} 1999). LPL hydrolyses triglycerides, and, depending upon its source, can be pro-atherogenic (macrophage-derived) or anti-atherogenic (derived from adipocytes and muscle), reviewed by Pentikäinen \textit{et al.} (2002).

Nilsson \textit{et al.} (2008) have demonstrated that sortilin can bind and mediate endocytosis of ApoA-V-containing complexes, including chylomicrons, when over-expressed in HEK293 cells. Fluorescently-labelled ApoA-V was used to show that sortilin mediates internalisation of ApoA-V within minutes of binding to sortilin at the cell surface. Co-localisation with fluorescent markers for lysosomes, late endosomes and the trans-Golgi network revealed the pathway through which ApoA-V is trafficked; it can be degraded in the lysosome or recycled in the Golgi complex. This function of sortilin is of interest because ApoA-V is a component of chylomicrons, VLDL and HDL, albeit at lower frequency than the other protein components of these particles (O’Brien \textit{et al.} 2005). Further, ApoA-V has been shown to reduce plasma triglyceride levels in mice (van Dijk \textit{et al.} 2004). As hypertriglyceridemia is a risk factor for CAD (Cullen, 2000), dysregulation of sortilin may contribute to altered ApoA-V metabolism and, in turn, increased CAD risk.

Sortilin has also been found to be involved in LDL uptake and clearance. Linsel-Nitschke \textit{et al.} (2010) over-expressed a human \textit{SORT1} cDNA in HEK293 cells by transfection and observed a 23% increase in uptake of radiolabelled LDL.
In vivo, liver-specific sortilin over-expression through SORT1-AAV8 injection in mice by Musunuru et al. (2010) resulted in a 73% decrease in LDL-cholesterol after 2 weeks, and after 6 weeks, over-expressing mice showed a 46% reduction in total cholesterol, a 73% reduction in VLDL and an 88% decrease in medium-small LDL particles. In addition, investigation of VLDL and triglyceride secretion by mouse liver revealed an approx. 57% decrease in the secretion of both VLDL and triglycerides.

Tveten et al. (2012) also over-expressed SORT1 in HeLa-TREX cells and observed increased binding of LDL to the cell surface (25-fold), and a comparatively small increase in uptake of 20%, whilst LDLR over-expression increased uptake by 156%. SORT1 siRNA knock-down in HeLa-TREX cells resulted in a 36% reduction in LDL uptake for one of the siRNAs tested; other siRNAs tested resulted in small effects. siRNA targeting LDLR resulted in a 71% decrease in LDL uptake. These data indicate that sortilin promotes cell surface binding and uptake of LDL. However, given that knock-down of SORT1 leads to only a small decrease in uptake, it is likely that sortilin mediates only a small proportion of total cellular LDL uptake.

Over-expression experiments in mice and cultured cell lines by Strong et al. (2012) revealed that sortilin is involved not only in LDL uptake, but also in catabolism of LDL. Knock-out experiments by this group in mice caused a decrease in LDL catabolism by 40% in Sort1 null mice, and a 50% reduction in Sort1+/−; Ldlr−/− mice. These data, and those obtained by Linsel-Nitschke et al. (2010) and Musunuru et al. (2010), suggest a role for sortilin in the uptake and lysosomal catabolism of LDL. Kjolby et al. (2010) present data that contradicts this conclusion. In a Sort1 null mouse, this group observed no difference in LDL uptake.

Sortilin has also been implicated in mediating uptake of LDL in macrophages. Patel et al. (2015) created a Sort1 null mouse on the Apobec1−/+; hAPOB background. After being fed for 18 weeks on a Western diet, the lipid profiles of these mice were compared with those of Sort1-expressing mice. There was no difference in total or LDL cholesterol levels in these two groups of mice. However, the Sort1−/− mice displayed a reduction in atherosclerotic lesion area (68-87%, depending upon the region of the aorta measured). Therefore, in these mice, atherosclerosis risk was reduced by Sort1 knock-out, despite no change in lipid profile. Bone marrow transfer experiments into irradiated Ldlr−/− mice showed that transfer of Sort1−/− bone marrow reduced atherosclerotic lesion area, indicating that Sort1 deficiency in macrophages may be mediating this protective effect. Indeed, culture of macrophages derived from both Sort1−/− and Sort1+/+ mice
found that LDL uptake and foam cell formation was reduced in the macrophages isolated from Sort1−/− mice when treated with LDL for 5 hours. LDL uptake was reduced by 39% in Sort1+/−;Ldlr−/− macrophages and by a similar amount in Sort1−/− macrophages, showing that this effect is due to reduced sortilin. These findings suggest that sortilin-mediated LDL uptake in macrophages promotes atherosclerosis through increased foam cell formation.

There is compelling evidence to suggest that sortilin is involved in LDL uptake and catabolism, but these findings have not been replicated by Kjolby et al. (2010) in mouse liver. The findings of Patel et al. (2015) differ from the observations of Mortensen et al. (2014), who reported no effect of Sort1 deficiency on murine macrophage LDL uptake, despite a reduction in atherosclerosis. This group provides evidence for a role of sortilin in promoting secretion of pro-inflammatory cytokines (such as IL-6 and IFN-γ). In this model, loss of sortilin reduces secretion of these cytokines, thereby decreasing the inflammatory response. These opposing findings mean that sortilin’s role in LDL metabolism and direction of effect with regards to atherosclerosis remains unclear. Disparities between different publications may arise from the experimental approaches taken, or may reflect a great complexity in both the function and regulation of sortilin.

Another process that has been associated with sortilin is secretion of VLDL and ApoB. Musunuru et al. (2010) reported a decrease in VLDL secretion in sortilin over-expressing mice. This finding was replicated by Ai et al. (2012); over-expression of sortilin by AAV resulted in a 34% decrease in ApoB secretion after 2.5 hours. Strong et al. (2012) also over-expressed Sort1 by AAV in both wild-type and Ldlr−/− mice and observed decreased ApoB secretion. A decrease of 30% was observed in wild-type mice, whilst this was increased to 50% in Ldlr null mice after 3 weeks. Strong et al. (2012) determined, through over-expression experiments with SORT1 mutants containing loss-of-function mutations in key motifs in the protein, that sortilin mediates lysosomal trafficking of ApoB. This finding was also replicated in primary mouse hepatocytes and HepG2 cells by Bi et al. (2013) using recombinant adenovirus transduction.

A TALEN-mediated knock-out model of SORT1 in human embryonic stem (HUES) cells also replicated the finding that sortilin reduces ApoB secretion (Ding et al. 2013). In this model, the knock-out of SORT1 resulted in increased ApoB:albumin ratio in hepatocyte-like cells (HLCs) differentiated from HUES cells. The magnitude of the effect varied between the HUES lines; HUES-1 HLCs displayed a 117% increase in
ApoB:albumin ratio, whilst HUES-9 HLCs showed a 65% increase in ApoB:albumin. The effect was abolished by transduction with lentivirus expressing SORT1, confirming the observations were caused by SORT1 knock-out.

Knock-out experiments in mice display a different phenotype from that observed in SORT1-deficient HUES HLCs. Strong et al. (2012), as previously discussed, observed decreased ApoB secretion in over-expressing mice. Sort1−/− mice also displayed decreased ApoB secretion (60%) in a Apobec1−/−;hAPOB background. This finding was also reported by Kjolby et al. (2010) with a different mouse model and diet.

The function of sortilin in ApoB/VLDL secretion has not been identified conclusively, with conflicting evidence reported by multiple groups using alternative experimental approaches. This is also the case for sortilin’s role in LDL metabolism. Because of these dissimilarities, the exact functions of sortilin remain uncertain. From these publications, it is strongly indicated that sortilin is involved in lipid metabolism and is therefore a likely candidate gene for mediating genetic risk of CAD at the 1p13 locus.

Not only has sortilin been implicated in direct lipid metabolism of either LDL or VLDL, it has also been shown to be involved with regulation of PCSK9, a protein that causes lysosome-dependent degradation of LDLR (Maxwell et al. 2005). Gustafsen et al. (2014) isolated primary hepatocytes from Sort1−/− mice and found that PCSK9 secretion was significantly reduced in Sort1-deficient hepatocytes. In mice, ELISA measurements of plasma Pcsk9 showed that Sort1−/− mice displayed a reduction in circulating Pcsk9 and a 2-fold increase in Ldlr. Adenoviral over-expression of Sort1 resulted in increased Pcsk9 and a 27% reduction in Ldlr. These findings suggest that sortilin promotes secretion of PCSK9, in turn promoting LDLR degradation.

This hypothesis has been questioned by Butkinaree et al. (2015), who found that knock-down of SORT1 in Huh7 (hepatocellular carcinoma) cells or in mice has no effect upon LDLR levels. In mouse, no difference in Pcsk9 protein was observed. Co-expression of PCSK9 and SORT1 in HEK293 resulted in a 90% decrease in sortilin.

Gustafsen et al. (2014) and Butkinaree et al. (2015) report disparate findings regarding the relationship between PCSK9 and sortilin. The former postulate that sortilin promotes PCSK9 secretion and is thus a pro-atherosclerotic protein. The latter, however, suggests that the opposite mechanism occurs, and that PCSK9 degrades sortilin, with no effect of sortilin over-expression upon LDLR.
A number of studies have attempted to elucidate the functional mechanisms sortilin is involved in through a number of approaches. Sortilin has been shown to be involved in a number of CAD-related pathways and mechanisms, so it is a strong candidate gene at the 1p13 locus. However, it is ambiguous as to exactly how this protein contributes to, or protects against, CAD. Further investigation of this gene and protein to resolve these conflicting findings is required.

4.1.3 Alternative Mechanisms at the 1p13 Locus

Functional analysis of the 1p13 locus in relation to CAD has predominantly focussed upon SORT1 as the candidate gene at this region. However, CELSR2 and PSRC1 may also be involved in CAD through a mechanism involving SNP genotype effects upon their expression in liver; eQTL data (table 4.3) suggests that SNPs associated with CAD or LDL cholesterol also associate with expression of CELSR2 and PSRC1 in addition to SORT1.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Lead GWAS SNP</th>
<th>Best eSNP</th>
<th>Best eSNP p-value</th>
<th>r² Between GWAS SNP and Best eSNP</th>
<th>Tissue</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELSR2</td>
<td>rs646776</td>
<td>rs646776</td>
<td>6.27x10^{-68}</td>
<td>Same SNP</td>
<td>Liver</td>
<td>Greenawalt et al. 2011</td>
</tr>
<tr>
<td></td>
<td>rs12740374</td>
<td>rs12740374</td>
<td>&lt;1x10^{-16}</td>
<td>Same SNP</td>
<td>Liver</td>
<td>Innocenti et al. 2011</td>
</tr>
<tr>
<td>PSRC1</td>
<td>rs646776</td>
<td>rs646776</td>
<td>3.05x10^{-86}</td>
<td>Same SNP</td>
<td>Liver</td>
<td>Greenawalt et al. 2011</td>
</tr>
<tr>
<td>SORT1</td>
<td>rs646776</td>
<td>rs646776</td>
<td>5.20x10^{-88}</td>
<td>Same SNP</td>
<td>Liver</td>
<td>Greenawalt et al. 2011</td>
</tr>
<tr>
<td></td>
<td>rs12740374</td>
<td>rs12740374</td>
<td>2.86x10^{-22}</td>
<td>Same SNP</td>
<td>Liver</td>
<td>Innocenti et al. 2011</td>
</tr>
<tr>
<td></td>
<td>rs599839</td>
<td>rs599839</td>
<td>1.52x10^{-56}</td>
<td>Same SNP</td>
<td>Liver</td>
<td>Schadt et al. 2008</td>
</tr>
</tbody>
</table>

Table 4.3: Liver eQTL Summary for 1p13 SNPs and CELSR2, PSRC1 and SORT1.

Published eQTL data for these 1p13 genes in liver. For all of these studies/genes, the most significant eSNPs are the same as those SNPs identified by GWAS.

These three genes also appear to show eQTLs in monocytes and macrophages. The best (most significant) SNPs are given in table 4.4, whilst table 4.5 displays the p-values for the GWAS SNPs in both macrophages and monocytes. These data indicate that 1p13 SNP variation effects expression of these genes in monocytes and/or macrophages; it is therefore plausible that the risk conferred to CAD by these SNPs may operate through mechanisms involving these cells.
Table 4.4: Monocyte and Macrophage eQTLs for 1p13 Genes - Best eSNPs. The most significant eSNPs are shown for monocytes and macrophages for the genes SORT1, CELSR2, and PSRC1. \( r^2 \) values are given for these SNPs and the lead SNPs identified by GWAS. “No LD” = no linkage disequilibrium between indicated SNPs. eSNP and related data taken from the Cardiogenics Consortium dataset (Garnier et al. 2013).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Best eSNP</th>
<th>( p )-value</th>
<th>Cell Type</th>
<th>Monocyte</th>
<th>Macrophage</th>
<th>Monocyte</th>
<th>Macrophage</th>
<th>Monocyte</th>
<th>Macrophage</th>
<th>Monocyte</th>
<th>Macrophage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SORT1</td>
<td>rs12726087</td>
<td>7.40x10^{-6}</td>
<td>Monocyte</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs7528419</td>
<td>1.65x10^{-5}</td>
<td>Macrophage</td>
<td>0.895</td>
<td>1</td>
<td>1</td>
<td>0.895</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CELSR2</td>
<td>rs35358959</td>
<td>1.26x10^{-11}</td>
<td>Monocyte</td>
<td>0.179</td>
<td>0.202</td>
<td>0.202</td>
<td>0.179</td>
<td>0.202</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs6667814</td>
<td>1.07x10^{-11}</td>
<td>Macrophage</td>
<td>0.111</td>
<td>0.124</td>
<td>0.124</td>
<td>0.111</td>
<td>0.124</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSRC1</td>
<td>rs660240</td>
<td>3.29x10^{-33}</td>
<td>Monocyte</td>
<td>0.83</td>
<td>0.927</td>
<td>0.927</td>
<td>0.83</td>
<td>0.927</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs4970836</td>
<td>4.4x10^{-30}</td>
<td>Macrophage</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td>No LD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5: GWAS SNP eQTL p-values for SORT1, CELSR2, and PSRC1 for Monocytes and Macrophages. The SNPs identified by GWAS also show significant eQTL associations with SORT1, CELSR2 and PSRC1 expression in monocytes and/or macrophages; \( p \)-values for these associations are given. “NA” = no association. eQTL data taken from the Cardiogenics Consortium dataset (Garnier et al. 2013).
4.1.4 Identification of Putative Functional Variants at the 1p13 Locus

GWAS identified the 1p13 locus as associated with CAD and related phenotypes, but the actual causative variants are unknown. Analyses by Musunuru et al. (2010) have presented compelling evidence to suggest that the 1p13 variant rs12740374 is the functional SNP at this locus.

Gene expression analysis comparing the three genotype groups of rs646776 was performed by Musunuru et al. (2010) for 960 liver samples. Genotype-specific differences were observed for SORT1 (p = 2x10^{-300}), CELSR2 (p = 5x10^{-94}) and PSRC1 (p = 2x10^{-271}), but no significant difference was found for 4 other local genes. For SORT1, CELSR2 and PSRC1, homozygous minor samples displayed the highest expression, with fold changes of up to ~4-fold (SORT1 and PSRC1). No genotype-specific effects were observed for any genes when examined in subcutaneous adipose or omental adipose samples, indicating that this may be driven by a regulatory mechanism that is active in liver but not in these adipose tissues.

Because rs646776 lies between CELSR2 and PSRC1, Musunuru et al. (2010) carried out association analysis between SNPs that lie in this interval and LDL-C using GWAS data. This identified two novel SNPs, in addition to the 4 GWAS SNPs in the region. These 6 SNPs are in high linkage disequilibrium and form two haplotypes with high frequencies in the CEU population (haplotype frequencies = 68% major, 29% minor).

Bacterial artificial chromosomes (BACs) containing these two haplotypes were cloned into luciferase vectors. Musunuru et al. (2010) transfected these into Hep3B cells, and found that the minor haplotype produced higher luciferase expression than the major haplotype vector. To identify which polymorphisms were responsible for this difference, a number of vectors were developed that contained single SNP alterations in the minor haplotype; the minor alleles were replaced with the major alleles and the luciferase assays repeated. From this, Musunuru et al. (2010) identified the major allele of rs12740374 was sufficient to cause a reduction in luciferase expression to a level similar to the construct containing the full major haplotype.

rs12740374 is located in a C/EBPα transcription factor binding site (TFBS); the minor allele (T) of rs12740374 creates the site whilst the major allele disrupts it (G). C/EBPα is a member of the CCAAT enhancer binding protein family of basic region leucine zipper (bZIP) transcription factors (Schrem et al. 2004). All six members of this family contain the sequence CCAAT in their recognition site (Landschulz et al. 1988).
Electrophoretic mobility shift assay (EMSA) experiments by Musunuru et al. (2010) found that the binding of C/EBPα is affected by SNP genotype of rs12740374. Minimal shifting (and therefore binding) was found for the major allele sequence compared to the minor allele sequence, which showed the same degree of shift as the C/EBPα consensus sequence control. Addition of C/EBPα antibodies inhibited the shift, confirming that C/EBPα is the protein binding to the site.

To further demonstrate that C/EBPα binding occurs at the rs12740374 site, luciferase assays using Hep3B cells that express a dominant-negative form of C/EBPα were carried out. Musunuru et al. (2010) observed a reduction in the effect of previously-observed haplotype-specific differences, indicating that C/EBPα binding promotes activation of the enhancer sequence. Conversely, when the dominant-negative C/EBPα is expressed in heterozygous hepatoma cells, the expression of SORT1 decreased 3-fold. Similarly, when over-expressed in homozygous minor HUES-1 differentiated into endoderm, over-expression of C/EBPα increased SORT1 expression. There was no effect of C/EBPα over-expression upon SORT1 in homozygous major HUES-9 cells.

Functional analysis of SORT1 by Musunuru et al. (2010) revealed that over-expression of Sort1 by AAV in mouse significantly reduced total cholesterol and LDL, whilst siRNA knock-down in mouse increased both total cholesterol and LDL. PSRC1 was also examined by AAV over-expression; there was no effect upon total cholesterol or LDL 2 or 4 weeks after AAV injection.

These findings indicate that rs12740374 genotype effects C/EBPα binding. This alteration in binding affinity between genotypes was suggested to result in altered SORT1 expression, leading to dysregulation of cholesterol and LDL levels.

4.1.5 Hypothesis and Experimental Approach

The 1p13 locus has been associated with CAD and related phenotypes in a number of GWAS studies. This locus contains a number of genes, 3 of which have been shown to be affected by 1p13 genotype by Musunuru et al. (2010). This group provides significant evidence for the causal SNP to be rs12740374. The proposed mechanism is that the major allele (G) of this SNP reduces binding of the C/EBPα transcription factor, which in turn decreases expression of SORT1, which is involved in lipid metabolism.

Although the evidence presented by Musunuru et al. (2010) is compelling, no evidence to directly show an effect of this specific SNP upon SORT1 expression has been given. Whilst it may be the case that the C/EBPα binding to this site regulates
SORT1 expression, the data presented by Musunuru et al. (2010) do not preclude the hypothesis that other C/EBPα sites, within the haplotype, are regulating SORT1. Further, the literature investigating SORT1 reaches opposing conclusions regarding the actual role of this protein in lipid metabolism. Some studies conclude that sortilin is associated with reduced risk of CAD, whereas others report evidence of a pro-atherogenic role. A greater understanding of the functional effects of rs12740374 and sortilin are required.

In this work, knock-in of rs12740374 in human HepG2 cells shall be performed. This will generate isogenic cell lines that differ at only the SNP of interest; all other variants remain constant between lines. This allows for direct assessment of the effect of rs12740374 genotype upon C/EBPα binding to the site and the expression of 1p13 genes, including SORT1, which is hypothesised to be the causal gene at this locus.

4.2 Results
4.2.1 Vector Design
When designing the vector to knock-in rs12740374 at the 1p13 locus, several design aspects (section 3.1.2) were considered. rs12740374 is located within the 3’ untranslated region (3’UTR) of the CELSR2 gene. So as to avoid vector-driven effects, the vector was designed so that the selection cassette is inserted outside of CELSR2.

Figure 4.2 shows the final vector design used for the 1p13 project. Track data obtained from the UCSC Genome Browser has been included to illustrate the location of rs12740374 in relation to the CELSR2 gene, and to also provide bioinformatic evidence of a regulatory role of rs12740374.

The ssDNA 1p13 targeting vector is 4300nt long, with a 1408nt left homology arm and a 1388nt right homology arm. The selection cassette is 1504nt long and comprises two loxP sites, two diagnostic PCR primer sequences, a PGK promoter and the G418 resistance gene. The vector has been designed so that 94bp of genomic DNA separates the last base pair of the CELSR2 3’UTR and the loxP site that flanks the left homology arm. rs12740374 is 883nt away from the left homology arm loxP site in the vector, and so it is 789bp from the end of the 3’UTR of CELSR2. ENCODE annotation (Dunham et al. 2012) of regulatory element-associated histone modifications suggests that there may be regulatory regions within the vicinity of the cloning scar; the location of the selection cassette was designed so as to attempt to avoid potential scar-mediated effects.
Figure 4.2: Vector Design for rs12740374 Knock-In, with Annotation. rs12740374 lies within the 3’UTR of CELSR2; the final three exons of CELSR2 are shown (black boxes). HepG2 track data obtained from the UCSC Genome Browser for the locus is also shown. “HepG2 DHS” refers to DNaseI hypersensitivity peaks (indicating a regulatory region). “HepG2 H3Kme1” refers to monomethylation of residue 4 of histone 3 (lysine), a chromatin mark associated with active/poised enhancers. “HepG2 H3K4me2” is another chromatin mark (dimethylation of histone 3 residue 4) associated with transcription factor binding sites. “HepG2 H3K27ac” refers to acetylation of histone 3 residue 27 (lysine) and marks active enhancer elements. “HepG2 CEBPB” represents ChIP-seq data from HepG2 cells for the C/EBPβ transcription factor (C/EBPα data was not available). The “HeLa-S3 CEBPB” track is HeLa-S3 ChIP-seq data. Note the absence of a C/EBPβ peak in the HepG2 track and presence in the HeLa-S3 track. This shows that this locus is capable of binding C/EBPβ. The selection cassette and loxP sites (green boxes) are recombined into the genome immediately downstream of the CELSR2 3’UTR. C/EBPα and C/EBPβ position weight matrices are given, with the T allele of rs12740374 highlighted. HepG2 and vector genotypes are given for the binding site.
4.2.2 Generation of Clonal, Heterozygous HepG2 Lines

4.2.2.1 Targeting and Diagnostic PCR Screening

Genomic DNA was collected from HepG2 cells and sequenced to determine the genotype of rs12740374 and to identify any other variants in the region. HepG2 cells were found to be homozygous for the risk allele (G) of rs12740374. The targeting vector sequence was therefore designed to contain the T (non-risk) allele of rs12740374 as the only modification. The targeting vector plasmid was produced by DNA synthesis. rAAV particles were generated by transfection of the targeting vector plasmid and helper plasmid (pDG-helper, Plasmid Factory, Germany) into HEK293T cells. The virus particles were purified and the titre estimated by qPCR. HepG2 cells were transduced in T75 flasks and cultured for 72 hours prior to seeding in 96-well plates for selection with G418. A total of 4 rounds of transduction were required to optimise the procedure for HepG2 cells. Details of seeding density, selection results and diagnostic PCR screening for all of these transduction events are given in table 4.6.

After the first rAAV transduction, cells were plated in 96-well plates with a seeding density of 1-200 cells/well. After 2 weeks of G418 selection at 0.8mg/ml, 17 wells containing colonies (with an average colony number of 1) were obtained. DNA was isolated from these samples using DirectPCR buffer (Viagen), and subjected to left homology arm diagnostic PCR. None of the samples were positive for integration of the vector at 1p13.

A second round of targeting was performed, utilising higher seeding densities (100-200 cells/well). After selection, a single well contained a colony of cells. Left homology arm diagnostic PCR was not performed for this single sample.

A third transduction was performed, which used greater seeding densities (1000-5000 cells/well) and more wells in order to improve recovery of colonies. G418 selection was performed as per the previous two transduction experiments. Following this, a greater number of colonies were obtained. As expected, the wells seeded with the highest density of cells contained the most colonies. Plates containing 5000 cells/well contained an average of 6 colonies per well post-selection, whilst those seeded with 1000-2000 cells/well contained ~2 colonies/well. The number of colonies/well obtained from all seeding densities was acceptable. However, wells containing fewer colonies are more amenable to diagnostic PCR and successful clonal derivation. Thus, it was determined that seeding densities of 1000-2000 cells/well are optimal, giving approximately 2 colonies per well after selection and a satisfactory number of wells.
containing colonies. DNA was isolated, as described previously, from all wells containing colonies. Left homology arm diagnostic PCR analysis of these samples identified a single sample as diagnostic PCR positive, as indicated by the presence of a knock-in specific band. A high PCR failure rate was observed, with 18.6% of samples failing to amplify. The efficiency of targeting for this transduction event was 0.15%, measured as the percentage of screened samples that were diagnostic PCR positive. A representative gel image is given in figure 4.3, showing the banding observed for both successfully targeted and non-targeted samples.

Figure 4.3: 1p13 Left Homology Arm Diagnostic PCR Screen of HepG2 Cells.
Representative agarose gel electrophoresis image of 1p13 left homology arm diagnostic PCR. PCR-positive samples are indicated by “*”. These samples amplify two products (A) and (B), 1730bp and 1509bp respectively. Product (A) is the wild-type band, amplified by all samples. Product (B) is knock-in specific, and is only amplified from samples containing cells that have integrated the vector at 1p13. Marker (“M”) used is Hyperladder 1kb (Bioline, UK). “G1-G7” refer to the lanes on the gel from which this image was taken. It should be noted that the knock-in specific product (“B”) shows poor amplification in this sample. This is due to there being a high number of colonies in this well; the actual number of which is unknown as only a subset of wells were colony counted to provide an average colony number for this seeding density (4000 cells/well). The average colony number for this seeding density is 4, with a range of 1-10 observed. The number of colonies in this well is known to be greater than 1, so the additional, non-targeted colonies in this well cause preferential amplification of the wild-type (product A) band.
A fourth targeting attempt was performed to increase the number of diagnostic PCR positive samples isolated. Cells were seeded at the optimal density of 2000 cells/well and selected for 2 weeks with G418. The number of wells containing colonies after selection was greater than expected, indicating higher transduction and/or recombination efficiency. Colonies were allowed to grow for an additional week (under selection) to increase colony size and improve DNA yield. DNA was isolated from over 1400 wells with DirectPCR buffer (Viagen, USA) and screened using the left homology arm diagnostic PCR. The PCR failure rate was much improved (6.78%), and 3 diagnostic PCR positive samples were identified (targeting efficiency was 0.22%).

The optimal conditions for 1p13 targeting of HepG2 cells have been developed using the iterative process described. The targeting efficiency with this cell line is lower than that experienced with other cell types (chapter 5). However, multiple clones derived from independent events have been isolated that are diagnostic PCR positive. Genotyping of these lines has confirmed successful integration of the T allele of rs12740374, creating knock-in HepG2 cells heterozygous for rs12740374.

4.2.2.2 Single Cell Dilution and pML-Cre Transfection

The diagnostic PCR positive sample from the third targeting round was identified in a well containing multiple colonies. As such, this sample was subjected to single cell dilution (SCD) (section 2.4.7) prior to pML-Cre transfection to remove the selection cassette. As the other PCR-positive lines (targeting attempt 4) were identified in wells containing lower cell numbers (1-2 colonies/well), a SCD was not deemed necessary.

The SCD performed for the third targeting round sample utilised a range of seeding densities (1-50 cells/well). A total of 1344 wells were plated. Of these, only 278 contained colonies after 2 weeks of growth; 93 of these contained a single colony. Given that higher seeding densities were more likely to give colonies, it can therefore be concluded that single cell expansion of HepG2 cells after rAAV infection and selection is an inefficient process.

The 93 single-colony samples were screened by diagnostic PCR. 10 samples were diagnostic PCR positive; 5 of these clonal lines were taken forward for expansion and transfection with pML-Cre.

pML-Cre transfection was performed for the 5 clones derived from targeting round 3 and the 3 cell samples from targeting round 4 (section 2.4.8). Following pML-Cre transfection, cells were plated at low cell numbers (1-10 cells/well) and cultured for 2-3
weeks to enable colony growth. Cre PCR screening was then carried out following DNA isolation (DirectPCR buffer, Viagen). A representative agarose gel image for the 1p13 Cre PCR is given in figure 4.4. For the 5 SCD clones from targeting round 3, 96 samples (derived from wells containing a single colony of cells) were screened by Cre PCR. Low PCR failure was observed (~1-10%). For two of the clones, no Cre PCR positive samples were found. For the remaining 3 clones, between 1 and 7 clones were identified as Cre PCR positive. 5 clones were taken forward for expansion and banking from two of the SCD clones, giving 10 Cre-positive, clonal lines derived from a single integration event in transduction experiment 3.

Because a SCD was not required for the 3 diagnostic PCR positive samples obtained from targeting round 4, pML-Cre transfection was carried out after expansion of the cells to T75 cultures. All wells containing single colonies were screened by Cre PCR (82-144 samples per line). For all lines, Cre positive clones were identified and expansion attempted. However, for two of the three lines, the clones failed to grow either due to infection or replicative senescence. From the third line, a total of 7 Cre positive clones were expanded and banked.

Figure 4.4: 1p13 Cre PCR Screening of Targeted HepG2 Cells Transfected with pML-Cre. Representative agarose gel image of 1p13 Cre PCR assay. “*” denotes a sample that is Cre PCR positive. Cre-negative samples are characterised by a 408bp band (A), which is amplified from the selection cassette. Cre-positive samples amplify a 224bp band (B). This is only amplified in samples where the selection cassette has been removed and a loxP cloning scar has been left behind. Both Cre-positive and Cre-negative samples amplify a 190bp band (C), which is amplified from non-targeted chromosomes. The band of ~300bp is a non-specific band. As it is only amplified in samples containing band (A), it is possibly an artefact of this product. This band does not prevent identification of Cre-positive samples. Marker used is PCRSizer 100bp ladder (Norgen, Canada).
<table>
<thead>
<tr>
<th>Targeting Attempt Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of 96-Well Plates</th>
<th>Total Number of Wells Plated</th>
<th>Number of Colony-Containing Wells</th>
<th>Mean Colony Number per Well (to Nearest Integer)</th>
<th>Total Number of DNA Samples Screened</th>
<th>Number of Successful PCR Samples</th>
<th>Number of Failed PCR Samples</th>
<th>Number of Positive PCR Samples</th>
<th>% PCR Positive Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>96</td>
<td>0</td>
<td>NA</td>
<td>17</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1</td>
<td>96</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1</td>
<td>96</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1</td>
<td>96</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1</td>
<td>96</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1</td>
<td>96</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>2</td>
<td>192</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>5</td>
<td>480</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1000</td>
<td>3</td>
<td>288</td>
<td>193</td>
<td>2</td>
<td>813</td>
<td>662</td>
<td>151</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>3</td>
<td>288</td>
<td>152</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>3</td>
<td>288</td>
<td>280</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4000</td>
<td>1</td>
<td>96</td>
<td>92</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>1</td>
<td>96</td>
<td>96</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
<td>16</td>
<td>1536</td>
<td>1431</td>
<td>3</td>
<td>1431</td>
<td>1334</td>
<td>97</td>
<td>3</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 4.6: Summary of Transduction Experiments Targeting 1p13 in HepG2 Cells. For each targeting attempt, the seeding density and total number of wells plated is given. Following selection, the number of wells containing colonies and the mean colony number was determined. Left homology arm diagnostic PCR screening was performed; details of the total number of samples screened, the number of failed and diagnostic PCR positive samples are given, along with the overall efficiency of targeting for each transduction.
4.2.3 Generation of Homozygous Non-Risk HepG2 Lines from Heterozygous Cells

A Cre positive line derived from experiment 3 (G7-E6-G5) was re-infected with the rAAV in order to obtain double knock-in lines that are homozygous T (non-risk) for rs12740374. For this experiment, a total of 608 wells across 7 96-well plates were seeded with 2000 cells/well post-infection. After G418 selection, 604 wells contained cells with a mean colony number of 6 per well. 384 DNA samples were screened by diagnostic PCR. 12 samples failed to amplify, and 82 were diagnostic PCR positive. Of these, 12 were expanded and single cell dilutions performed using 1 96-well plate per cell pool. A total of 86 single-colony wells were obtained, of which 62 were used for a second diagnostic PCR screen. 55 of the PCRs successfully amplified, with 16 being diagnostic PCR positive. These were expanded and genotyping attempted. However, both restriction digest genotyping and sequencing both failed to provide conclusive genotype data for the samples. The restriction digest samples repeatedly produced a smear when run on an agarose gel, obscuring the banding pattern. Sanger sequencing across the region produced either unreadable traces or sequence traces that were homozygous for the G allele of rs12740374. It is believed that the PCR used to generate the sequence data may have been preferentially amplifying G-containing DNA, so that heterozygous DNA appears to be homozygous G. It is unclear why this has occurred.

Due to the unresolvable issues with genotyping the potential double knock-in samples, the genotype of the lines could not be confirmed, and the clonal lines obtained were not taken further. For those samples whose sequencing data was readable, the genotype appeared to have remained heterozygous or reverted to homozygous risk (G). The former can happen when integration of the vector occurs into the chromosome into which it occurred previously. The latter is due to the aforementioned preferential amplification of one DNA strand over another.

4.2.4 Complete List of Clonal Cell Lines Obtained for Functional Analysis of the 1p13 SNP rs12740374

A total of 17 clonal, Cre-positive HepG2 lines have been developed that are heterozygous for rs12740374, derived from parental HepG2 cells (homozygous for the G allele). A complete list of all lines obtained is given in table 4.7. A traceability system that utilises the plate locations of the colonies taken forward at each step of the procedure is used – this system allows the derivation of clonal lines to be traced easily and also provides a unique identification number. A simplified nomenclature can also
be used, based upon this system; all references to particular cell lines given in the remainder of the thesis used the simplified nomenclature.

<table>
<thead>
<tr>
<th>Targeting Event</th>
<th>Clonal Line Identification Number</th>
<th>Simplified Identification Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>HepG2-G7-E6-B5</td>
<td>E6-B5</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-E6-C7</td>
<td>E6-C7</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-E6-F5</td>
<td>E6-F5</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-E6-G5</td>
<td>E6-G5</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-E6-H7</td>
<td>E6-H7</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-G5-A10</td>
<td>G5-A10</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-G5-A11</td>
<td>G5-A11</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-G5-B11</td>
<td>G5-B11</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-G5-C5</td>
<td>G5-C5</td>
</tr>
<tr>
<td></td>
<td>HepG2-G7-G5-F12</td>
<td>G5-F12</td>
</tr>
<tr>
<td>4</td>
<td>HepG2-(6)-E2-A9</td>
<td>6-A9</td>
</tr>
<tr>
<td></td>
<td>HepG2-(6)-E2-B1</td>
<td>6-B1</td>
</tr>
<tr>
<td></td>
<td>HepG2-(6)-E2-C1</td>
<td>6-C1</td>
</tr>
<tr>
<td></td>
<td>HepG2-(6)-E2-C10</td>
<td>6-C10</td>
</tr>
<tr>
<td></td>
<td>HepG2-(6)-E2-C11</td>
<td>6-C11</td>
</tr>
<tr>
<td></td>
<td>HepG2-(6)-E2-D11</td>
<td>6-D11</td>
</tr>
<tr>
<td></td>
<td>HepG2-(6)-E2-E7</td>
<td>6-E7</td>
</tr>
</tbody>
</table>

Table 4.7: List of HepG2 Lines Obtained via rAAV Genome Editing of rs12740374.

Cre-positive, clonal cell lines have been generated from editing of the 1p13 SNP rs12740374. All cell lines are heterozygous for this SNP, and were created from two independent transduction/homologous recombination events.

4.2.5 Confirmation of Successful Targeting by Sequencing rs12740374

Sanger sequencing was used to confirm successful rs12740374 genotype alteration. Figure 4.5 shows a representative sequence trace from parental HepG2 and a heterozygous HepG2 line; as can be seen, parental HepG2 cells are homozygous for the G allele of rs12740374. Successful knock-in of the T allele of rs12740374 has generated heterozygous cell lines.
129

Figure 4.5: Parental and Targeted HepG2 Genotyping of rs12740374 by Sanger Sequencing. Sequencing of parental HepG2 cells shows that they are homozygous G for rs12740374. Targeting of this SNP by rAAV has been successful; heterozygosity of rs12740374 is seen in the line E6-G5. “*” = position of rs12740374. “N” = heterozygosity of rs12740374.

4.2.6 Microarray Genotyping of HepG2 Cell Lines

Microarray SNP genotyping was performed using the Axiom™ Genome-Wide UKB WCSG Genotyping Array at Affymetrix (UK) (section 2.2.2.6). Parental HepG2, a pre-Cre heterozygous line and a heterozygous post-Cre recombination line (E6-G5), were genotyped. The genotype of rs12740374 was identical between the sequencing, restriction digestion and microarray methods for all lines examined.

Concordance between all of the lines (genotyped by microarray) was calculated to be 99.92% (in this context, concordance refers to the percentage of SNPs that are identical between the edited and parental lines). This finding demonstrates that targeted cell lines can be considered identical to parental cells, according to this measure, as any differences are likely due to genotyping or calling errors. Assessing concordance is an important consideration so as to ensure that any phenotypic differences that are genotype-specific can be confidently concluded as such, and not simply due to confounding variants that have arisen between parental and targeted lines during the procedure.
4.2.7 Targeting rs12740374 in Mesenchymal Stem Cells (MSCs)

As an additional model in which to investigate rs12740374 function, MSCs were a potential candidate because they are capable of differentiating into hepatocyte-like cells (Lee et al. 2004); such cells would be a good model of hepatocytes in a non-cancer cell line.

A subset of a cohort of MSCs, derived from the Wharton’s jelly of the umbilical cord, was genotyped by sequencing rs12740374. One of the lines genotyped was found to be heterozygous for rs12740374. As rAAV vectors were available for both the G and T alleles, this line was selected for rAAV genome editing.

After infection with either the G- or T-allele rAAV vector, MSCs were plated at high density (500-4000 cells/well) with G418 for selection of targeted cells. Of the wells plated, between 45-86% contained colonies of cells (depending upon seeding density). DNA samples were isolated as described previously, and subjected to left homology arm diagnostic PCR. A total of 444 DNA samples from cells infected with the G allele virus were screened; of the 318 samples that showed amplification, none were successfully targeted.

677 DNA samples from T-allele virus infected cells were screened. Only 288 of these displayed successful amplification. 2/288 samples were diagnostic PCR positive (efficiency of 0.69%). Expansion of these two cell populations was attempted. Unfortunately, one of the samples failed to grow to a T75 culture, whilst the other was capable of this but failed to grow after SCD to generate clonal lines prior to pML-Cre transfection. In both cases, replicative senescence was observed.

This experiment suggests that MSCs, derived from this donor, are not suitable for rAAV-mediated genome editing due to a lack of long-term proliferative ability. Further, findings from 9p21 experiments using MSCs derived from a different donor (section 5.2.2.7) also corroborate this conclusion.

4.2.8 Quality Control of Genome-Edited and “Aged” Controls: Proliferation Comparison

In order to establish if the proliferation of the genome edited and ”aged” cells has been altered by the procedure, compared to the parental cells, the SRB proliferation assay was used to determine the doubling time of the cells. This was used as a proxy for overall cell health and viability.
The SRB assay was performed as described in section 2.2.1, except 5 wells were plated for each line and the cells were left to recover for 48 hours after plating before fixation of the first (day 0) plate. Subsequently, a further 4 days were tested, giving data for a total of 5 days. SRB absorbance data for days 1-4 were normalised to the day 0 values after subtraction of background from control wells. A total of three genome edited lines were analysed (E6-G5, 6-C1 and 6-C10), as well as one aged control line (G2-G8), and these were compared to parental cells. Data for each line is presented in figures 4.6A (normalised absorbance) and 4.6B (log-transformed normalised absorbance, used for doubling time calculation).

Proliferation over the time period analysed appears highly similar between all cell lines compared to parental cells. This indicates that the rAAV genome editing procedure has not influenced the proliferation rate of the cells. Calculating the doubling time of each cell line confirms this observation. The doubling time of the parental HepG2 cells in this experiment is 22.6 hours, the heterozygotes have doubling times of 22.9, 22.4 and 21.5 hours for the E6-G5, 6-C1 and 6-C10 lines respectively. The “aged” control line had a doubling time of 23.6 hours. Because the doubling times of all lines in this experiment are consistent with each other, it may be concluded that neither the genome editing nor “ageing” procedures have influenced cell proliferation.

However, the doubling time of HepG2 cells was determined to be 15.0 hours in the parental characterisation experiment (section 3.2.2). The difference in doubling time between the two experiments may be due to the inclusion of a 48 hour recovery period in this experiment prior to analysis. It is therefore possible that the modification to the protocol contributed to the difference in doubling time between the two experiments.
Figure 4.6: Proliferation Assay for a Subset of the HepG2 Genome-Edited and “Aged” Control Lines Compared to Parental Cells. Cells were allowed to recover from plating for 48 hours prior to beginning the assay. (A) Data shown are normalised absorbance values (to time 0) with background subtracted, calculated from the mean of 5 wells. Error bars represent standard deviation. (B) Data shown are log-transformed values of normalised absorbance values with background subtracted, calculated from the mean of 5 wells.
4.2.9 Assessment of SNP rs12740374 Functionality: Quantifying C/EBPα Transcription Factor Binding by Chromatin Immunoprecipitation-qPCR (ChIP-qPCR)

Upon establishing a bank of cell lines heterozygous for the SNP rs12740374, functional analysis of the SNP was undertaken, starting with quantification of C/EBPα binding at the locus. Such quantification was performed by chromatin immunoprecipitation, followed by qPCR of the immunoprecipitated DNA (see sections 2.11 and 2.13 for a detailed discussion of the protocol used).

% input data was calculated for all samples, and all data were normalised to the parental % input values. The resulting data from this analysis (figure 4.7) reveals an approximate 15-fold increase in C/EBPα binding to the locus in heterozygotes compared to parental cells. This difference is statistically significant, as assessed by the Mann-Whitney test ($p = 0.0277$).

---

**Figure 4.7: rs12740374 Genotype Influences C/EBPα Binding.** ChIP-qPCR analysis comparing C/EBPα binding between heterozygous and homozygous G HepG2 cells reveals a significant increase in binding in rs12740374 heterozygotes compared to homozygous G (parental) cells. Data are mean values, normalised to parental values. Error bars represent the standard error of the mean. $n = 3-4$ for each group. “*” = $p \leq 0.05$. 

---
4.2.10 Assessment of SNP rs12740374 Functionality: 1p13 Gene Expression Analysis

After observing a genotype-specific difference in C/EBPα binding, investigation of a functional relationship between rs12740374 genotype and gene expression was performed by analysing the expression of 8 local 1p13 genes (figures 4.8 and 4.9). Expression of SORT1, CELSR2 and PSRC1 was observed by Musunuru et al. (2010) to be increased in liver samples from rs12740374 heterozygotes compared to homozygous risk individuals. No gene expression difference was observed for the other 1p13 genes examined by Musunuru et al. (2010) when comparing haplotypes. Here, analysis of SORT1, CELSR2 and PSRC1 expression was compared between genotypes of HepG2 cells (figure 4.8). Additionally, to clarify if rs12740374 genotype specifically influences expression of any of the other 1p13 genes, 5 additional genes were examined (figure 4.9).

Three lines, heterozygous for rs12740374, were analysed for all genes, and compared to parental controls across three independent experiments. Normalisation to parental data was performed for comparison between the two genotypes.

There is a statistically significant genotype-dependent effect upon SORT1, CELSR2, PSRC1 and, interestingly, MYBPHL gene expression. For all four genes, heterozygous cells show increased expression compared to the parental controls (assessed by Mann-Whitney test). SORT1 expression is approximately 41% greater ($p = 0.0161$), whilst CELSR2 expression is approximately 33% higher ($p = 0.0156$). PSRC1 showed the most modest expression difference (24% increase) but also the most significant ($p = 0.0103$) due to the least amount of variation in the data. MYBPHL expression is increased by ~31% in heterozygous cells ($p = 0.0463$), as shown in figure 4.9.
Figure 4.8: Expression Analysis of SORT1, CELSR2 and PSRC1: Comparison of rs12740374 Homozygous Risk and Heterozygous Isogenic Cell Lines. Gene expression was measured by qPCR in parental (GG) cells and three heterozygous (GT) cell lines (E6-G5, 6-C1 and 6-C10). All data was normalised to parental cells. RNA isolation from all cell lines and subsequent expression analysis was performed three times. Error bars represent standard error of the mean. “*” = p≤0.05.

Figure 4.9: Expression Analysis of SARS, MYBPHL, PSMA5, and WDR47: Comparison of rs12740374 Homozygous Risk and Heterozygous Isogenic Cell Lines. Gene expression was measured by qPCR in parental (GG) cells and three heterozygous (GT) lines. Experiments were performed three times. Data is presented normalised to parental cells. Error bars represent SEM. “*” = p≤0.05.
4.3 Discussion

The risk allele, G, of the 1p13 SNP rs12740374 has been reported to disrupt the binding of C/EBPα to the binding site at this locus by Musunuru et al. (2010). This is hypothesised to lead to alterations in local gene expression. *SORT1* was of particular interest to Musunuru et al. (2010) (section 4.1.4). *SORT1* has been shown to be involved in multiple lipid metabolism pathways, although the exact functions of the sortilin protein are unclear due to conflicting evidence in the literature. It is therefore conceivable that dysregulation of sortilin through rs12740374 disruption of C/EBPα binding could contribute to altered lipid metabolism and, in turn, to CAD. Whilst the evidence presented by Musunuru et al. (2010) indicates that rs12740374 genotype directly affects *SORT1* expression, it is possible that the observed changes in *SORT1* expression are due to additional C/EBPα sites in the locus, background variation or cell type-mediated effects.

In this work, rAAV genome editing was utilised to generate isogenic HepG2 cell lines, harbouring different genotypes of rs12740374, whilst keeping all background variation constant. From homozygous risk (GG) cells, heterozygous (GT) cell lines were successfully generated through homologous recombination of the vector. Unfortunately, we were unable to produce homozygous non-risk (double knock-in) cells. The absence of cell lines homozygous for the non-risk allele of rs12740374 is undesirable. This is because additional data regarding the influence of rs12740374 genotype upon C/EBPα binding and 1p13 gene expression would have been obtained, potentially providing further evidence of a functional role of rs12740374. Functional studies of lipid metabolism, which were not performed in this work, would benefit from having both heterozygous and homozygous non-risk HepG2 lines in addition to homozygous risk (parental) cells. This is because a greater understanding of the influence of this SNP would have been obtained from the analysis of all three genotypes. For all functional analyses of this SNP (C/EBPα, gene expression and lipid metabolism), analysis of all three genotypes would have enabled the data collected to be used for determining if the effects of the SNP are additive, and the extent to which these parameters are affected by a single SNP. In addition, by assessing homozygous non-risk cells alongside the heterozygous and homozygous risk lines, further evidence to confirm the observed effects are due to SNP editing and are not derived from genome editing procedures, such as prolonged culture, may have been obtainable. In order to account for this, age-matched controls have been utilised in this work. Generating homozygous
non-risk HepG2 lines would have therefore validated the data obtained from heterozygous cells, facilitating interpretation of the data in the analyses performed in this study.

Functional analysis of rs12740374 with the cells generated by genome editing confirmed the previous findings of Musunuru et al. (2010). Binding of C/EBPα to the rs12740374 site was assessed by ChIP-qPCR. This revealed a significant 15-fold increase in binding in heterozygous cells compared to homozygous risk controls. This observation confirms that the T allele of rs12740374 increases the affinity of C/EBPα to the locus. Since binding was still present in GG homozygous cells, this would suggest that the T allele is not essential for binding. However, analysis of the region covered by the ChIP-qPCR assay using PROMO 3.0 (ALLGEN Research Software, available online, Meseguer et al. 2002; Farré et al. 2003) identified two predicted, complete, C/EBPα/β binding sites approx. 60bp upstream of rs12740374. If these sites are real and active in HepG2 cells, this may account for the observed C/EBPα binding in GG homozygotes. Nevertheless, the 15-fold increase in binding shows that the T allele of rs12740374 affects binding significantly, as the sequence of the upstream binding sites is not altered in the edited (heterozygous) cell lines.

After observing a genotype-specific effect upon C/EBPα binding, local gene expression was analysed to determine if the alteration of a single SNP, rs12740374, is capable of affecting gene expression. A total of 8 genes were analysed; three of these (SORT1, CELSR2 and PSRC1) were shown by Musunuru et al. (2010) to be differentially expressed in liver in a genotype-dependent manner. Expression of the remaining 5 genes was assessed in order to identify any other rs12740374 genotype-specific expression differences.

Significant up-regulation in heterozygous (compared to homozygous risk) cells was observed for SORT1 (41%), CELSR2 (33%), PSRC1 (24%), and MYBPHL (31%). These observations indicate that rs12740374 affects expression of these genes by affecting C/EBPα binding. This therefore confirms the finding of Musunuru et al. (2010), and provides further evidence that SORT1 may be involved in CAD.

Unfortunately, we were unable to expand the study to investigate the effect of rs12740374 genotype upon lipid metabolism. Further work to investigate LDL uptake using labelled LDL (permitting quantification of uptake, either fluorescently or colourimetrically), VLDL/ApoB secretion (using triglyceride secretion as a proxy for VLDL secretion or ApoB secretion quantification using ELISA) and other phenotypes
associated with sortilin, utilising these isogenic cell lines and comparing between genotypes, may offer an understanding of the causal mechanisms at this locus and how rs12740374 genotype influences them.

Extensive research has been performed to explore the role of sortilin in CAD, but little attention has been paid to CELSR2 and PSRC1. The evidence presented herein shows that expression of these genes is affected by rs12740374 genotype, potentially implicating their involvement in CAD and warranting a thorough examination of these genes and their protein products. Such a finding is not unexpected, given the eQTL data discussed in section 4.1.3 that suggests that 1p13 SNPs modulate expression of these genes.

In addition to identifying genotype-specific expression of SORT1, CELSR2 and PSRC1, we also observed a similar, significant effect upon MYBPHL. This gene is located within the intervening sequence between PSRC1 and SORT1, and encodes a 354 amino acid protein from 9 exons. The function of the MYBPHL (Myosin Binding Protein H-Like) protein is unknown. The association of MYBPHL expression with rs12740374 genotype is a novel finding, suggesting a potential role for this gene in CAD; this work suggests that MYBPHL is a new candidate gene that may contribute to CAD risk at the 1p13 locus. Further investigation of this gene is warranted, and may be carried out using a number of techniques, including over-expression and knock-down/knock-out experiments in order to identify any overt phenotypic effect. In addition, immunofluorescence may provide an indication of where in the cell MYBPHL localises, and protein-protein interaction studies (such as co-immunoprecipitation and pull-down studies) may help to determine what other proteins MYBPHL associates with.

In this work, effects of rs12740374 genotype have been considered only in hepatocellular carcinoma cells. Future studies utilising other cell types are necessary in order to discover the effect of this SNP in other CAD-related tissues. An example of appropriate cell types would be monocytes and macrophages, which express C/EBPα (Lee et al. 2014). Patel et al. (2015) have provided evidence to suggest that sortilin knock-out reduces foam cell formation in mice. It would be of interest to determine if rs12740374 genotype is capable of influencing SORT1 expression in these cells and the consequence of this upon foam cell formation, which may be measured using cellular lipid concentration as a proxy. Further, eQTL data (tables 4.3 and 4.4) indicates that
1p13 genotype influences expression of SORT1, CELSR2 and PSRC1 in monocytes and macrophages.

From the evidence obtained in this work, SORT1 is not the only gene affected by rs12740374 genotype. In future studies involving liver-derived and other cells, such as VSMCs, a role of CELSR2, PSRC1 and MYBPHL should also be considered and explored experimentally. This may include over-expression and/or knock-out experiments coupled with functional studies of pathways known to be involved in CAD that CELSR2, PSRC1 and MYBPHL may play a role in, such as proliferation.

C/EBPα binding at rs12740374 may exert cis and trans effects, influencing expression of genes outside of the region. Consideration of long-range interactions should therefore be made in future studies. Such analyses would utilise chromatin conformation capture approaches, such as 3C (typically local interactions) and 4C, which typically produces genome-wide interaction data for a specific locus of interest. These analyses may provide evidence to suggest which loci the rs12740374 region interacts with, both within the 1p13 locus and further afield. Such, potentially novel, interaction data, may lead to the generation of additional hypotheses regarding the regulatory role of the region containing rs12740374. Experimental interrogation of these hypotheses may reveal additional regulatory roles for rs12740374 at other loci. The affected gene(s) may be of interest to CAD as well as other diseases. Chromatin conformation capture experiments may also be performed in the absence and presence of cytokines and other signalling molecules, and may reveal how the interactions between genomic regions are modified as a result. This may also provide important data with respect to the regulatory role of 1p13.3 and rs12740374.

Future studies investigating the 1p13 locus may also utilise allele-specific expression analyses. In this approach, the level of expression of a gene of interest from either chromosome can be quantified through the utilisation of tag variants that differ between the two copies of the chromosome (i.e. heterozygotes) and RNA-seq to quantify the number of reads containing either allele of the tag variant. In this manner, one may observe an association between variant genotype and the expression of the gene, suggesting that variants in high linkage are responsible for the difference in expression between the two chromosomes. This approach would provide data showing the effect of the two haplotypes upon local gene expression.

In this work, we have examined the functional effects of a single SNP, rs12740374, in the 1p13 locus. Whilst we have demonstrated a functional effect upon C/EBPα
binding and local gene expression, it is possible that additional variants in the locus (occurring both frequently and rarely) also contribute to CAD. In order to identify such variants, multiple fine mapping techniques have been employed that aim to identify additional, putatively-causal variants. These include utilisation of the Metabochip SNP genotyping array, analysis of the data produced by the 1000 Genomes Project, and sequencing projects, including exome sequencing.

The Metabochip is a custom-designed SNP genotyping array capable of genotyping 196,725 SNPs. This array was designed to enable the follow-up of putative SNP associations found by GWAS for multiple metabolic traits. The Metabochip also contains SNPs to help refine the metabolic trait-associated loci and identify new associations which may direct and inform future functional studies.

Several studies have utilised the Metabochip array in an attempt to both replicate and refine the 1p13 association found by GWAS. The CARDIoGRAMplusC4D Consortium (Deloukas et al. 2013) replicated the association of the 1p13 locus with coronary artery disease using a large cohort of individuals of European or south Asian descent. This group identified an association of SNP rs602633 with CAD (p = 1.47 x 10^{-25}), which has an r² with rs599839 (the initial GWAS lead SNP) of 1.00. rs602633 has an r² of 0.895 with rs12740374. This work has replicated the association with 1p13, but has not been able to refine the interval associated with CAD.

However, similar work using the Metabochip array and a European cohort by Sanna et al. (2011) identified the 1p13 SNP rs583104 (p = 1.28 x 10^{-9}) as associated with LDL-C levels. This SNP has an r² value of 1.000 with rs599839 and is reported to be 31kb from SORT1 and 1kb from PSRC1. This group calculated that the 1p13 locus accounts for 0.61% of the variation in LDL-C levels (when replacing rs599839 with rs583104 in the calculation); with rs599839, the variance explainable by rs599839 alone is 0.63%. This study has also replicated the association of 1p13 with CAD, but has also not been able to refine the associated region further.

The 1000 Genomes Project aims to catalogue the variation in the human genome by performing whole-genome sequencing of a large cohort of individuals from several populations. A meta-analysis of GWAS data from multiple studies has utilised the data produced by the 1000 Genomes Project for SNP genotype imputation (Nikpay et al. 2015). In this work, the 1p13 locus SNP rs7528419 was identified as associated with CAD. This SNP is in high linkage disequilibrium with rs646776, with an r² of 0.85. This work suggests that the same linkage block is involved in genetic risk of CAD as that
found by GWAS. This study has also been unable to refine the association or propose novel putative-causal variants.

Resequencing of disease-associated loci may also help to identify potentially causal variants. This can be targeted sequencing of specific genes of interest, or genome-wide using, for example, exome sequencing. These approaches may identify rare variants that could be involved in disease – further examination of such variants can be used to characterise their effects. Sequencing of the \textit{SORT1} gene by Sanna \textit{et al.} (2011) identified a total of 119 (predominantly) rare variants. The variants are composed of both SNPs and indels, and were found in 3 populations (Sardinians, HapMap CEU and HapMap YRI). Interestingly, the majority of these variants were only observed in one or two of these populations; only 15 of the variants were common to all 3 populations studied (12.6%). The type/location of variant was reported by Sanna \textit{et al.} (2011). Variants assigned as “non-coding” were the most common, accounting for 61.3% of all variants identified in \textit{SORT1}. Other types of variant included 3′UTR (30.3%), non-synonymous (5.0%) and synonymous (3.4%). It is possible that a subset of these variants contribute to CAD through altered sortilin protein function (non-synonymous variants) or expression (non-coding and 3′UTR variants). Further bioinformatic and experimental examination of these variants is needed to identify if any have a functional effect. It should be noted that if any variants within the \textit{SORT1} sequence are functional, this may suggest two regions of 1p13 are involved in the association with CAD. This is because one of the variants identified by Sanna \textit{et al.} (2011) is a known SNP, rs413582, with an $r^2$ of 0.277 with rs599839 in the CEU population, indicating relatively low linkage disequilibrium between these SNPs. The absence of a GWAS signal here demonstrates that any causal variant(s) within the \textit{SORT1} sequence are likely to have a low frequency in the population, and would thus not be detected by GWAS.

Whole exome sequencing of \textit{SORT1} by Sánchez \textit{et al.} (2015) has identified a non-synonymous mutation in exon 4 that alters residue 171 of the sortilin protein from glycine to alanine. When over-expressed in HEK293 cells by transfection with plasmids either expressing wild-type or mutant \textit{SORT1}, this mutation resulted in a 3-fold decrease in expression, whilst also decreasing sortilin protein levels by two-fold.

The work performed by Sánchez \textit{et al.} (2015) was used to investigate the genetic causes of the neurological disease, essential tremor. The mutation identified segregated with the disease in a family of 6 individuals, of which 4 were unaffected. This mutation was not identified in an additional 90 cases of essential tremor, suggesting it is a rare
cause of essential tremor. Whilst this work is not directly relevant to CAD, it does provide evidence for a disease-segregating mutation associated with reduced sortilin expression and protein levels. This, and other variants that influence sortilin expression and function, may be involved in CAD pathogenesis. This work demonstrates the utility of exome sequencing, and suggests that variants that contribute to CAD may be discovered using this approach.

In conclusion, fine mapping using data derived from the Metabochip array and the 1000 Genomes Project have not helped to refine the region associated with CAD, but has replicated the association previously identified by GWAS. However, gene resequencing and exome sequencing work has identified mutations and rare variants, respectively, in the SORT1 gene. It is currently unknown if any of these variants are involved in CAD genetic risk. Further work examining these variants, as well as additional fine mapping, is required to help characterise the 1p13 locus and fully explore its involvement in CAD.

In summary, successful rAAV genome editing of rs12740374 in HepG2 cells (albeit with a lower efficiency than expected) has yielded heterozygous lines from parental (homozygous risk) cells. These cells were used to examine C/EBPα binding and confirmed the findings of Musunuru et al. (2010) directly; the risk allele, G, reduces C/EBPα binding. Next, gene expression analysis also confirmed the involvement of this SNP in SORT1 regulation. Importantly, expression of CELSR2 and PSRC1 was also affected in a similar manner, suggesting these genes may also be involved in CAD. A novel observation has also been made that also connects MYBPHL with rs12740374 genotype. In short, by using rAAV genome editing, this SNP could be examined directly to confirm that rs12740374 is a functional SNP, capable of regulating local gene expression in an allele-specific manner by affecting C/EBPα binding.
Chapter 5: Functional Analysis of the 9p21 Locus Using rAAV-Mediated Genome Editing

5.1 Introduction

5.1.1 Overview of the 9p21 Coronary Artery Disease-Associated Locus

The 9p21 locus was the first region to be associated with coronary artery disease (CAD) and myocardial infarction (MI). This association was identified by several groups in Caucasian populations in 2007 (Helgadottir et al. 2007; McPherson et al. 2007; Samani et al. 2007) and has since been replicated in other ethnic groups, including South Korean (Shen et al. 2008), Japanese (Hinohara et al. 2008), Indian (Shanker et al. 2014), and Hispanic (Assimes et al. 2008) populations. This makes the 9p21 locus the most widely-replicated and robust CAD-associated region.

The 9p21 lead SNPs that have been associated with CAD in the above ancestral groups have not been identified in studies of African populations. However, Kral et al. (2011) found that the 9p21 SNP rs3217989 associates with CAD in an African cohort. This SNP is in low linkage disequilibrium with the lead SNPs identified in the other ethnicities ($r^2 = 0.0008$, $D' = 0.111$ with rs1333049 in the YRI population, calculated using SNAP; Johnson et al. 2008), suggesting a different causal mechanism.

In addition to being associated with CAD, the 9p21 locus has also been associated with a number of other diseases. Such diseases include ischaemic stroke (Smith et al. 2009; Dichgans et al. 2014; Helgadottir et al. 2008), abdominal and intracranial aneurysms (Bown et al. 2008; Yasuno et al. 2010; Helgadottir et al. 2008), chronic lymphocytic leukaemia (Berndt et al. 2013), metabolic syndrome (Bayoglu et al. 2013), endometriosis (Nyholt et al. 2012), and glaucoma (Ramdas et al. 2010; Burdon et al. 2015).

The locus in which the CAD-associated SNPs are located is a non-coding region. However, the genes CDKN2A, CDKN2B, ANRIL and MTAP are nearby and are therefore possibly involved in contributing to CAD.

The CDKN2A and CDKN2B (cyclin dependent kinase inhibitor 2A/2B) genes are also referred to as INK4A and INK4B (inhibitor of CDK4 A/B), respectively. These two genes encode three proteins involved in cell-cycle regulation: p16 and p14/ARF (Serrano et al. 1993; Stone et al. 1995) are encoded by CDKN2A, whilst p15 (Hannon & Beach, 1994), is encoded by CDKN2B.
Figure 5.1: Genes and Key SNPs of the 9p21 Region. The CAD-associated linkage block (black box) is located within the CDKN2B-AS1 (ANRIL) gene. The lead SNP for the 9p21 association, rs1333049 (highlighted green), is shown, as well as the two SNPs investigated in this work (rs10811656 and rs10757278; highlighted red). This image is taken from the UCSC Genome Browser using publicly available data and a custom track containing the CAD linkage block and SNP information.
p16 and p15 both inhibit CDK4 and CDK6 (cyclin dependent kinase 4/6)-mediated phosphorylation of retinoblastoma protein, causing the cell cycle to arrest in G1 (Serrano et al. 1993; Hannon & Beach, 1994). p14 (which is also called ARF as it is transcribed from an alternate reading frame), inhibits cell cycle progression by preventing the ubiquitination of p53 by MDM2 (Honda & Yasuda, 1999) and causes relocalisation of MDM2 to the nucleolus (Weber et al. 1999). p14 therefore promotes p53 activity by preventing MDM2-mediated inhibition of p53.

ANRIL (anti-sense non-coding RNA in the INK4 locus), also referred to as CDKN2BAS (CDKN2B anti-sense), is a long non-coding RNA (IncRNA) that was discovered by Pasmant et al. (2007). This gene, for part of its sequence, is anti-sense to CDKN2B. With a total of 19 “classical” exons (some rare exons are derived from retained intronic sequences), ANRIL has multiple splice forms, including novel circular forms (Folkersen et al. 2009; Burd et al. 2010).

The MTAP gene encodes the enzyme methylthioadenosine phosphorylase. This enzyme is responsible for cleaving S-methyl-5’-thioadenosine (MTA, a byproduct of polyamine synthesis) to adenine and 5-methylthioribose-1-phosphate (Carrera et al. 1984).

5.1.2: Functional Analysis of the 9p21 CAD-Associated Locus
As the most widely replicated CAD-associated locus, a large amount of work has been undertaken in order to understand the role of the 9p21 genes in CAD and how variants in the locus may contribute to the pathogenesis of the disease. In the following subsections, a discussion of the work that has been performed to investigate the genes and proteins of the 9p21 region in the context of CAD is given.

5.1.2.1 CDKN2A & CDKN2B
CDKN2A and CDKN2B have been shown to be differentially expressed in atherosclerotic coronary arteries. Holdt et al. (2011) found, by immunostaining, that p16 is expressed in normal coronary artery smooth muscle and endothelial tissues. p14 was found to be present in the smooth muscle but not the endothelium. In atherosclerotic arteries, all 3 proteins were present in the intimal, medial and adventitial layers, with p15 staining constant between all 3 layers. Staining for p16 was lower in the media, and p14 was reduced in the intima in regions rich in macrophages. Further, p14 was not present in the endothelial cells of the vasa vasorum in the adventitia. All 3 proteins were
present in subsets of macrophages (of which some were foam cells) and smooth muscle cells in the intima.

Several groups have attempted to determine if the genotype of 9p21 affects local gene expression. Holdt et al. (2011) failed to identify an association between 9p21 genotype and protein levels (although a non-significant correlation between ANRIL isoform DQ485454 and p16 was observed). However, Motterle et al. (2012) found that the risk genotype of 9p21 results in decreased expression of p16, p15, and ANRIL in cultured vascular smooth muscle cells (VSMCs). This was associated with increased VSMC proliferation, and, in plaques, increased VSMC content. Liu et al. (2009) also found the same association in peripheral blood T-cells, with decreased expression of p14 in addition to those genes identified by Motterle et al. (2012).

Functional work investigating the CDKN2A and CDKN2B genes has been performed by multiple groups in an attempt to identify the role of these genes in CAD and related phenotypes. siRNA knock-down of CDKN2B in human coronary artery smooth muscle cells (HCASMCs) resulted in increased proliferation, migration, and apoptosis (Leeper et al. 2013), implicating CDKN2B in atherosclerosis. However, Cdkn2b KO mice showed no difference in plaque size, whilst double KO Cdkn2a mice were shown to have reduced plaque sizes compared to heterozygous and WT mice (Kim et al. 2012). Further, mice carrying an extra copy of the Cdkn2a and Cdkn2b genes showed increased expression in liver (significant) and aorta (non-significant) (Fuster et al. 2012). Cultured macrophages from these mice displayed increased apoptosis in response to UV light, but there was no effect upon lesion size, compared to normal mice. Though there was a small increase in VSMC content, no difference in proliferation or apoptosis in cultured VSMCs was observed. Thus, the findings of Kim et al. (2012) and Fuster et al. (2012) differ from those of Leeper et al. (2013).

Deletion of 70kb in mouse, corresponding to the 58kb CAD-associated region in humans, by Visel et al. (2010) resulted in a ten-fold decrease in Cdkn2a and Cdkn2b expression in double-null mice. Back-crossing and subsequent allele-specific expression analysis revealed that the regulation of Cdkn2a and Cdkn2b is a cis mechanism. Further, increased proliferation rate and absent senescence were observed for cultured mouse embryonic fibroblasts (MEFs) and aorta-derived smooth muscle cells (aSMCs) isolated from mice carrying the deletion, compared to WT.
5.1.2.2 ANRIL

Extensive analysis of ANRIL has been performed in order to identify the function(s) of this lncRNA, and also to elucidate any mechanisms through which ANRIL may contribute to CAD.

Holdt et al. (2010) identified an association between 9p21 genotype and expression of some ANRIL transcripts, with the risk genotype increasing the expression of ANRIL. However, no association with genotype was found for CDKN2A, CDKN2B or MTAP. Congrains et al. (2012a) also observed this association with ANRIL. This group also found that siRNA knock-down of ANRIL results in a reduction in CDKN2A and a >2-fold increase in CDKN2B, suggesting that ANRIL negatively regulates CDKN2B. These expression differences were sufficient to reduce the proliferation rate of VSMCs significantly. However, Folkerson et al. (2009) reported a positive correlation between ANRIL and CDKN2A and CDKN2B, and a negative correlation with MTAP.

siRNA- or shRNA-mediated knock-down of ANRIL has led to the identification of multiple genes regulated by this lncRNA. Examples include ADFP (an inducer of foam cell formation) and PPARD (involved in lipid metabolism and cell proliferation), which were identified by Congrains et al. (2012b). Also, Bochenek et al. (2013) have shown that ANRIL is involved in regulation of ADIPOR1 (involved in lipid metabolism), VAMP3 (part of the glucose metabolic pathways) and C11orf10, whilst Bai et al. (2014) have demonstrated that ANRIL regulates CARD8 (involved in apoptosis).

The mechanisms utilised by ANRIL to regulate these genes have begun to be determined. ANRIL has been shown to bind to the chromobox 7 (CBX7) component of the polycomb repressor complex (PRC), which binds to the CDKN2A/CDKN2B locus and represses expression of these genes (Yap et al. 2010). Also, ANRIL binds a member of the PRC2 complex, SUZ12, leading to p15 repression (Kotake et al. 2011). siRNA knockdown of ANRIL reduced SUZ12 binding to CDKN2B, permitting an increase in CDKN2B expression and a reduction in proliferation. Non-PRC mechanisms of ANRIL function at 9p21 have also been identified. The transcription factor E2F1 induces ANRIL expression (Sato et al. 2010) in response to DNA damage (Wan et al. 2013), decreasing expression from the INK4A/ARF/INK4B locus.

Additionally, Zhang et al. (2014) observed binding of the transcription factor EZH2 to the promoters of CDKN2A and CDKN2B in response to over-expression of ANRIL in gastric cancer cells. As it is known that miRNAs are targets of EZH2, Zhang et al. (2014) investigated the possibility that ANRIL may also regulate these miRNAs. They
found that miR-99a and miR-449a were both regulated by ANRIL, which bound PRC2 and resulted in repression of the miRNA genes. An inverse correlation between ANRIL and miR99a/miR449a expression was observed in gastric cancer cells. ANRIL knock-down resulted in decreased expression of miR99a/miR449a-regulated genes (mTOR and CDK6) and increased expression of p15 and p16. Further, a reduction of ANRIL also resulted in decreased E2F1; over-expression of E2F1 resulted in increased expression of both E2F1 and ANRIL. E2F1 was shown, by ChIP, to bind to the promoter of ANRIL. Also, miR-449a over-expression reduced ANRIL expression. Thus, the authors postulate a positive feedback mechanism that operates in the gastric cancer cells used that may also be relevant to other diseases. They propose that increased ANRIL expression reduces INK4A, INK4B, and miR-449a. Lowered miR-449a enables expression and translation of CDK6. The reduction in expression of INK4A and INK4B also results in less inhibition of CDK6. This promotes proliferation and E2F1 expression, which further drives ANRIL expression, hence resulting in a positive feedback loop.

Alu elements within the ANRIL sequence have been shown to mediate interactions with proteins associated with gene induction and repression (Holdt et al. 2013). Expression of ANRIL through stable transfection revealed that ANRIL was capable of regulating a number of genes in trans (9p21 cis effects would not be observed due to the experimental methodology utilised). 708 genes showed a >2-fold increase and 219 showed a >0.5-fold decrease in expression. This finding, taken together with the aforementioned regulatory role of ANRIL, indicates that ANRIL is capable of both cis and trans regulation of genes.

5.1.2.3 Summary of the 9p21 Genes & Their Association With CAD
The literature concerning the association between the genotype of 9p21 SNPs and gene expression is variable with respect to CDKN2A, CDKN2B and ANRIL. Mouse models investigating the functional effects of these genes and CAD have also produced inconclusive findings. There are several challenges when modelling human coronary artery disease in mice, including practical considerations such as the number of mice required and costs of maintenance. Another consideration is the genetic background utilised, which can vary between groups and may have an influence upon experimental findings. With respect to modelling human 9p21 in mice, the orthologous mouse region on chromosome 4 appears to be syntenic with human 9p21, in terms of the presence of
protein-coding genes such as Cdkn2a and Cdk2b, with the interferon cluster up-stream. However, the mouse lacks ANRIL, which means interpretation of experiments assessing the effects of alterations to this region in mice may be misleading when considering human disease. Further, analysis of the 9p21 locus in mice is complicated by the fact that the gene(s) affecting risk of CAD are unknown, as is the exact mechanism by which this occurs, resulting in the absence of specific candidate genes for mouse genetics studies. In addition, the exact regulatory mechanism(s) affected by 9p21 variants in humans may not be found in the orthologous mouse region, as may not the causal variants. A further consideration is that the interaction between 9p21 proteins, whose expression could be affected by 9p21 variants, and other proteins is likely, and this may be part of the aetiology of CAD in humans. In mice, however, these functional interactions may be absent or altered by genetic differences between mice and humans. The inconclusive findings reported in the literature with respect to mouse models of 9p21 genes and pathways could be being influenced by one or more of these factors. The mouse model has been an invaluable tool in the investigation of a number of loci in a number of disciplines. However, it is possible that mouse models have limited applicability to studies investigating the role of human 9p21 in coronary artery disease for the aforementioned reasons.

Recently, Zhao et al. (2015) have analysed genome-wide expression in transformed β-lymphocytes in a cohort of non-Hispanic Caucasian individuals. They report that 9p21 SNP genotype is associated with ANRIL expression, as well as transcripts on other chromosomes. These include DUT (15q21), EIFAY (Yq11), CASP14 (19p13) and DHR59 (2q31), as well as the chromosome 9 gene ABCA1 (9q31).

Given ANRIL’s ability to control gene expression in trans, ANRIL is an interesting candidate gene at the 9p21 locus. If 9p21 genotype does affect ANRIL function, which in turn influences expression of other genes, including the genes identified by Zhao et al. (2015), then the genes normally regulated by ANRIL could be affected. It is important that any variants involved in this mechanism are discovered and characterised.

5.1.3 Putative Functional Variants at 9p21
The identification of putative functional SNPs in the 9p21 locus was first reported by Harismendy et al. (2011). In this study, several chromatin marks were analysed (enrichment of H3K4me1, reduction of H3K4me3, p300 and MED1 binding, and
DNase hypersensitivity) in HeLa cells. This revealed 9 predicted enhancers within the 53kb CAD-associated linkage block, termed ECAD1-9, and a total of 33 enhancers throughout 9p21, making this gene desert the second-most enhancer-enriched region in the genome; it is 6x denser than the genome average. This observation indicates that this region may play a significant role in gene regulation.

Harismendy et al. (2011) sequenced 50 Europeans and identified a total of 131 SNPs with an $r^2$ of 0.5 or more with up to 8 CAD-associated SNPs. 41 of these variants had an $r^2$ of 1.0 with at least one CAD-associated SNP; these SNPs spanned a 44kb region. Of all of the SNPs identified, 33 were located in the enhancers ECAD1-9, with ECAD9 containing 11 variants. Previously, Jariño et al. (2009) had identified 4 conserved regions at 9p21, one of which exhibited enhancer activity assessed by luciferase assay. The conserved enhancer was 2069bp long, and shares 715bp with ECAD8.

Harismendy et al. (2011) identified a number of SNPs that were bioinformatically predicted to create or disrupt transcription factor binding sites (TFBSs) with either the risk or non-risk genotype. ECAD9 contained the greatest number of such SNPs, with 8 expected to alter transcription factor sites. Because of this, Harismendy et al. (2011) focussed on ECAD9, and identified two variants, rs10811656 and rs10757278 (which are separated by just 5bp and have an $r^2$ of 1.0 in Caucasians) that are predicted to affect a STAT1 binding site. The risk alleles of these two SNPs (T and G, respectively) disrupt the STAT1 site, whilst the protective alleles (C and A, respectively) create the site.

This finding is significant, because STAT1 is an effector of interferon-$\gamma$ (IFN-$\gamma$) signalling, which is involved in the immune response. There is significant evidence to suggest that IFN-$\gamma$ is involved in the initiation and/or progression of atherosclerosis (reviewed by Gotsman & Lichtman, 2007).

STAT1 (Signal Transduction and Activator of Transcription 1) is activated via phosphorylation by Janus kinases (JAK), which are themselves activated by phosphorylation due to binding of IFN-$\gamma$ to the receptors IFNGR1 and IFNGR2. Association of IFN-$\gamma$ with the receptors causes their oligomerisation, which triggers phosphorylation of JAK. Once activated, STAT1 forms homodimers and is translocated to the nucleus where it binds a specific DNA sequence (the gamma activated sequence, GAS), where it is able to regulate gene expression (Ramana et al. 2002).

Harismendy et al. (2011) investigated these SNPs further using ChIP and gene expression studies. Firstly, HUVECs were treated with IFN-$\gamma$, which resulted in a
substantial increase in STAT1 binding to the locus. This indicates that the STAT1 TFBS is functional and activated by IFN-γ.

Following this, HUVEC and HeLa cells (which were both reported as heterozygous for the SNPs), were treated with IFN-γ. Expression of ANRIL and CDKN2B was analysed; ANRIL expression increased and CDKN2B decreased in both cell lines. HUVEC cells showed a greater increase in ANRIL compared to HeLa cells (4-fold compared to 2-fold), and also a greater CDKN2B decrease (2-fold).

Lymphoblastoid cell lines (LCLs) were then used to assess the effect of SNP genotype upon STAT1 binding in the same cell type. Homozygous non-risk cells displayed 2.7-fold STAT1 enrichment over IgG control, whereas no binding was observed in homozygous risk cells. siRNA knockdown of STAT1 in homozygous non-risk LCLs resulted in increased (~7-fold) ANRIL expression, whilst there was a 2-fold increase in homozygous risk cells.

Taken together, these findings indicate that differential STAT1 binding, mediated by the genotype of these two SNPs, regulates the expression of ANRIL and CDKN2B in a cell-type specific manner. This therefore implicates an immune system-driven mechanism in controlling expression of these genes, and possibly contributing to CAD.

Harismendy et al. (2011) also examined long-range interactions with loci within 2Mb of the ECAD9 region using chromatin conformation capture (3C). This analysis, performed in HUVEC cells, identified 9 sites that interact with the ECAD9 enhancer. These are located within the CDKN2A/CDKN2B locus, the MTAP gene, downstream of IFNA21 and downstream of the adjacent type 2 diabetes-associated linkage block (which is immediately 3’ of the CAD-associated linkage block).

Fluorescence in situ hybridisation (FISH) analysis of HUVECs revealed that in the absence of IFN-γ stimulation, 37% of analysed chromosomes displayed the interaction between IFNA21 and ECAD9, compared to 58% upon stimulation with IFN-γ.

The interaction between ECAD9 and CDKN2A/CDKN2B or MTAP was also modulated in response to IFN-γ, as assessed via PCR. In the presence of IFN-γ, the interaction between ECAD9 and MTAP decreases, whilst the interaction between CDKN2A/CDKN2B and ECAD9 is increased.

The reported association between the enhancer and IFNA21 may also provide an additional hypothesis as to how the 9p21 region contributes to CAD. The IFNA genes are clustered together and encode the interferon-α proteins. Like IFN-γ, the IFN-α
proteins are also involved in the inflammatory response; dysregulation of these genes could conceivably influence CAD.

These data suggest that IFN-γ stimulation is capable of remodelling the chromatin at this locus, promoting the interaction between IFNA21 and CDKN2A/B. Taken together with their previous observations, Harismendy et al. (2011) propose that the risk alleles of rs10811656 (T) and rs10757278 (G) reduce STAT1 binding in response to IFN-γ stimulation. This, in turn, affects the expression of the gene(s) regulated by ECAD9, contributing to atherosclerosis. However, no direct association between the genotype of these two SNPs and differential gene expression has been investigated.

Almontashiri et al. (2013) analysed expression of CDKN2A and CDKN2B in response to IFN-γ treatment (using the same conditions as Harismendy et al. 2011) in HeLa cells. They report no effect upon CDKN2A and a small increase in CDKN2B. This observation differs from that of Harismendy et al. (2011), and no genotype-driven effect could be found.

Erridge et al. (2013) attempted to determine if there was an effect of 9p21 genotype upon plasma levels of IFN-α or IFN-α21, as well as leukocyte induction of IFN-α, IFN-α21, IFN-β, CXCL10 and total type I IFN at the mRNA, protein and activity levels. No associations with 9p21 genotype and these phenotypes were identified.

These findings question those of Harismendy et al. (2011). However, these studies have not looked at these SNPs in isolation of other variants, either within the 9p21 locus or elsewhere. Therefore, it is possible that whilst STAT1 binding is influenced by genotype, this is not influencing expression of these genes. Instead, other variants at the locus (within the risk haplotype) could be mediating the effect by altering other regulatory elements. Alternatively, the functional effects of other variants may mask/alter the effects of the 9p21 SNPs in these studies through complex interactions of multiple pathways. Because of these possibilities, it is important to confirm that rs10811656 and rs10757278 are the causal SNPs at this locus directly, and attempt to address their functional importance.

5.1.4 Hypothesis and Experimental Methodology

The work by Harismendy et al. (2011) suggests a mechanism by which the CAD risk-associated alleles of two SNPs within an enhancer (rs10811656 and rs10757278) disrupt binding of the STAT1 transcription factor, thereby inhibiting the function of this interaction. This, in turn, is predicted to result in altered 9p21 gene expression in
response to IFN-γ signalling. In this work, this mechanism shall be investigated. In order to examine the functional effect of these two SNPs in isolation, genome editing shall be performed using rAAV-mediated homologous recombination of a targeting vector containing the alternative alleles of the SNPs. In these experiments, HeLa-S3 and HCT116 cells shall be used. HeLa-S3 cells have been selected because HeLa cells were used by Harismendy et al. (2011) to identify the enhancers at the 9p21 locus and experimentally examine their function. HCT116 cells shall be utilised due to their previous use in rAAV genome editing by Horizon Discovery Group plc. These cells will be a useful control for rAAV genome editing, as well as providing an additional cell type in which the 9p21 locus can be examined.

Because both cell lines are homozygous for the protective alleles, a rAAV vector containing the risk alleles of these two SNPs was designed. After alteration of these two SNPs, the effect of SNP genotype was assessed by ChIP-qPCR to assess the level of STAT1 binding at the SNPs. Following this, qPCR analysis of local gene expression was carried out to determine if the genotype of just rs10811656 and rs10757278 is capable of influencing gene expression alterations induced by IFN-γ signalling.

We have chosen to investigate these two SNPs together and perform genome editing of both SNPs. The work of Harismendy et al. (2011) indicates that the STAT1 site these SNPs are located in is affected by these SNPs, influencing binding. In order to assess the impact of the risk haplotype SNPs, both risk alleles shall be knocked-in to the locus simultaneously in order to replicate the STAT1 site sequence in the two haplotypes. This work is therefore a way in which the functional effect of these two SNPs together can be investigated.

Our hypothesis is that the risk alleles of rs10811656 and rs10757278 reduce the affinity of STAT1 to this site. In turn, this may alter the gene expression response observed for local genes upon stimulation with IFN-γ.

5.2 Results

5.2.1 Vector Design

The vector design strategy that was used for the 9p21 locus aimed to ensure that the selection cassette, and hence the loxP site that remains after Cre recombination, was not inserted into the regulatory regions at the locus. The vector design strategy utilised is given in figure 5.2.
It should be noted that the first vector that was used to target the SNPs at this locus yielded very few successfully-targeted cells, and, of those that were targeted, homologous recombination at the location of the SNPs was not successful. This resulted in the integration of the selection cassette but the genotype of the SNPs of interest remained homozygous non-risk.

To circumvent this issue, a second vector was designed. This vector, referred to as the 2nd generation 9p21 vector, was modified in two aspects compared to the first design; the selection cassette was moved closer to the SNPs of interest to increase the probability of successful SNP targeting, and shRNA genes targeting the G418 resistance gene were included (to enrich for cells during selection that have recombined the vector into the target locus). Both homology arms, adjacent to the two ITRs, contain a U6 promoter followed by the antisense and sense shRNA molecules. Each homology arm contains 1099nt of homology, with just the alleles of rs10811656 and rs10757278 different. The selection cassette is 1504nt long.

5.2.2 Efficiency of Targeting and Cre Recombination
Multiple attempts were required to generate heterozygous HeLa-S3 cells, containing a single knock-in (KI) of the targeting vector. In comparison, fewer rounds were required to isolate heterozygous HCT116 cells, which were then used to create homozygous risk (double knock-in, DKI) cell lines. In the following sub-sections, the efficiencies for the two cell lines are discussed.

The nomenclature system used for describing the genotype of the cell lines obtained from 9p21 genome editing (and which will be used extensively in this chapter) is based upon independent chromosomes, with SNP alleles given in sequential order as the genome has been assembled (5′-3′ of forward strand). The system is best explained with examples: “CA/CA” would refer to a cell line homozygous C for rs10811656 and homozygous A for rs10757278. The SNP genotypes are displayed in order along the chromosome, with chromosomes separated by a “/”. Therefore, according to this system, heterozygous lines are always referred to as “CA/TG”, regardless of which chromosome has been targeted (as this has not been discerned). Homozygous risk cells are denoted as “TG/TG”, referring to homozygosity of the T allele of rs10811656 and homozygosity of the G allele of rs10757278. Whilst there are nucleotides separating the two SNPs, these have been omitted from all genotype assignments to clarify the genotype of the SNPs of interest as the intervening DNA is unchanged between lines.
Figure 5.2: Schematic Representation of 2nd Generation 9p21 Targeting Vector with shRNA Directed Against the G418 Resistance Gene. (A) Genomic region encompassing ECAD9 is shown, together with STAT1 and STAT3 transcription factor binding sites in HeLa-S3 cells. Chromatin marks suggestive of regulatory regions are also shown. (B) The vector design, showing left and right homology arms (grey bars) that are separated by a selection cassette containing two loxP sites (green bars) and two diagnostic PCR primers (yellow bars) and a G418 resistance gene. A U6 promoter, sense and antisense shRNA molecules are located in both homology arms (orange bars). (C) Position weight matrix for human STAT1 and the genotypes of the cell lines of interest indicate that binding is permitted in the non-risk genotype and is predicted to be reduced in the risk genotype (targeting vector).
5.2.2.1 Generation of Heterozygous HeLa-S3 Cell Lines

Initial experiments with the first 9p21 vector design yielded no successfully targeted cell lines. Several attempts were made using this virus (detailed in table 5.1). The first attempt with this vector required two plating events. The first, denoted in table 5.1 as 1A, consisted of plating cells at the indicated density and performing selection on the cells. The second, named 1B, involved plating cells at the indicated seeding densities after 2 weeks of G418 selection in a T75 flask. This was performed to ascertain if this method improves the yield of colonies. No colonies grew on any plates seeded for 1A, but the majority of wells contained colonies for plating event 1B. Due to this, diagnostic PCR screening of the large number of wells was performed by pooling up to 4 wells together for the PCR (DNA was stored as individual samples and only combined in the PCR reaction). Because of this pooling strategy, some values in table 5.1 contain values in brackets; these values represent the number of DNA samples that were tested, whereas the value not in parentheses refers to the number of PCR reactions performed.

By utilising DNA sample pooling, the number of PCRs required is reduced. A total of 4 PCR samples, derived from 16 DNA samples, were diagnostic PCR positive. The 16 DNA samples were screened individually using the diagnostic PCR assay, and of these, 3 were positive. In table 5.1, the “% PCR positive samples” column refers to the % of individual DNA samples tested (not pooled samples). All 3 of these samples were genotyped, but none showed alteration of rs10811656 and rs10757278 genotype.

A second round of targeting was performed. In this experiment, a range of seeding densities were used (1-100 cells/well). Of all wells screened (240), only 19 contained colonies. The diagnostic PCR assay was performed on all 19 samples, but none were positive.

Based upon observations from previous experiments, it was determined that the seeding densities utilised in the previous targeting experiments with HeLa-S3 cells were not generating enough colonies for screening, due to low targeting efficiency and the number of plates seeded. To increase the number of wells containing colonies after selection, the seeding density was increased for the third targeting experiment. A range of cell densities were used (200-2000 cells/well), in order to increase the number of colonies available for screening. The largest number of wells containing colonies was obtained, as expected, from the plates containing 2000 cells/well. The average colony number per well for this seeding density was 4, which is acceptable for diagnostic screening. Across all seeding densities tested in this experiment, a total of 592 wells
contained colonies. All samples were subjected to diagnostic PCR analysis, and of these, one sample was positive. Genotyping by sequencing revealed that despite the integration of the selection cassette, the genotype of the SNPs of interest remained the same as the parental cells.

Because of the low efficiency of identifying positive colonies, and the absence of recombination at the position of the SNPs, a new vector was designed and implemented (as described in section 5.2.1). Targeting of HeLa-S3 cells was performed using the same protocol as used for the 1st generation vector. For this experiment, a single seeding density was used (1000 cells/well). An increased number of plates were seeded at this density, in order to maximise recovery of targeted cells. A total of 2084 wells were seeded, with 2082 containing colonies after G418 selection. The diagnostic PCR assay was performed on all samples. It should be noted that 611 samples failed the screening; this was due to either absence of amplification or gel electrophoresis issues (smearing of samples rather than banding). Of the 1471 PCR samples that were successful, 3 were diagnostic PCR positive. Genotyping by sequencing revealed that the cells were heterozygous for rs10811656 and rs10757278. These three cell pools were named after the plate number from which they were grown. As they originated from three different plates, this was sufficient to enable distinction between them. The identification number of the pooled samples was “4”, “5” and “12”. A representative agarose gel image is given in figure 5.3.

Single cell dilutions were performed to obtain clonal lines prior to pML-Cre transfection to remove the selection cassette. Table 5.2 shows the results of the single cell dilution experiment for each of the cell pools identified by diagnostic PCR. From all 3 pools, a total of 288 samples were screened by diagnostic PCR. Of these, 52 samples were positive, indicating that these are clonal lines that have been successfully targeted. 1 clonal line from each pool (“4”, “5” and “12”) was transfected with pML-Cre and screened by PCR (table 5.3).

Seeding of cells after pML-Cre transfection was performed using 5 plates at 1 cell/well. For line 5-D6, the number of wells containing single colonies was suitable for screening, yielding a total of 7 Cre recombination positive lines, 4 of which were expanded and banked.

For line 4-C6, only 22 wells contained cells after 2 weeks of culture. Within all wells, growth was slower than had previously been observed, with only 10 wells containing colonies of sufficient size for PCR screening. Additional culture time of the
cells did not increase the number of colonies available (the remaining cells either senesced or died). The 10 colonies that did grow successfully were Cre PCR screened, but no positives were identified. Due to the compromised proliferative ability of these cells, it was determined that the 4-C6 line should not be carried forward for further work.

Initially, 5 plates of line 12-D4 were seeded. However, no wells on any plate contained colonies after two weeks of culture. The plating of previously pML-Cre-transfected cells was repeated, using 10 plates at 1 cell/well. Of the 182 wells that contained colonies after 2 weeks, 160 were screened by PCR assay. 6 Cre PCR positive lines were found, of which 3 were expanded and banked. Figure 5.4 presents a representative agarose gel for the 9p21 Cre PCR assay.

From the targeting events performed with the HeLa-S3 cell line, a total of 7 Cre-positive lines were derived from 2 independent targeting events. An optimised protocol for targeting HeLa-S3 cells has been developed during the iterative process described above.
### Table 5.1: Summary of 9p21 Targeting Attempts in Order to Obtain Heterozygous HeLa-S3 Cells from Parental, Homozygous Non-Risk Cells

Details of seeding density and colony numbers (following antibiotic selection) are given, along with diagnostic PCR data and targeting efficiencies. For experiment 1B, diagnostic PCR was performed by pooling DNA samples together. A total of 4 DNA samples were pooled and analysed in a single PCR reaction. Bracketed values represent the total number of DNA samples analysed by sample pooling.

<table>
<thead>
<tr>
<th>Vector Design</th>
<th>Targeting Attempt Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of 96-Well Plates</th>
<th>Total Number of Wells Plated</th>
<th>Number of Colony-Containing Wells</th>
<th>Mean Colony Number per Well (to Nearest Integer)</th>
<th>Total Number of Samples Screened</th>
<th>Number of Successful PCR Samples</th>
<th>Number of Failed PCR Samples</th>
<th>Number of Positive PCR Samples</th>
<th>% PCR Positive Samples</th>
<th>Number of Lines with Recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1A</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>384</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2</td>
<td>192</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>192</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1B</td>
<td>1</td>
<td>1</td>
<td>480</td>
<td>282</td>
<td>2</td>
<td>71 (282)</td>
<td>66 (262)</td>
<td>5 (20)</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>288</td>
<td>283</td>
<td>5</td>
<td>71 (283)</td>
<td>70 (279)</td>
<td>1 (4)</td>
<td>1 (up to 4)</td>
<td>0.35</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3</td>
<td>288</td>
<td>288</td>
<td>7</td>
<td>72 (288)</td>
<td>70 (280)</td>
<td>2 (8)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>1</td>
<td>96</td>
<td>96</td>
<td>16</td>
<td>24 (96)</td>
<td>24 (96)</td>
<td>0 (0)</td>
<td>3 (up to 12)</td>
<td>2.08</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5</td>
<td>480</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>288</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1</td>
<td>96</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>1</td>
<td>96</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1</td>
<td>96</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>4</td>
<td>384</td>
<td>15</td>
<td>1</td>
<td>15</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>2</td>
<td>192</td>
<td>19</td>
<td>2</td>
<td>19</td>
<td>16</td>
<td>3</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2</td>
<td>192</td>
<td>57</td>
<td>3</td>
<td>57</td>
<td>50</td>
<td>7</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>3</td>
<td>288</td>
<td>130</td>
<td>3</td>
<td>130</td>
<td>112</td>
<td>18</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>3</td>
<td>288</td>
<td>159</td>
<td>3</td>
<td>159</td>
<td>138</td>
<td>21</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>3</td>
<td>288</td>
<td>227</td>
<td>4</td>
<td>227</td>
<td>194</td>
<td>33</td>
<td>1</td>
<td>0.52</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>21.71</td>
<td>2084</td>
<td>2082</td>
<td>6</td>
<td>2082</td>
<td>1471</td>
<td>611</td>
<td>3</td>
<td>0.2</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of 9p21 Targeting Attempts in Order to Obtain Heterozygous HeLa-S3 Cells from Parental, Homozygous Non-Risk Cells.
### Diagnostic PCR Screen

<table>
<thead>
<tr>
<th>Vector Design</th>
<th>Targeting Attempt Number</th>
<th>Diagnostic PCR Positive Line Well Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of Plates</th>
<th>Total Number of Wells Plated</th>
<th>Number of Wells with Single Colonies</th>
<th>Number of Wells with &gt;1 Colony</th>
<th>Number of Samples Screened</th>
<th>Number of Diagnostic Positive Single Clones</th>
<th>Number of Clones Transfected with pML-Cre</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>960</td>
<td>91</td>
<td>96</td>
<td>182</td>
<td>96</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>960</td>
<td>62</td>
<td>96</td>
<td>36</td>
<td>96</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>1</td>
<td>10</td>
<td>960</td>
<td>36</td>
<td>96</td>
<td>34</td>
<td>96</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 5.2: Summary of Single Cell Dilution Procedure and Observed Efficiency for Heterozygous HeLa-S3 Cell Line Development

<table>
<thead>
<tr>
<th>Vector Design</th>
<th>Targeting Attempt Number</th>
<th>Line ID Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of Plates Seeded</th>
<th>Total Number of Wells Plated</th>
<th>Number of Single-Colony Wells</th>
<th>Number of PCR Samples Screened</th>
<th>Number of Successful PCR Samples</th>
<th>Number of Failed PCR Samples</th>
<th>Number of Cre-PCR Positive Samples</th>
<th>% Cre-PCR Positive</th>
<th>Number of Cre +ve Clones Taken Forward</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>4-C6</td>
<td>1</td>
<td>5</td>
<td>480</td>
<td>22</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5-D6</td>
<td>1</td>
<td>5</td>
<td>480</td>
<td>51</td>
<td>51</td>
<td>51</td>
<td>0</td>
<td>7</td>
<td>13.73</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12-D4</td>
<td>1</td>
<td>10</td>
<td>960</td>
<td>182</td>
<td>160</td>
<td>156</td>
<td>4</td>
<td>6</td>
<td>3.85</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5.3: pML-Cre Recombination Procedure Summary and Observed Efficiencies for HeLa-S3 Heterozygous Cell Lines
2.2 Generation of Heterozygous HCT116 Cell Lines

HCT116 cells were targeted once with the first 9p21 targeting vector, and although this yielded diagnostic PCR positive lines, none of these showed successful recombination, and thus no change in SNP genotype. As such, the second 9p21 virus was then used, and this allowed for the development of a total of 15 heterozygous lines, derived from 3 independent targeting events. Details regarding all targeting experiments carried out using HCT116 cells are given below and in tables 5.4 and 5.5.

The first targeting attempt with HCT116 cells utilised the first 9p21 vector. Three seeding densities were used (200, 500 and 1000 cells/well). Selection was performed using 0.3mg/ml of G418 (as recommended by Horizon Discovery plc, UK). However, this concentration was insufficient, as the cells proliferated with little cell death. The
G418 concentration was increased to 1.1mg/ml, which inhibited growth in the majority of wells. Due to the large number of cells in each well, subsequent growth was identified by media colour change (from red to orange-yellow), indicating a decrease in pH associated with continued growth in the higher G418 concentration. Wells that showed reduced pH were used for the diagnostic PCR screen. This PCR screen of 281 samples identified 3 PCR-positive samples, but, as observed for the HeLa-S3 cells (section 5.2.2.1), none had successfully undergone recombination of the SNPs. 1.1mg/ml G418 was taken forward as the concentration to use for future HCT116 selection.

As the second design of vector was available, this was used for HCT116 targeting. A range of cell densities (100-1000 cells/well) were plated in order to enable optimisation of HCT116 seeding density. A total of 1001 wells contained colonies, with a mean colony number of 1-2 colonies/well for all seeding densities. All of these samples were analysed by diagnostic PCR (a representative gel image is given in figure 5.5). A total of 22 samples (of 847 samples that showed successful amplification) were diagnostic PCR positive. This gives an average efficiency of 2.60% for the entire experiment.

In order to genotype 22 samples, initially, restriction endonuclease digestion was utilised (section 2.4.10.1). The digest results suggested that 12/22 samples were targeted successfully; these results were confirmed by sequencing a subset of digest-positive and digest-negative samples. It should be noted that the number of colonies in each well was 1-2, and therefore sequencing data was expected to be able to provide genotype data for all samples. Three heterozygous samples were expanded and the selection cassette removed by pML-Cre transfection. These three samples were identified by the plate number from which they were derived and the well location on that plate (“1-D3”, “3-F10” and “6-E4”). Due to the low number of colonies in each well, the single cell dilution procedure was omitted. Further, one cell sample (2-C2) that was targeted but did not show recombination of the SNPs was also taken forward to be used as a control (to control for the presence of the loxP site and the “aged” nature of the edited cells). Table 5.5 provides procedural and efficiency data for the Cre recombination process.
Table 5.4: Summary of 9p21 Targeting Attempts in Order to Obtain Heterozygous HCT116 Cells from Parental, Homozygous Non-Risk Cells. Details of seeding density and colony numbers (following antibiotic selection) are given, along with diagnostic PCR data and targeting efficiencies. TNTC = too numerous to count (selection conditions were initially sub-optimal).

<table>
<thead>
<tr>
<th>Vector Design</th>
<th>Targeting Attempt Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of 96-Well Plates</th>
<th>Total Number of Wells Plated</th>
<th>Number of Colony-Containing Wells</th>
<th>Mean Colony Number per Well (to Nearest Integer)</th>
<th>Total Number of Samples Screened</th>
<th>Number of Successful PCR Samples</th>
<th>Number of Failed PCR Samples</th>
<th>Number of Cre PCR Positive Samples</th>
<th>% PCR Positive</th>
<th>Number of Lines with Recombination</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>200</td>
<td>5</td>
<td>480</td>
<td>480</td>
<td>TNTC</td>
<td>95</td>
<td>41</td>
<td>54</td>
<td>1</td>
<td>2.44</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>2</td>
<td>192</td>
<td>192</td>
<td>TNTC</td>
<td>74</td>
<td>52</td>
<td>22</td>
<td>2</td>
<td>3.85</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>2</td>
<td>192</td>
<td>192</td>
<td>TNTC</td>
<td>112</td>
<td>77</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>100</td>
<td>5</td>
<td>480</td>
<td>94</td>
<td>1</td>
<td>94</td>
<td>78</td>
<td>16</td>
<td>1</td>
<td>1.28</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>5</td>
<td>480</td>
<td>233</td>
<td>2</td>
<td>233</td>
<td>233</td>
<td>0</td>
<td>5</td>
<td>2.15</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>5</td>
<td>480</td>
<td>290</td>
<td>2</td>
<td>290</td>
<td>282</td>
<td>8</td>
<td>8</td>
<td>2.84</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>5</td>
<td>480</td>
<td>384</td>
<td>2</td>
<td>384</td>
<td>254</td>
<td>130</td>
<td>8</td>
<td>3.15</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.5: pML-Cre Recombination Procedure Summary and Observed Efficiencies for HCT116 Heterozygous and LoxP Control Cell Lines

<table>
<thead>
<tr>
<th>Vector Design</th>
<th>Targeting Attempt Number</th>
<th>Line ID Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of Plates Seeded</th>
<th>Total Number of Wells Plated</th>
<th>Number of Single-Colony Wells</th>
<th>Number of PCR Samples Screened</th>
<th>Number of Successful PCR Samples</th>
<th>Number of Failed PCR Samples</th>
<th>Number of Cre PCR Positive Samples</th>
<th>% Cre PCR Positive</th>
<th>Number of Cre +ve Clones Taken Forward</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1-D3</td>
<td>12</td>
<td>1152</td>
<td>130</td>
<td>96</td>
<td>95</td>
<td>1</td>
<td>43</td>
<td>45.26</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2-C2</td>
<td>12</td>
<td>1152</td>
<td>81</td>
<td>81</td>
<td>81</td>
<td>0</td>
<td>28</td>
<td>34.57</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3-F10</td>
<td>12</td>
<td>1152</td>
<td>239</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>35</td>
<td>36.46</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6-E4</td>
<td>12</td>
<td>1152</td>
<td>144</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>28</td>
<td>29.17</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

163
For all HCT116 lines transfected with the pML-Cre plasmid, 12 plates at 1 cell/well were seeded. This large number of wells yielded a small number of single colony wells after 2 weeks of culture (at most 20.75%). Up to 96 wells per line were used for Cre PCR screening (figure 5.6). Cre recombination positive clones were obtained for each line (efficiency = 29.17-45.26% of wells). Of the available PCR positive samples, 5 were taken forward for expansion and banking.

Figure 5.5: 9p21 Left Homology Arm HCT116 Diagnostic PCR for Generation of Heterozygous (Single Knock-in) Cells. Representative agarose gel electrophoresis image of 9p21 left homology arm diagnostic PCR screening of HCT116 cells. “*” denotes a targeted sample. Targeted samples display two bands; 2250bp (A) and a 1563bp (B), amplified from wild-type and targeted chromosomes, respectively. Non-targeted samples only amplify band A. Marker used is Hyperladder 1kb (Bioline, UK).

Figure 5.6: 9p21 Cre PCR Screening of Heterozygous HCT116 Cells. Representative agarose gel electrophoresis image. “*” indicates samples with successful Cre recombination. Cre-negative samples can be identified by the amplification of band (A) (400bp), which is amplified specifically from the selection cassette. Cre-positive samples amplify a 236bp product (band B); this can only be amplified if the selection cassette has been removed – this PCR amplifies across the loxP site. Because all samples are heterozygous, all samples amplify a 202bp band (C). This is amplified from the non-targeted chromosome and can be used as a PCR control. The marker ladder used is PCRSizer 100bp ladder (Norgen, Canada).
5.2.2.3 Generation of Homozygous Risk HCT116 Cell Lines

Heterozygous HCT116 cells were developed by rAAV genome editing (section 5.2.2.2). Another round of targeting and screening using a heterozygous line enabled the generation of homoygous risk cell lines. Experimental data pertaining to the targeting procedure is given in table 5.6, with information regarding the Cre recombination process presented in table 5.7.

One heterozygous line, (1-D3-A11) was re-infected with the second 9p21 rAAV vector in order to generate homoygous risk cells. 15 96-well plates were seeded with 2000 cells/well to maximise the yield of colony-containing wells (and therefore potentially increase the number of targeted cells that are identified by diagnostic PCR). A large number of samples was required because generation of homoygous risk HCT116 cells requires integration and recombination of the vector into the chromosome that has not been targeted previously during the generation of heterozygous cells. As this is expected to occur at a frequency of at most 50%, it is therefore possible that screening of twice as many samples will be required to identify cells that are DKIs.

2000 cells/well were plated into 1440 wells in 96-well plates. The cells were treated with G418 for 2 weeks, producing 1218 wells containing colonies, with a mean colony number of 2. 384 of these samples were examined by diagnostic PCR; 15 samples were found to be diagnostic PCR positive from the 239 samples that showed successful PCR amplification. Of these 15 samples, 4 were derived from a single colony of cells, and therefore already clonal. Digestion genotyping was performed, and it was found that 1 of the samples was homoygous risk; a DKI line had been developed (this was confirmed by sequencing). The remaining 11 samples that were comprised of multiple colonies were single cell diluted (4-5 plates were seeded with 1 cell/well). This was performed in order to generate clones from the multi-colony samples to determine if any contained DKI cells. 96 wells containing a single colony (or as many as were available) for each of the 11 cell pools were used for diagnostic PCR. An example gel image is shown in figure 5.7. Of the samples from each pool that were diagnostic positive, 5 were chosen at random and genotyped by restriction endonuclease digestion. None of the clones tested were homoygous risk – they were all heterozygous. Three of the lines were taken forward for use as heterozygous lines that have undergone two full rounds of targeting, and could therefore be useful controls where this is necessary.

After a second round of targeting, two efficiency calculations can be performed using the number of PCR samples. The first is the percentage of total DNA samples that
are diagnostic PCR positive (6.28%), and the second is the percentage of total DNA samples that display successful targeting of both chromosomes after two rounds of infection (0.39%). This indicates preferential re-targeting of previously-edited chromosomes at this locus in HCT116 cells.

Cre recombination of the 1 DKI line plus the 3 heterozygous aged control lines was carried out. Cre PCR screening was used to identify clonal, Cre-recombinant lines (table 5.7, sample gel image figure 5.8). From screening 96 single-colony wells, a high proportion of samples showed successful Cre recombination as assessed by PCR (up to 29%). For all lines, 5 Cre positive clones were expanded and banked.

Figure 5.7: HCT116 9p21 Left Homology Arm Diagnostic PCR for Generation of Homozygous Risk (Double Knock-in) Lines. "*" indicates diagnostic PCR-positive samples, comprising both bands A (single knock-in) and B (knock-in on the other chromosome). Hyperladder 1kb (Bioline, UK) was used as a size marker.

Figure 5.8: Representative Agarose Gel Image of HCT116 9p21 Cre PCR for Generation of Homozygous Risk (Double Knock-in) Lines. "*" indicates a Cre positive sample. Presence of band A (400bp) shows that Cre recombination has not occurred (amplification of selection cassette). Positive samples show a single band (B; 236bp) for lines that have undergone Cre recombination. There is no 202bp band for any Cre-positive DKI samples as both chromosomes have now been targeted. Whilst non-specific bands are present in some of the Cre-negative samples, the Cre-positive samples do not show these bands. Successful selection cassette removal can be detected from this PCR assay. PCRSizer 100bp ladder (Norgen, Canada) was used as a ladder.
<table>
<thead>
<tr>
<th>Vector Design</th>
<th>Targeting Attempt Number</th>
<th>Line ID Number</th>
<th>Seeding Density (cells/well)</th>
<th>Number of Plates Seeded</th>
<th>Total Number of Wells Plated</th>
<th>Number of Single-Colony Wells Screened</th>
<th>Number of Successful PCR Samples</th>
<th>Number of Failed PCR Samples</th>
<th>Number of Cre PCR Positive Samples</th>
<th>% Cre PCR Positive</th>
<th>Number of Cre +ve Clones Taken Forward</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-A11-4-F12</td>
<td>1</td>
<td>10</td>
<td>960</td>
<td>277</td>
<td>96</td>
<td>93</td>
<td>3</td>
<td>27</td>
<td>29.03</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1-G12-A3</td>
<td>1</td>
<td>5</td>
<td>480</td>
<td>99</td>
<td>96</td>
<td>93</td>
<td>3</td>
<td>17</td>
<td>18.28</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3-F2-C12</td>
<td>1</td>
<td>5</td>
<td>480</td>
<td>102</td>
<td>96</td>
<td>94</td>
<td>2</td>
<td>21</td>
<td>22.34</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4-H8-H3</td>
<td>1</td>
<td>5</td>
<td>480</td>
<td>92</td>
<td>96</td>
<td>88</td>
<td>8</td>
<td>22</td>
<td>25.00</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.6: Procedural Summary of Cre Transfection and Subsequent Screening PCR Assay Efficiency for HCT116 Homozygous Risk and Aged Control Lines
5.2.2.4 Complete List of Clonal Cell Lines Obtained for Functional Analysis of the 9p21 SNPs rs10811656 and rs10757278

Successful targeting of the 9p21 SNPs of interest by rAAV genome editing has generated several lines for both HeLa-S3 and HCT116 cells. Table 5.7 lists the cell lines obtained for the HeLa-S3 line, and table 5.8 lists the heterozygous HCT116 cell lines and the associated controls. Table 5.9 details the homozygous HCT116 cells and the control lines for these cells. The simplified identification number for the HCT116 aged control lines is composed of three parts. The first of these is the genotype of the cell line, followed by “S” or “D”, which corresponds to what “age” the lines are a control for. “S” refers to a single round of targeting (equal to the age of heterozygous cells), whereas “D” refers to controls for DKI cell lines (homozygous risk). The third component part of the simplified identification system is a unique number assigned to each line. Further references to the HeLa-S3 and HCT116 genome edited cell lines shall use the simplified nomenclature system.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Clonal Line Identification Number</th>
<th>Simplified Identification Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA/TG</td>
<td>HeLa-S3 (5)-D6-A5</td>
<td>5-A5</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 (5)-D6-B8</td>
<td>5-B8</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 (5)-D6-D3</td>
<td>5-D3</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 (5)-D6-D4</td>
<td>5-D4</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 (12)-D4-B8</td>
<td>12-B8</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 (12)-D4-C10</td>
<td>12-C10</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 (12)-D4-D11</td>
<td>12-D11</td>
</tr>
<tr>
<td>CA/CA &quot;aged&quot; controls</td>
<td>HeLa-S3 D4-A9-A11</td>
<td>D4-A11</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 D4-A9-D8</td>
<td>D4-D8</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 D4-F12-D6</td>
<td>D4-D6</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 D4-F12-F12</td>
<td>D4-F12</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 G2-D6-E4</td>
<td>G2-E4</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 G2-D6-G1</td>
<td>G2-G1</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 G2-F12-D9</td>
<td>G2-D9</td>
</tr>
<tr>
<td></td>
<td>HeLa-S3 G2-F12-F5</td>
<td>G2-F5</td>
</tr>
</tbody>
</table>

Table 5.7: List of HeLa-S3 Lines Obtained by Genome Editing and Aged Controls
<table>
<thead>
<tr>
<th>Genotype</th>
<th>Clonal Line Identification Number</th>
<th>Simplified Identification Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA/TG</td>
<td>HCT116 (1)D3-A9</td>
<td>CA/TG-1</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)D3-A11</td>
<td>CA/TG-2</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)D3-C1</td>
<td>CA/TG-3</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)D3-F1</td>
<td>CA/TG-4</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)D3-H10</td>
<td>CA/TG-5</td>
</tr>
<tr>
<td></td>
<td>HCT116 (3)F10-C5</td>
<td>CA/TG-6</td>
</tr>
<tr>
<td></td>
<td>HCT116 (3)F10-C8</td>
<td>CA/TG-7</td>
</tr>
<tr>
<td></td>
<td>HCT116 (3)F10-D1</td>
<td>CA/TG-8</td>
</tr>
<tr>
<td></td>
<td>HCT116 (3)F10-E8</td>
<td>CA/TG-9</td>
</tr>
<tr>
<td></td>
<td>HCT116 (3)F10-H1</td>
<td>CA/TG-10</td>
</tr>
<tr>
<td></td>
<td>HCT116 (6)E4-E11</td>
<td>CA/TG-11</td>
</tr>
<tr>
<td></td>
<td>HCT116 (6)E4-F5</td>
<td>CA/TG-12</td>
</tr>
<tr>
<td></td>
<td>HCT116 (6)E4-F8</td>
<td>CA/TG-13</td>
</tr>
<tr>
<td></td>
<td>HCT116 (6)E4-H1</td>
<td>CA/TG-14</td>
</tr>
<tr>
<td></td>
<td>HCT116 (6)E4-H12</td>
<td>CA/TG-15</td>
</tr>
<tr>
<td>CA/CA &quot;aged&quot; controls for heterozygous cells</td>
<td>HCT116 (2)C2 - C7</td>
<td>CA/CA-S-1</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)C2 - D1</td>
<td>CA/CA-S-2</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)C2 - D4</td>
<td>CA/CA-S-3</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)C2 - E3</td>
<td>CA/CA-S-4</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)C2 - G10</td>
<td>CA/CA-S-5</td>
</tr>
</tbody>
</table>

Table 5.8: List of HCT116 Heterozygous Lines Obtained by Genome Editing and Aged Controls
<table>
<thead>
<tr>
<th>Genotype</th>
<th>Clonal Line Identification Number</th>
<th>Simplified Identification Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>TG/TG</td>
<td>HCT116 (1)A11-(4)F12-A6</td>
<td>TG/TG-1</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)A11-(4)F12-B8</td>
<td>TG/TG-2</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)A11-(4)F12-E12</td>
<td>TG/TG-3</td>
</tr>
<tr>
<td></td>
<td>HCT116 (1)A11-(4)F12-G7</td>
<td>TG/TG-4</td>
</tr>
<tr>
<td>CA/CA &quot;aged&quot;</td>
<td>HCT116 (2)E3-E2-E10-B9</td>
<td>CA/CA-D-1</td>
</tr>
<tr>
<td>controls for</td>
<td>HCT116 (2)E3-E2-E10-F3</td>
<td>CA/CA-D-2</td>
</tr>
<tr>
<td>homozygous risk</td>
<td>HCT116 (2)E3-E2-H5-E7</td>
<td>CA/CA-D-3</td>
</tr>
<tr>
<td>cells</td>
<td>HCT116 (2)E3-E2-H5-F3</td>
<td>CA/CA-D-4</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)E3-G5-B2-B1</td>
<td>CA/CA-D-5</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)E3-G5-B2-F9</td>
<td>CA/CA-D-6</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)E3-G5-F4-A6</td>
<td>CA/CA-D-7</td>
</tr>
<tr>
<td></td>
<td>HCT116 (2)E3-G5-F4-G9</td>
<td>CA/CA-D-8</td>
</tr>
<tr>
<td></td>
<td>(1)G12-A3-A6</td>
<td>CA/TG-D-1</td>
</tr>
<tr>
<td></td>
<td>(1)G12-A3-B8</td>
<td>CA/TG-D-2</td>
</tr>
<tr>
<td></td>
<td>(1)G12-A3-C4</td>
<td>CA/TG-D-3</td>
</tr>
<tr>
<td></td>
<td>(1)G12-A3-E8</td>
<td>CA/TG-D-4</td>
</tr>
<tr>
<td></td>
<td>(1)G12-A3-F6</td>
<td>CA/TG-D-5</td>
</tr>
<tr>
<td></td>
<td>(3)F2-C12-B3</td>
<td>CA/TG-D-6</td>
</tr>
<tr>
<td></td>
<td>(3)F2-C12-C6</td>
<td>CA/TG-D-7</td>
</tr>
<tr>
<td></td>
<td>(3)F2-C12-D3</td>
<td>CA/TG-D-8</td>
</tr>
<tr>
<td></td>
<td>(3)F2-C12-E7</td>
<td>CA/TG-D-9</td>
</tr>
<tr>
<td></td>
<td>(3)F2-C12-G5</td>
<td>CA/TG-D-10</td>
</tr>
<tr>
<td></td>
<td>(4)H8-H3-A5</td>
<td>CA/TG-D-11</td>
</tr>
<tr>
<td></td>
<td>(4)H8-H3-A11</td>
<td>CA/TG-D-12</td>
</tr>
<tr>
<td></td>
<td>(4)H8-H3-C6</td>
<td>CA/TG-D-13</td>
</tr>
<tr>
<td></td>
<td>(4)H8-H3-D2</td>
<td>CA/TG-D-14</td>
</tr>
<tr>
<td></td>
<td>(4)H8-H3-H2</td>
<td>CA/TG-D-15</td>
</tr>
</tbody>
</table>

Table 5.9: List of HCT116 Homozygous Risk Lines Obtained by Genome Editing and Aged Controls

5.2.2.5 Confirmation of Successful Targeting by Sequencing rs10811656 and rs10757278

Sequencing of SNPs rs10811656 and rs10757278 was performed as per the parental cell line genotyping procedure (section 2.2.5) in order to confirm successful editing of the SNPs. Figures 5.9 and 5.10 display representative sequence traces of the SNPs of interest for parental and edited HeLa-S3 and HCT116 cells, respectively.
Figure 5.9: Sanger Sequencing of Parental (Homozygous Non-Risk) and Heterozygous HeLa-S3 Lines to Confirm SNP Genotype Editing. Sequence traces from parental HeLa-S3 and a representative trace from a heterozygous HeLa-S3 line. “*” = rs10811656, “*” = rs10757278, “#” = rs1333047. Rs1333047 is heterozygous in parental HeLa-S3 cells, whereas rs10811656 and rs10757278 are homozygous. This suggests that rs10811656 and rs1333047 are in low linkage disequilibrium, despite only being 32bp apart. Heterozygous HeLa-S3 line 12-C10 is heterozygous for rs10811656 and rs10757278, whilst rs1333047 has remained heterozygous; any phenotypic differences between genotypes can therefore be attributed to rs10811656 and rs10757278. “N” indicates heterozygosity.
Figure 5.10: Sanger Sequencing of Parental, Heterozygous and Homozygous Risk HCT116 Lines to Confirm SNP Genotype Editing. Sequence traces from parental and edited HCT116 cells (representative). “*” = rs10811656, “**” = rs10757278, “#” = rs1333047. Successful generation of heterozygous (“N”) and homozygous risk HCT116 cell lines from homozygous non-risk parental cells has been performed.
5.2.2.6 Microarray Genotyping of HeLa-S3 and HCT116 Cell Lines
In order to confirm the sequencing data obtained for a subset of the cell lines, samples were submitted for genome-wide SNP genotyping (Axiom™ Genome-Wide UKB WCSG Genotyping Array at Affymetrix, section 2.2.2.6). The genotypes of parental HeLa-S3, parental HCT116 and three pre-Cre heterozygous HCT116 lines were determined using the array (no heterozygous HeLa-S3 lines were available at the time of submission). Array genotyping data for rs10757278 agreed with the sequencing data for all samples.

Further, as the array genotyped SNPs from across the genome, concordance between the parental and edited lines was calculated. The concordance between all three heterozygous lines and the parental HCT116 line was calculated to be 99.92%, indicating that targeted lines can be considered identical to the parental cell lines, as measured by SNP genotype.

5.2.2.7 rs10811656 & rs10757278 Targeting in Mesenchymal Stem Cells (MSCs)
Alongside targeting of rs12740374 (section 4.3.8), MSCs were also targeted with the 9p21 targeting vector (containing the shRNA). Donor 18 was selected as this line was homozygous non-risk for rs10811656 and rs10757278, and so the same vector as that used for the HeLa-S3 and HCT116 experiments could be utilised.

Targeting followed by G418 selection of the MSCs failed to yield any colonies. Six 96-well plates were seeded with 3000 cells/well prior to selection. After 2 weeks of selection, a total of 11 wells contained individual cells. No wells contained more than 2 living cells and no colonies. This indicated that the cells were unable to proliferate following selection.

Experiments targeting the 1p13 locus also failed due to senescence (section 4.2.7). From these observations, it was concluded that MSCs derived from Wharton’s jelly are unsuitable for rAAV genome editing. No further targeting attempts were made using MSCs.

5.2.3 Proliferation Comparison: HeLa-S3
Proliferation of two heterozygous HeLa-S3 and two aged control lines were compared with parental HeLa-S3 cells. The lines chosen were also selected for functional work discussed in sections 5.2.5 and 5.2.6 The lines chosen were 5-B8 and 12-C10 (heterozygous) and D4-F12 and G2-G1 (aged controls). The protocol for this
experiment is a modified version of the parental characterisation SRB assay procedure - 5 wells were plated for each line instead of 8, and the cells were incubated for 2 days after seeding prior to initiating measurement (see section 2.2.1). Data for the analysis of the HeLa-S3 cells is given in figure 5.11A and 5.11B (normalised absorbance and log-transformed normalised absorbance, respectively).

**Figure 5.11: Proliferation Assay for a Subset of the HeLa-S3 Genome-Edited and “Aged” Control Lines Compared to Parental Cells.** (A) Data shown are normalised absorbance values (to time 0) with background subtracted, calculated from the mean of 5 wells. Error bars represent standard deviation. (B) Data shown are log-transformed values of normalised absorbance values with background subtracted, calculated from the mean of 5 wells. These data are used for doubling time calculation.
The doubling time of each line was calculated from the log\(_{10}\)-transformed data. The doubling time of parental HeLa-S3 cells was determined to be longer in this experiment than that observed during the parental cell line characterisation experiments. This was also observed for the HepG2 cells (section 4.2.8). As the protocol utilised for this experiment differed from that used in the characterisation procedure, it is likely the modification of the protocol affected the measurement of the doubling time. Therefore, only comparisons using data from this experiment can be made, in which the protocol was identical for all lines.

The doubling time of parental HeLa-S3 cells was 21.7 hours, whilst it was 20.0 and 21.0 hours for aged control lines D4-F12 and G2-G1, respectively. This indicates that the ageing of the cells does not affect the proliferation rate of HeLa-S3 cells. The doubling time of the heterozygous lines was slightly more variable (21.6 and 23.2 hours for 5-B8 and 12-C10 respectively). These data indicate that the combination of ageing (at least to the degree to which it has been induced in this project) and rAAV genome editing has little effect upon doubling time.

5.2.4 Proliferation Comparison: HCT116
Proliferation of genome edited and aged control HCT116 lines was performed as described previously for HepG2 cells (section 4.2.8) and HeLa-S3 cells (section 5.2.3). A total of 7 lines were chosen for analysis of proliferation in addition to parental HCT116 cells. These consisted of heterozygous, homozygous risk, and aged controls (for both heterozygotes and homozygous risk lines). These lines were also used for additional, functional experiments as discussed in sections 5.2.5 and 5.2.6. The heterozygous lines selected were CA/TG-2 and CA/TG-8. The aged controls for the heterozygotes were CA/CA-S-3 and CA/CA-S-4. Homozygotes TG/TG-3 and TG/TG-5 were used; line CA/CA-D-3 was chosen as an aged control for the homozygous lines. Figures 5.12A and 5.12B show the results of this experiment, with normalised absorbance data given in figure 5.12A and log-transformed data provided in figure 5.12B. A summary of calculated doubling time is given in table 5.10.
Figure 5.12: Proliferation Assay for a Subset of the HCT116 Genome-Edited and “Aged” Control Lines Compared to Parental Cells. (A) Data shown are normalised absorbance values (to time 0) with background subtracted, calculated from the mean of 5 wells. Error bars represent the standard deviation. (B) Doubling times were calculated from log-transformed values of normalised absorbance values with background subtracted, calculated from the mean of 5 wells.
Table 5.10: Calculated Doubling Time of HCT116 Cell Lines from SRB Proliferation Assay Data. Doubling times were calculated from log-transformed data, using values within the linear part of each curve.

The doubling time of the majority of cell lines investigated in this experiment are similar to the parental HCT116 cells. The aged control CA/CA-D-3 has a calculated doubling time of 36.5 hours, which is approximately 2.5x the doubling time of parental HCT116 cells. The remaining lines are consistent with the parental cells (with doubling times differing by up to 3.1 hours).

The observed discrepancy between the parental cells and line CA/CA-D-3 may be due to experimental issues or a problem with the line. With reference to figure 5.12A and 5.12B, it can be seen that this line does not proliferate as well as the others. This may be due to a seeding issue, such that the cell number is low and the cells are remaining in the lag phase of growth for longer than the other cells, which would therefore reduce the doubling time calculated. However, this may represent an issue with this cell line’s proliferative ability, questioning its suitability for further experimentation. Since either of these explanations is possible, the line should not be discarded based solely on these data.

5.2.5 Assessment of SNP rs10811656 & rs10757278 Functionality: Quantifying STAT1 Transcription Factor Binding by Chromatin Immunoprecipitation-qPCR (ChIP-qPCR)

Investigation of the functional effects of SNP genotype upon STAT1 transcription factor binding at the SNP locus was undertaken using chromatin immunoprecipitation (ChIP) coupled with qPCR (see sections 2.12 and 2.13). This technique was used to quantify binding of STAT1 in response to IFN-γ stimulation.
In these experiments, heterozygous HeLa-S3 cells were compared to parental HeLa-S3 cells. Similarly, parental HCT116 were compared to heterozygous and homozygous risk cells. Data for the 9p21 ChIP-qPCR experiments are presented in figure 5.13 for both the HeLa-S3 and HCT116 cell lines and were calculated using the % input method. All edited lines are normalised to the parental control (homozygous non-risk). A statistical comparison between the different genotype groups was performed using the Mann-Whitney test. Statistical significance was accepted at \( p = 0.05 \) or lower. This test was performed as a non-parametric approach was required due to the small sample size.

![Figure 5.13: Relative STAT1 Binding at 9p21 SNPs of Interest in HeLa-S3 and HCT116 cells.](image)

**Figure 5.13:** Relative STAT1 Binding at 9p21 SNPs of Interest in HeLa-S3 and HCT116 cells. Data are normalised to parental cells. Error bars represent standard error of the mean (SEM). \( n = 3-4 \) per group. Data were analysed using the Mann-Whitney test. * \( = p<0.05 \)

ChIP-qPCR analysis revealed a significant decrease in STAT1 binding in heterozygous HeLa-S3 cells, compared to the parental (CA/CA) cells, of approximately 60\% \( (p = 0.0463) \). Similarly, heterozygous HCT116 cells showed a significant ~47\% reduction in binding \( (p = 0.0277) \), whilst homozygous risk (TG/TG) HCT116 cells exhibited a 50\% reduction in STAT1 binding \( (p = 0.0277) \). For both cell types, no STAT1 binding was
detectable in vehicle-only treated cells, confirming that the binding observed in IFN-γ-treated cells is directly caused by exposure to IFN-γ.

These findings indicate that STAT1 binding is affected by rs10811656/rs10757278 genotype; CAD risk-associated alleles reduce affinity of STAT1 to the site. However, the non-risk alleles do not appear to be essential for binding in HCT116 cells, as TFBS occupancy is still observed in homozygous risk cells.

5.2.6 Assessment of SNP rs10811656 & rs10757278 Functionality: 9p21 Gene Expression Analysis by qPCR

5.2.6.1 9p21 Gene Expression Analysis in Response to Interferon-γ
STAT1 binding at the 9p21 locus is affected by rs10811656 and rs10757278 genotype in both HeLa-S3 and HCT116 cells (section 5.2.5). In this experiment, identification of the gene(s) regulated by this interaction was attempted using qPCR (section 2.10). Cell lines were treated with 100ng/ml IFN-γ for 24 hours prior to RNA isolation for gene expression quantification. After qPCR data acquisition, two comparative analyses were made for each gene/cell type. Firstly, the effect of IFN-γ stimulation was compared between untreated and treated cells of the same genotype. Subsequently, the effect of IFN-γ treatment was assessed between genotypes, comparing to aged homozygous non-risk control cells. Data presented in figures 5.14 (IRF-1, CDKN2A (p16) and CDKN2A (p14/ARF)), 5.15 (CDKN2B, ANRIL exon 1-5 and ANRIL exon 4-6), 5.16 (ANRIL exon 17-18 and 18-19 and MTAP), 5.17 (total IFN-α and IFN-β) displays all genotype groups analysed, normalised to untreated cells. Data from HeLa-S3 and HCT116 are presented together, with each graph displaying the data from a single gene.

Statistical analysis comparing untreated and treated cells was performed using the Mann-Whitney test. This test was utilised due to the small number of samples in each data set; a non-parametric test was required. The analysis of expression between untreated and treated cells is given in section 5.2.6.2, whilst the discussion of the effect of genotype upon IFN-γ exposure is given in figure 5.2.6.3.
Figure 5.14: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for IRF-1, CDKN2A (p16) and CDKN2A (p14/ARF). Data are normalised to untreated values. Error bars represent SEM. SEM of treated values were calculated after normalisation to the reference gene and the untreated samples. * = p-value <0.05, ** = p-value <0.01.
Figure 5.15: Gene Expression Analysis in Response to IFN-γ Treatment for Hela-S3 and HCT116 Cells for CDKN2B, ANRIL (exons 1-5) and ANRIL (exons 4-6). Data are normalised to untreated values. Error bars represent SEM. SEM of treated values were calculated after normalisation to the reference gene and the untreated samples. * = p-value <0.05, ** = p-value <0.01.
Figure 5.16: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for ANRIL (exons 17-18), ANRIL (exons 18-19) and MTAP. Data are normalised to untreated values. Error bars represent SEM. SEM of treated values were calculated after normalisation to the reference gene and the untreated samples. * = p-value <0.05, ** = p-value <0.01.
**Figure 5.17: Gene Expression Analysis in Response to IFN-γ Treatment for HeLa-S3 and HCT116 Cells for Total IFN-α and IFN-β.** Data are normalised to untreated values. Error bars represent SEM. SEM of treated values were calculated after normalisation to the reference gene and the untreated samples. * = p-value <0.05, ** = p-value <0.01.

### 5.2.6.2 Effect of IFN-γ Treatment Upon 9p21 Gene Expression

Stimulation of HeLa-S3 and HCT116 cells was assessed using expression of the *IFR-1* gene, which is known to be directly up-regulated by IFN-γ signalling via STAT1 activation (Coccia et al. 1995). Modulation of expression of this gene in response to IFN-γ confirms that the IFN-γ/STAT1 pathway is functional and that the stimulation has worked.

For HeLa-S3 cells, IFN-γ stimulation induces a significant ~30-fold increase in expression of *IRF-1* compared to untreated cells, regardless of genotype (p = 0.0039 for...
both genotypes examined). Significant increases were also observed for CDKN2A (p16) (p = 0.0039 for both genotypes), total IFN-α for homozygous non-risk cells only (p = 0.0103) and IFN-β for both genotypes (p = 0.0039 for both genotypes). A significant decrease in MTAP was found for CA/CA cells (p = 0.0063).

For HCT116 cells, IRF-1 was also significantly up-regulated in response to IFN-γ (by ~16-fold, p = 0.0037, 0.0039 and 0.0039 for CA/CA, CA/TG and TG/TG cells respectively). Further, CA/CA and CA/TG cells exhibited a significant increase in CDKN2A (p14/ARF), with p = 0.0158 and p = 0.0303 respectively. The expression of CDKN2B decreased significantly in response to IFN-γ (p = 0.0364) for CA/CA cells, whilst it increased in heterozygotes (p = 0.0198). ANRIL expression was also affected by IFN-γ exposure, with exon 4-6 increasing significantly in CA/TG and TG/TG cells (p = 0.0039 and 0.0127 respectively). Further, exon 17-18 expression was induced for all three genotypes (p = 0.0037, 0.0064 and 0.0039 for CA/CA, CA/TG, and TG/TG respectively), as was exon 18-19 (p = 0.0050, 0.0039, and 0.0103, respectively). IFN-β was also increased, but only in heterozygous cells (p = 0.0250).

The observation of down-regulation in response to IFN-γ of expression of MTAP and total IFN-α is of interest. STAT1 has been shown to negatively regulate other genes including COL1A1 (Yuan et al. 1999) and MMP13 (Ala-Aho et al. 2000).

DMRTA1 (a local 9p21 gene) expression was also analysed for all lines of both cell types. However, HCT116 cells exhibited very low or absent expression of this gene (below the limit of detection of the assay), whilst HeLa-S3 cells showed highly variable expression of this gene. Because of this, this gene was not analysed further.

These data indicate that IFN-γ has multiple, predominantly modest, effects upon 9p21 gene expression. These alterations differed between the two cell types; such variation is likely to be driven by cell-type specific differences in other regulatory factors that modulate this signalling pathway.

The alteration in gene expression at the 9p21 locus may be caused by the STAT1 binding site containing rs10811656 and rs10757278. Alternatively, additional STAT1 binding sites (and other gene expression regulators that are themselves regulated by STAT1) may also be involved in the regulation of these genes. Determination of the effect of the rs10811656/rs10757278 SNPs, whose risk alleles reduce STAT1 binding (section 5.2.5), was assessed and is discussed in section 5.2.6.3.
5.2.6.3 rs10811656/rs10757278 Genotype-Driven Effects upon the 9p21 Gene Expression Response to IFN-γ Signalling

Following the analysis discussed in section 5.2.6.2, the effect of rs10811656 and rs10757278 genotype upon the regulation of 9p21 gene expression in response to IFN-γ was assessed by comparing the fold-change in expression induced by IFN-γ between genotype groups (for HCT116, comparisons were made between CA/CA cells and CA/TG or TG/TG cells).

For HeLa-S3, the expression of CDKN2A (p16) was significantly different between treated CA/CA and treated CA/TG cells ($p = 0.0374$). In contrast, expression of CDKN2B and ANRIL exon 18-19 were both significantly affected by SNP genotype in HCT116 cells, compared to treated CA/CA cells. For CDKN2B, $p = 0.0039$ and 0.0103 for CA/TG and TG/TG cells, respectively, whilst for ANRIL exon 18-19, $p = 0.0196$ for CA/TG cells compared to CA/CA. No significant difference was observed between CA/CA and TG/TG cells for ANRIL exon 18-19. These data suggest that the genotype of rs10811656/rs10757278 affects IFN-γ-mediated regulation of 9p21 genes, and that the effects are cell type-specific.

Additional normalisation of the data to control for variation in IFN-γ treatment variation can be applied. To do this, the IRF-1 expression level is used to normalise all samples after 36B4 reference gene normalisation. This approach assumes that identical stimulation conditions would yield identical IRF-1 up-regulation across lines. If this analysis is carried out and the statistical analysis repeated, there is no genotype-driven effect upon IFN-γ regulation of genes in either cell type. This suggests that variation in treatment between lines contributed to the significant differences previously observed. Indeed, for the HeLa-S3 cells, the mean fold-induction of IRF-1 differed by ~3.5-fold, whilst the largest difference between HCT116 IRF-1 inductions is ~1.5-fold. These large differences could have contributed to the difference in 9p21 gene expression between genotypes. Thus, these data, after analysis, indicate that the genotype of rs10811656/rs10757278 does not affect the expression of the genes examined in response to IFN-γ signalling.

5.3 Discussion

In this work, two SNPs have been investigated at the 9p21 locus. These two SNPs have been chosen for examination together as they are both located within the same STAT1 transcription factor binding site and it is this site that is of interest.
Genome editing of the two SNPs (rs10811656 and rs10757278), located 5bp apart at the 9p21 locus, has been successfully achieved in two cell lines. HeLa-S3 cells, which are homozygous non-risk (CA/CA), were used to generate heterozygous lines (CA/TG). Low efficiency of successful vector recombination in this line resulted led to an extensive optimisation period that resulted in an inability to develop homozygous risk (TG/TG cell lines).

However, HCT116 cells (also CA/CA genotype) were also targeted. With a higher efficiency of successful targeting, heterozygous and homozygous risk lines were produced.

The genome editing protocol was optimised to improve efficiency with both of these cell lines in an iterative manner, allowing development of rAAV targeting protocols for these lines for 9p21 targeting. Targeting efficiency of HeLa-S3 cells was 0.2% (as assessed by the number of diagnostic PCR positive samples). Cre transfection efficiency was variable for HeLa-S3 cells (3.9-13.7%). HCT116 targeting efficiency was higher (1.28-3.85%, mean 2.24%), as was Cre transfection efficiency (up to 45.3%) compared to HeLa-S3.

Upon completion of cell line development, functional analyses were performed. These comprised, following IFN-γ treatment, ChIP-qPCR to quantify STAT1 binding site occupancy at rs10811656/rs10757278, and qPCR to measure the expression of local genes.

For HeLa-S3 cells, where only homozygous non-risk and heterozygous cell lines were available, a significant, genotype-specific effect upon STAT1 binding was observed. Heterozygous cells showed ~60% less STAT1 binding ($p = 0.0463$), indicating reduced affinity of STAT1 to the locus.

A significant reduction in STAT1 binding of approximately 40% was observed for both heterozygous and homozygous risk HCT116 cells compared to parental homozygous non-risk lines ($p = 0.0277$ for both edited genotypes). As for HeLa-S3 cells, there was no observable binding of STAT1 in the absence of IFN-γ.

From the ChIP-qPCR data obtained from both cell lines, it may be concluded that STAT1 binding at this locus is affected by rs10811656 and/or rs10757278 genotype. The CAD risk-associated alleles of these two SNPs are sufficient to reduce affinity for STAT1 at the locus, but homozygosity of the risk-associated alleles does not prevent STAT1 binding.
The ChIP-qPCR data presented in figure 5.13 suggests a dominant effect of SNP genotype upon STAT1 binding, as the TG/TG lines had a higher binding frequency than both the CA/TG and CA/TG HCT116 lines (which showed similar levels of binding). This observation contrasts with the findings from a meta-analysis of GWAS data, which found that an autosomal-additive model of inheritance explained the 9p21 association the most satisfactorily (Schunkert et al. 2008). The observation of an apparently dominant effect would suggest that both rs10811656 and rs10757278 do not contribute to CAD variation at 9p21. However, it is also possible that the dominant effect observed is an artefact of the experiment performed. When stimulating with IFN-γ in this work, 100ng/ml was used and treatment was allowed to progress for 24 hours. This concentration and duration was selected in order to replicate the work of Harismendy et al. (2011) and Almontashiri et al. (2013). However, the duration and concentration used when stimulating cultured cells with IFN-γ varies within the literature. Erridge et al. (2013) stimulated PBMCs with 100ng/ml IFN-γ for 2 hours, whilst Liu et al. (2011) stimulated MSCs with 50ng/ml for up to 4 hours. Torvinen et al. (2007) treated PBMCs and lung epithelial cells with 0.1-10ng/ml IFN-γ for 18 hours. Klampfer et al. (2007) examined the effects of IFN-γ stimulation using Hke-3 cells (a cell line derived from HCT116). In this work, this group treated the cells with 10ng/ml of IFN-γ for 1 and 3 hours. These differences in protocol demonstrate the variability between publications and may also reflect differences in sensitivity to IFN-γ exhibited by the cells investigated. Thus, it may be postulated that the concentration and duration used in this work may not be optimal for HeLa-S3 and HCT116 cells, contributing to the observed dominant effect with respect to STAT1 binding at the SNPs of interest.

An additional factor which may contribute to the apparently dominant effect is the presence of additional STAT binding sites near to the site of interest. Using the online transcription factor binding site prediction tool PROMO 3.0 (Messeguer et al. 2002; Farré et al. 2003), analysis of 250bp either side of rs10811656 (500bp in total) was performed. This revealed the expected STAT binding site at rs10811656 and rs10757278 (this was annotated as a STAT3 site by PROMO 3.0; the consensus sequences of STAT1 and STAT3 are highly similar; Qing & Stark, 2004). In addition, two further putative STAT sites were predicted, both upstream of the site of interest. These were STAT5B (the start of this site was 142bp upstream of rs10811656) and STAT1B (90bp upstream of rs10811656). Both of these sites are not within the PCR amplicon used by the ChIP-qPCR assay. However, it is possible that the occupancy of
the upstream STAT1 site led to immunoprecipitation (and subsequent PCR amplification) of DNA that did not have STAT1 bound to the rs10811656/rs10757278 site due to the random shearing of the DNA by sonication. This, in turn, may have affected the ability of the qPCR assay to detect subtle differences in STAT1 binding at the site of interest, and effectively reducing the window of the assay by potentially increasing background amplification. The raised background could, in combination with experimental variation, prevent the detection of a difference in STAT1 binding at rs10811656/rs10757278 between heterozygous and homozygous risk cells. We therefore postulate that the combination of background PCR template, in conjunction with sub-optimal treatment conditions, led to the apparently dominant effect of rs10811656/rs10757278 genotype upon STAT1 binding.

Gene expression analysis of local genes was employed to elucidate what effect this reduction in STAT1 binding has on 9p21 gene expression. HeLa-S3 IFN-γ stimulation resulted in up-regulation of the majority of genes within the locus by a small amount in both genotype groups, and a slight decrease in expression of some genes. The largest expression difference was found for IFN-β, which was up-regulated by between 4- and 5-fold in HeLa-S3. Total IFN-α expression was reduced by approximately 20-30%, whilst CDKN2A (p16) expression increased ~40-60%. Expression changes induced by IFN-γ in all other genes tested are modest, with small up- and down-regulations observed. Analysis of HCT116 9p21 gene expression revealed small, significant differences in expression induced by IFN-γ, with IFN-β showing the greatest fold change (52% in heterozygous cell lines).

Whilst the effect of IFN-γ on 9p21 gene expression is of interest, determination of whether SNP genotype at the locus can influence the effect of IFN-γ signalling is of greater importance. Normalisation to the IRF-1 gene (which is known to be up-regulated by IFN-γ signalling via STAT1) allowed for correction of variation caused by slight differences in IFN-γ treatment. With this normalisation, significant genotype-specific effects that were observed are no longer statistically significant. Thus, it is suspected that small variations in stimulation conditions contributed to the small differences between genotypes. Additionally, this could have been exacerbated by the small effect of IFN-γ upon gene expression.

Taken together, the findings presented herein confirm that STAT1 TFBS occupancy is affected by rs10811656/rs10757278 genotype. The risk alleles have been shown to
reduce affinity for STAT1, but homozygosity of these alleles does not prevent binding (indicating that the non-risk alleles are not essential for binding).

Further, IFN-γ treatment results in modulation of 9p21 gene expression in a cell type-specific manner. However, this modulation is not affected by rs10811656/rs10757278 genotype when controlling for IFN-γ treatment variation by normalisation to a known IFN-γ induced gene.

The difference in gene expression response to IFN-γ stimulation demonstrates that cell type-specific effects may be an important consideration for studies of this type. For the majority of genes examined, the effect of IFN-γ stimulation is different between HeLa-S3 and HCT116 cells, with IRF-1, CDKN2A (p16), total IFN-α and IFN-β exhibiting the largest discrepancies. This means that it is difficult to appreciate the relationship between these SNPs, interferon-γ stimulation and gene expression on CAD risk in vivo, as the two cell types used are not relevant to CAD pathogenesis. HeLa-S3 and HCT116 cells may therefore have limited suitability when investigating CAD functional genetic studies. However, these cells have permitted the first use of rAAV genome editing to directly assess the effects of SNPs associated with CAD. As this technology has been demonstrated in this work as a valid method of investigating individual regulatory variants, future studies may consider using alternative, more relevant, cell types to investigate CAD-associated variants. These cells may include iPSCs, which would be capable of differentiation into multiple cell lineages associated with CAD pathogenesis.

The data presented herein demonstrates that these two SNPs are functional as they are capable of affecting the binding of STAT1, but this variation in binding does not result in a detectable difference in gene expression between genotype groups. It is therefore likely that additional variants in the haplotype contribute to CAD through alternative mechanisms, which may or may not involve this STAT1 site. As such, this site could have a functional effect upon CAD risk in combination with other variants and interacting pathways that were not assessed in this work.

Additional functional experiments may also be performed using the edited cell lines generated in this work. These studies were beyond the scope of this study, but may include proliferation in the presence and absence of interferon-γ to identify if CDKN2A and CDKN2B function is affected by the genotype of these SNPs in response to interferon-γ. Further, migration may also be assessed through the utilisation of the scratch assay (Liang et al. 2007), and could also be affected by 9p21 gene expression
dysregulation. In addition, proliferation and migration are two phenotypes associated with atherosclerosis development, and so examination of the role of 9p21 variants in these cellular functions is warranted. Also, as CDKN2A and CDKN2B are involved in regulation of mitosis, the effect of cellular stress caused by DNA damage and reactive oxygen species may be examined with a view to identifying if 9p21 variants affect these cellular responses. This may be examined by monitoring cell proliferation or apoptosis.

ANRIL is an interesting feature of the 9p21 locus, unique to humans and other primates. Characterisation of the functions and the diversity in isoforms of this lncRNA will be informative in that novel disease-associated functions may be found. It is possible that variants within the sequence of ANRIL are capable of either affecting the expression of ANRIL (or other genes), or the stability and function of ANRIL transcripts. Alternatively, these variants may affect splicing, preventing synthesis or creating novel isoforms of ANRIL which may affect function.

Because the exact role (if any) of ANRIL in CAD is unknown, further experimental characterisation is needed, consisting of both discovering all ANRIL isoforms and determining their function. We may start to investigate ANRIL function by identifying where ANRIL localises within the cell, as well as what proteins and genomic loci it interacts with. To identify which sub-cellular regions contain ANRIL, FISH analysis may be used. Evidence suggests that ANRIL is capable of binding proteins; to identify which proteins form an association with ANRIL, RNA pull-down techniques may be used that enable the identification of associated proteins with mass spectrometry analysis. A modified version of FISH that utilises the co-localisation of RNA probes and antibodies may also be used, but this technique is not readily amenable to high-throughput analyses as it requires the use of specific antibodies against proteins of interest. To determine the genetic loci ANRIL binds, methods such as CHART (capture hybridisation analysis of RNA targets), developed by Simon et al. (2011), may be used. This technique, developed for lncRNA analysis, purifies RNA:DNA heteroduplexes using affinity-tagged oligonucleotides following reversible cross-linking. Interacting DNA can be identified using sequencing; the technique was shown by Simon et al. (2011) to be amenable to genome-wide analysis, as well as suitable for isolating proteins that interact with the RNA of interest. These approaches may reveal additional regulatory roles of ANRIL that could be involved in diseases such as CAD.

Whole-haplotype analyses may need to be performed to identify the effect of all variants in the risk and protective haplotypes. In this work, two SNPs have been
investigated in isolation from all other variants. Whilst this offers information regarding the functional effects of these SNPs, it does not provide information as to how their role is attenuated by other variants in the locus. The results obtained here may therefore not truly reflect the roles of these SNPs \emph{in vivo}. Genome editing of a larger number of SNPs in the haplotype (individually and in combination, forming an allelic series experiment) followed by assessment of functional effects may permit the elucidation of the complex network of interactions that are likely to be acting at the 9p21 regulatory region. Such an approach would reveal haplotype-dependent effects and allow for the modelling of the risk and non-risk-associated haplotypes in a genetically-identical background.

An alternative approach to investigating rs10811656 and rs10757278 together would also be to investigate the effect of each separately. Possible through genome editing, this would allow for the generation of SNP allele combinations that are not common due to linkage disequilibrium. Such an investigation would permit examination of, whether one SNP or the other has more of an effect upon STAT1 binding. Also, the role of each SNP in the regulation of genes through other transcription factor-mediated mechanisms may also be investigated.

This work has focussed on investigating the role of rs10811656 and rs10757278 in the context of gene regulation through interferon-$\gamma$ signalling, as this was the mechanism proposed by Harismendy \emph{et al.} (2011). Consideration of the effect of SNP genotype upon STAT1-mediated regulation of gene expression after interferon-$\gamma$ stimulation, including \emph{ANRIL}, was given. Future work should also focus on the functions of \emph{ANRIL} in combination with the regulation of gene expression of the 9p21 enhancers, including, but not limited to, the rs10811656/rs10757278 and IFN-$\gamma$ pathway explored in this work. If \emph{ANRIL} gene expression is altered through variant genotype-mediated effects through modified enhancer activity, it is possible that the expression level of some or all \emph{ANRIL} transcripts will be affected. As it may be hypothesised that the multiple transcripts of \emph{ANRIL} have different regulatory functions, enhancer effects that influence the expression profile of \emph{ANRIL} could be diverse. The exact nature of these effects would need to be explored after the elucidation of the functions of \emph{ANRIL}. In order to discover these, many molecular genetics approaches may need to be taken, as described above. Genome editing may also be used for the examination of \emph{ANRIL} function through many experimental methodologies. These may include the insertion of expression reporters, such as GFP or luciferase, to reveal the signalling pathways that regulate expression of \emph{ANRIL} by monitoring the expression of the reporter as a proxy of
ANRIL expression. Knock-out of ANRIL or removal/alteration of specific exons or motifs within exons may also help to identify the role(s) of ANRIL.

Another way in which our understanding of the 9p21 locus could be improved is by fine mapping of the locus. Fine mapping of the 9p21 locus has been attempted using a number of approaches with a view to refine the association with CAD and also identify novel, potentially functional variants for further characterisation. Fine mapping may also be able to identify independent association signals. The Metabochip array has been utilised for this purpose, as has data from the 1000 Genomes Project and re-sequencing.

The Metabochip array has been utilised by Deloukas et al. (2013) to perform an association analysis to identify loci association with metabolic traits and fine-map those loci previously identified as associated with CAD. In this work, Deloukas et al. (2013) replicated the association of the 9p21 locus with CAD. However, in this work, the SNP rs3217992 was identified ($p = 7.75 \times 10^{-57}$). This SNP, located within the 3’UTR of CDKN2B and intron 1 of ANRIL, has an $r^2$ of 0.50 with rs1333049. This finding suggests causal variants may be in high LD with rs3217992, and indicates that examination of variants in this region may be warranted.

Coronary artery calcification is a marker of atherosclerosis in the coronary arteries, and was the subject of a Metabochip-based association analysis by Pechlivanis et al. (2013) in a European cohort. In this work, the 9p21 SNP rs1537373 remained associated with coronary artery calcification after adjustment and conditional analysis ($p = 0.001$). This SNP is in high LD with rs10757278 ($r^2 = 0.93$), replicating previous association results. Further, fine mapping of 9p21 using specially-selected SNPs on the Metabochip identified a number of variants with strong associations ($p \leq 1.0 \times 10^{-5}$). As this work has replicated the association with the same linkage block as that identified previously by GWAS, the locus has not been refined. However, fine mapping of the region has identified variants with strong associations with coronary artery calcification. Some of these variants may be functional; further bioinformatic analysis is needed to identify any potentially causal variants.

A meta-analysis of GWAS data using the 1000 Genomes Projects SNPs for imputation has replicated the association of the 9p21 locus with CAD (Nikpay et al. 2015). The lead SNP from this analysis was rs2891168 ($p = 2.29 \times 10^{-98}$). This variant is in high LD with rs1333049 ($r^2 = 0.967$), indicating that the causal variants at 9p21 are in high LD with rs1333049 and rs2891168. Further, false discovery rate variants were also
reported, with 5 reported in 9p21. Of these, 1 is located within the intergenic region upstream of IFNA21, with the rest in high LD with the GWAS lead SNP rs1333049.

Analysis by Shea et al. (2011) of the 9p21 region using targeted resequencing, genotyping and 1000 Genomes Project data imputation identified 145 variants associated with MI at 9p21 with \( p \) values of <0.05. None of these variants had a stronger association with MI than the GWAS lead SNP rs4977574. Annotation information was provided in this publication for the variants identified, with the majority located within ANRIL. 1 variant was annotated as being located within the 3' UTR of CDKN2B, whilst 16 variants had no annotation provided. 1 variant, rs1412830, was predicted to be in a FOX03 or FOX04 transcription factor binding site. rs1412830 has an \( r^2 \) value of 0.21 with rs4977574 and was only just significant (\( p = 0.042 \)). This work provides an extensive list of variants associated with MI that require further analysis in order to identify functional 9p21 variants. However, as no stronger associations with MI were detected than those already identified by GWAS, this suggests that the causal variant(s) are in high to moderate LD with the GWAS lead SNPs. However, further work using the approach taken by Shea et al. (2011) with a larger cohort of individuals may help to refine the 9p21 locus through greater power to detect disease associations with low-frequency variants.

Johnson et al. (2013) performed re-sequencing of 9p21 in a total of 281 individuals. Of these, 94 had a history of MI, 94 had high levels of coronary artery calcification, and 93 were controls (no history of MI and normal coronary artery calcium content). A total of 250,000bp was Sanger sequenced, encompassing the CDKN2A, CDKN2B and ANRIL genes. Initial analysis of the sequence data revealed a total of 1418 SNPs and complex variants, of which 1011 were novel. The novel variants identified typically had low minor allele frequencies. 140 SNPs were chosen for follow-up genotyping in over 7000 individuals based upon association analysis from the initial data. 126 SNPs were genotyped successfully in this cohort; genotype data was used to assess associations with multiple risk factors and indicators of cardiovascular disease. The top variants associated with prevalent MI (rs4977574), coronary artery calcification (rs1333045) and abdominal aortic diameter (rs1333046) are the only associations to remain statistically significant following multiple-test correction, with \( p \)-values of 6.2x10^{-4} or below. These top variants have either been associated with CAD by GWAS (rs4977574) or are in high LD with them (\( r^2 = 0.76 \) or above). This work replicates the 9p21 association with CAD, but failed to find any variants that have functional effects upon protein coding.
genes. However, a number of variants were found (both common and rare) that associate with expression of ANRIL and/or CDKN2B in the cohort studied.

Multiple fine mapping approaches have been used in an attempt to refine the association of the 9p21 locus with CAD, including Metabochip genotyping, 1000 Genomes Project data SNP genotype imputation and locus sequencing. These approaches have identified a number of novel variants, but the majority of these are in high LD with the SNPs identified by GWAS. Further, no variants have been found that are more strongly associated with CAD than the GWAS lead SNPs.

The study presented herein, in combination with the potential future work described above, may yet reveal novel roles for rs10811656 and rs10757278 in genetic risk of CAD at 9p21. However, it is also important to consider alternative variants within the region, which may be identified bioinformatically or by fine mapping, and investigated in a similar manner to rs10811656 and rs10757278. Also, ChIP-seq analyses for additional transcription factors and regulatory marks at the 9p21 locus is also an important approach to take. This methodology may produce information regarding additional regulatory regions that are present at the 9p21 locus. It may also yield new insights into the functions and mechanisms controlled by the known regulatory regions at 9p21 (such as ECAD1-9, identified by Harismendy et al. 2011), and may help to direct future studies by identifying possibly causal variants in additional, differentially-regulated, transcription factor binding sites. Genome editing approaches, which were used successfully to investigate specific SNPs in this study, are likely to play an important role in the future examination of the 9p21 locus.
Chapter 6: Discussion

6.1 Summary of Key Findings
This thesis has focussed on the investigation of three principle aims. The first of these was the establishment of the rAAV genome editing technique as a method by which disease-associated variants can be examined directly. Once developed, this technique was used to generate isogenic cell lines that were identical, except for the genotype of putative functional SNPs at the 1p13 (rs12740374) and 9p21 (rs10811656/rs10757278) loci, which have been associated with CAD by GWAS. After generation of these cell lines, functional analyses to investigate causality were performed; ChIP-qPCR was used to assess transcription factor binding, whilst qPCR was used to measure gene expression, comparing between genotypes. Our findings support the hypothesis that rs12740374 is a causal variant at the 1p13 locus. However, whilst capable of influencing transcription factor binding, our gene expression data suggests that the functional SNPs and causal mechanisms are yet to be identified for the 9p21 locus. The ability to use WJ-MSCs for rAAV genome editing was also assessed, as an additional aim of the project.

6.1.1 Generation of Isogenic Cell Lines by rAAV-Mediated Genome Editing
In this work, rAAV-mediated genome editing was used to investigate SNP genotype-driven effects upon local gene expression within the 1p13 and 9p21 CAD-associated loci. When this project was initiated, genome editing had not been utilised to investigate SNPs or non-coding regions. However, since this project began, genome editing has been used to investigate SNPs and regulatory regions. Beaudoin et al. (2015) edited a SNP on chromosome 6p24 (rs9349379) using a CRISPR approach, and showed that the binding of the transcription factor MEF2 is affected, leading to altered expression of the PHACTR gene. Further, Yoshimi et al. (2014) demonstrated the ability to use CRISPRs to alter the genotype of a SNP in rats that causes albinism; edited rats exhibited albinism as expected.

In this work, SNPs in the 1p13 (rs12740374) and 9p21 (rs10811656 and rs10757278) loci were examined. To do this, isogenic cell lines were successfully generated that differ only at the SNPs of interest, allowing comparisons to be made between genotypes, as background variation remains constant.
HepG2 cells were targeted for the 1p13 project, creating rs12740374 heterozygous lines from homozygous risk parental cells. HeLa-S3 and HCT116 cells were used for the 9p21 project, generating heterozygous lines of both cell types from rs10811656/rs10757278 homozygous non-risk parental cells. In addition, homozygous risk HCT116 cells were also generated.

WJ-MSCs were also targeted at the 1p13 and 9p21 SNPs. Isogenic cell lines could not be derived due to replicative senescence – WJ-MSCs are unsuitable for rAAV genome editing.

6.1.2 Functional Examination of the 1p13 SNP rs12740374 Using Genome-Edited Cell Lines
Musunuru et al. (2010) provided strong evidence to suggest that the causal SNP at the 1p13 locus is rs12740374 (discussed in chapter 4). In this publication, the authors present evidence to support the hypothesis that the genotype of rs12740374 influences the binding of a liver-expressed transcription factor, C/EBPα. Further evidence suggests that C/EBPα binding regulates the SORT1 gene, which encodes the sortilin protein. Sortilin was demonstrated by this group to increase LDL uptake in both cultured cells and mouse models. Musunuru et al. (2010) therefore hypothesised that the risk-associated allele of rs12740374, G, reduces C/EBPα binding, which in turn leads to decreased SORT1 expression and a decrease in LDL uptake. However, whilst this hypothesis was supported by experimental evidence, no data was presented to conclusively show that the genotype of rs12740374 directly influences SORT1 expression.

In order to identify if the causal SNP at the 1p13 locus is rs12740374, and to compare the risk (G) and non-risk (T) alleles of this SNP for differential C/EBPα binding and SORT1 expression, rAAV-mediated genome editing was utilised using HepG2 cells. HepG2 hepatocellular carcinoma cells were used for the 1p13 locus as a liver-expressed transcription factor, and associated pathway, were being investigated. These cells were chosen because HepG2 extracts were used by Musunuru et al. (2010) to investigate C/EBPα binding to the SNP of interest during EMSA experiments. Further, HepG2 cells have been shown to be suitable as models for investigating apoB secretion (Meex et al. 2010) and LDL uptake (Jackson et al. 2006), and are therefore a suitable hepatocyte model.
HepG2 cells, homozygous for the risk allele of rs12740374, were targeted with a rAAV vector containing the non-risk allele of rs12740374. This enabled the successful generation of multiple rs12740374 heterozygous cell lines. These isogenic cell lines, which differ only at rs12740374, were used to investigate the functional effects of this SNP.

ChIP analysis was used to quantify C/EBPα binding at the rs12740374 binding site, comparing between genotypes. Cell lines heterozygous for rs12740374 displayed a ~15-fold increase in binding, compared to homozygous risk cells. This shows that the genotype of rs12740374 has a profound effect upon C/EBPα binding at this locus.

Analysis of 1p13 gene expression identified four genes that show statistically-significant differences in expression between the two genotype groups. In rs12740374 heterozygotes, expression of *SORT1, CELSR2, PSRC1,* and *MYBPHL* was increased compared to the homozygous risk cells. *SORT1* expression was increased by 41%, whilst *CELSR2* expression was 33% higher. *PSRC1* was up-regulated by 24%, with a similar effect observed for *MYBPHL* (31%).

The results obtained from this work investigating the 1p13 SNP rs12740374 support the hypothesis proposed by Musunuru *et al.* (2010); rs12740374 is a functional SNP whose genotype is capable of influencing the binding of C/EBPα and, in turn, local gene expression.

Evidence provided by Musunuru *et al.* (2010) suggested that binding of C/EBPα is reduced in major (G) haplotype cells compared with heterozygous cells; this finding was confirmed in this work. This group also presented data of over-expression and repression of C/EBPα and concomitant alterations in *SORT1* expression that would suggest that C/EBPα binding affinity for rs12740374 influences *SORT1* expression. Musunuru *et al.* (2010) did not, however, show definitively that rs12740374 genotype directly affects *SORT1* expression – it is possible that the effect observed by Musunuru *et al.* (2010) was mediated by alternative, functional C/EBPα binding sites, and that rs12740374 genotype, whilst capable of causing differential C/EBPα binding, was not regulating *SORT1*.

The approach taken in this work was able to ascertain a direct functional link between rs12740374 genotype and local gene expression because the only variant to differ between the cell lines is the SNP of interest. This confirms that rs12740374 genotype regulates *SORT1*, as well as *CELSR2, PSRC1* and *MYBPHL*. 

197
Confirmation of this mechanism has several implications. Firstly, it questions the findings of other groups that have found that increased \( \text{SORT1} \) expression exacerbates phenotypes associated with contributing to CAD. Kjolby \textit{et al.} (2010) and Gustafsen \textit{et al.} (2014) have reported that VLDL and PCSK9 secretion are both positively regulated by sortilin, respectively. Both of these phenotypes are associated with increased risk of disease. However, these experiments required over-expression or knock-out of \( \text{SORT1} \), neither of which is likely to truly reflect protein function in humans. In the experiments presented in this thesis, the expression difference observed for \( \text{SORT1} \) is dependent on rs12740374 genotype and is smaller than previously reported by other groups comparing the entire haplotype (Musunuru \textit{et al.} 2010). This finding is congruent with this previous work because one would expect a small gene expression change when only a single SNP and transcription factor binding site are being investigated - it is highly likely that other mechanisms are involved in regulating gene expression in addition to the C/EBP\( \alpha \) binding site in which rs12740374 is located. Thus, these mechanisms, such as additional transcription factor binding sites and epigenetic modifications, may also be affected by other variants in the locus. Determining which variants and the regulatory mechanisms affected by them in the 1p13 region would provide further insight into how genetic polymorphisms in this locus regulate gene expression, and potentially confirm the role of \( \text{SORT1} \) (or other genes in the locus) in CAD.

Whilst much of the focus of functional research into the 1p13 region and CAD has been on the \( \text{SORT1} \) gene, other genes in the locus should also be considered as possible candidates. As rs12740374 genotype affects \( \text{CELSR2} \), \( \text{PSRC1} \) and, unexpectedly, \( \text{MYBPHL} \), this indicates that C/EBP\( \alpha \) is involved in regulation of these genes, and that their reduced expression in cells homozygous for the risk allele of rs12740374 may also be contributing to CAD risk. Identifying the role of these three proteins in CAD, and defining the variants and signalling pathways affecting their expression, may therefore be informative. Musunuru \textit{et al.} (2010) showed that PSRC1 does not affect total cholesterol or LDL-C by over-expressing \( \text{PSRC1} \) in mouse liver. However, \( \text{PSRC1} \), \( \text{CELSR2} \) and \( \text{MYBPHL} \) may contribute to CAD through other mechanisms.

Further studies investigating the 1p13 locus should be performed (discussed in detail in section 4.3). These include assessment of genotype-dependent effects upon lipid metabolism and genetic studies to examine the 1p13 locus in greater detail. For example, identification of additional regulatory elements and SNPs, determination of
whether the regulatory mechanisms act in *cis* or *trans* (e.g. allele-specific expression analysis), and characterisation of chromatin conformation are all approaches that will provide insight into the contribution of 1p13 variation to CAD. As discussed previously, the functions of *CELSR2, PSRC1* and *MYBPHL* should also be investigated in order to establish if these genes are involved in CAD pathogenesis.

This work demonstrates that genome edited cell lines can be successfully used to investigate the functional effects of individual SNPs. Further, these findings demonstrate that individual SNPs can have subtle effects upon gene expression, and that these effects can be observed and quantified using genome-edited cell models.

### 6.1.3 Functional Examination of the 9p21 SNPs rs10811656 and rs10757278 Using Genome-Edited Cell Lines

Harismendy *et al.* (2011) present data to suggest that the 9p21 SNPs rs10811656 and rs10757278 affect the response to interferon-γ stimulation of local gene regulation by affecting STAT1 transcription factor binding, subsequently influencing local gene expression. In this study, Harismendy *et al.* (2011) suggest that the risk-associated alleles of these SNPs reduce STAT1 binding, thereby altering expression of local genes, such as *CDKN2B, ANRIL*, and the more distant interferon-α genes.

In order to determine if these are the causal SNPs at this locus, and to discern if there is a genotype-driven effect upon STAT1 binding and local gene expression, we utilised rAAV-mediated genome editing to target these SNPs in HeLa-S3 and HCT116 cells. Harismendy *et al.* (2011) investigated these variants using HeLa cells, and so HeLa-S3 cells were chosen for use in this study. HCT116 cells were also used to provide another cell type with which to examine the 9p21 locus, and to provide a control for genome editing. HeLa-S3 and HCT116 are both homozygous for the non-risk alleles of rs10811656 and rs10757278. Targeting with a vector containing the risk alleles enabled the generation of heterozygous HeLa-S3 and HCT116 cells. Re-targeting of heterozygous HCT116 cells produced homozygous risk lines.

The cell lines generated by genome editing were used for functional analysis of STAT1 binding by ChIP-qPCR following stimulation with IFN-γ. We observed a decrease in STAT1 binding in heterozygous cells compared to homozygous non-risk lines. Heterozygous HeLa-S3 cells displayed a 60% reduction in STAT1 binding, whilst heterozygous HCT116 cells exhibited a 40% decrease. Similarly, homozygous risk
HCT116 cells showed a 40% decrease in STAT1 binding. For all lines, no STAT1 binding could be detected in the absence of IFN-γ.

Gene expression analysis following IFN-γ treatment revealed no genotype-specific differences after correction for interferon-γ stimulation by normalisation to the interferon-γ-responsive gene IRF-1.

The data presented in this work regarding the functional analysis of the 9p21 region both confirm and disagree with the work of Harismendy et al. (2011). The work presented herein confirms that IFN-γ stimulation (using the same treatment conditions as Harismendy et al. 2011) induces STAT1 binding to the rs10811656/rs10757278 TFBS. This interaction was reduced in heterozygous (HeLa-S3 and HCT116) and homozygous risk (HCT116) cells compared to homozygous non-risk cells. However, binding was not abolished in homozygous risk cells, suggesting that the non-risk alleles are not essential for binding.

Harismendy et al. (2011) reported increased ANRIL expression in HeLa-S3 cells by ~2-fold, whilst CDKN2B decreased slightly. In this work, ANRIL increased (up to ~30%) with no change in expression of CDKN2B, despite the use of the same concentration, duration and assays as those used by Harismendy et al. (2011).

From their findings, Harismendy et al. (2011) suggest that the risk genotype of rs10811656/rs10757278 reduces STAT1 binding at the site, which results in altered gene expression of the locus, including potentially the IFN-α genes. No genotype-specific effect was observed when controlling for IFN-γ treatment conditions. This conclusion was also reached by Almontashiri et al. (2013) using LCLs. However, this group reported that IFN-γ stimulation had no effect upon CDKN2A expression and caused a small increase in CDKN2B, as measured by Northern blot. This is in disagreement with the data presented herein, where an increase in CDKN2A and no change in CDKN2B was observed. Therefore, these results are not concordant with previous work by Harismendy et al. (2011) and Almontashiri et al. (2013).

The method employed here utilised isogenic cell models derived from genome editing to investigate two SNPs in isolation of other variants. This approach differs from that taken by Harismendy et al. (2011) and Almontashiri et al. (2013), potentially explaining the discrepant findings.

It should be noted that the findings by Harismendy et al. (2011), which show that HUVEC and HeLa cells express ANRIL and CDKN2B to different levels in response to IFN-γ, suggests a cell type-specific effect. Whilst this is likely the case (it is the case in
this work using HeLa-S3 and HCT116 cells), it is also possible that this is caused by a genotype difference between the two cell lines used. Heterozygous HUVEC cells were compared to HeLa cells; in this work, direct Sanger sequencing of the rs10811656 and rs10757278 SNPs and SNP genotyping microarray both confirm that HeLa-S3 cells (derived from HeLa cells) are homozygous non-risk for these two SNPs. Further, Sanger sequencing of these SNPs in HeLa cells also showed that they too are homozygous non-risk (data not shown). Harismendy et al. (2011) sequenced an adjacent region and genotyped other SNPs as tag/proxy SNPs for rs10811656 and rs10757278. No direct sequencing across these two SNPs was reported in their publication. Due to the African ancestry of HeLa cells, linkage disequilibrium between the SNPs of interest and the tag SNPs is low. This highlights the importance of, where possible, direct genotyping (e.g. by sequencing) of the polymorphisms of interest rather than using tag SNPs, in order to account for linkage disequilibrium. In cases where direct genotyping is not possible, genotyping of tag SNPs should be performed after considering linkage for the population from which the sample originates.

Our data suggests that these SNPs do not contribute to CAD through alteration of gene expression regulation by IFN-γ. However, consideration of study limitations should be made. Firstly, only a single signalling pathway was examined – other pathways may be involved, or potentially interact with IFN-γ signalling, in vivo to result in a genotype-specific effect that could not have been seen in this work. Further, cell models were utilised that are not derived from tissues involved in CAD. As such, pathways that could alter the functionality of these SNPs could exist in such tissues, but are not present in the cell lines used. An additional confounding issue is the investigation of just two SNPs at once. Whilst this approach would provide information regarding specific SNP allele functions, it could not replicate the interactions between multiple variants in the risk haplotype. Therefore, it is possible that additional SNPs contribute to CAD also, increasing the effect of rs10811656 and rs10757278 risk alleles. Such variants may be present in ANRIL and affect expression, splicing, stability or function, thereby affecting the genes controlled by ANRIL. A final point to consider is the potential for long range interactions – this STAT1 site may control expression of genes not investigated in this work that are situated at a greater distance from the enhancer or trans interactions with other chromosomes.

Thorough examination of the 9p21 locus is needed to fully understand the association with CAD. As discussed above and in section 5.3, additional genetic and
functional studies are required. These include further characterisation of the regulatory elements at 9p21, as well as the identification and assessment of other candidate variants – ideally, exploration of the interaction between variants and the (likely) multiple signalling pathways affected by them would be performed. The findings of such work should help to explain the complex effects of entire haplotypes upon gene expression regulation.

Our understanding of ANRIL is currently limited. Additional studies to identify more splice variants and the functions of this lncRNA are needed in order to fully appreciate the role ANRIL plays in gene regulation and, ultimately, in CAD.

In terms of rs10811656 and rs10757278, future studies should consider examining genotype-dependent effects upon proliferation and migration. Also, evidence presented recently by Almontashiri et al. (2015) suggests other transcription factors (TEAD3 and TEAD4) may bind to the rs10811656/rs10757278 site and regulate CDKN2A and CDKN2B – using the edited cell lines produced in this work, this mechanism could be fully explored.

The findings of this study illustrate the complex regulatory nature of the 9p21 locus. We have shown that the genotype of rs10811656 and rs10757278 is capable of affecting STAT1 binding following IFN-γ exposure, but this does not affect gene expression (when correcting for treatment conditions). This observation, made possible by genome editing, indicates that additional (currently unknown) variants and mechanisms in the locus are involved in CAD. It is possible that the effects of multiple variants together elicit functional effects that contribute to disease.

6.2 Challenges of rAAV-Mediated Genome Editing and In Vitro Disease Modelling

rAAV genome editing is a powerful technology, capable of making highly precise alterations to the genome. However, as with most technologies, rAAV editing poses some technical challenges that must be overcome. The first of these is low efficiency of successful targeting. HepG2 cells, used in the 1p13 project, displayed a maximum targeting efficiency of 0.22%. Although this was sufficient to generate 17 clonal lines from multiple integrations, several rounds of targeting and iterative optimisation were required in order to establish a protocol for 1p13 targeting of HepG2 cells.

With respect to 9p21 targeting, HeLa-S3 cells exhibited less efficiency for the procedure than HCT116 (0.2-0.5% for HeLa-S3, compared to up to 3.15% for the first targeting event of HCT116 cells). HCT116 cells showed 0.39% efficiency for the
second targeting event (for the generation of homozygous risk cells from heterozygous lines). Whilst this reduction in efficiency for the second targeting attempt may be due to reduced cell survivability caused by previous targeting, this does not appear to be the case. This is because 6.28% of the samples screened during the generation of DKI HCT116 lines were diagnostic PCR positive. If the on-target integration event was random, a 50% double KI efficiency would be expected (if this were the case, 3.14% of the total number of samples screened would be DKI). As this is not the case, it would suggest that homologous recombination of the vector in previously-targeted cells favours the edited chromosome. Greater homology is shared between the vector and the targeted chromosome than between the vector and non-targeted DNA, although this is only small (two SNPs and a single loxP site).

Evidence from other members of the group that have subsequently utilised this technology suggests that optimisation of the procedure is required for each locus, as well as each cell line. The group has observed greater efficiency when targeting coding regions than regulatory regions (such as those in this work), despite utilising the same cell lines. Because of this observation, it is likely that locus-specific and cell line-specific effects result in the requirement for extensive optimisation for each targeting experiment.

As homologous recombination and subsequent identification of targeted cells occurs at very low rates, methods to improve the efficiency of the technique are needed. Recently, Kaulich et al. (2015) showed that the rate of HR could be increased 10-fold by combining CRISPR and rAAV technologies. Further, development of RNA-guided FokI nucleases (RFNs) by Tsai et al. (2014) reduces off-target cleavage by utilising two guide RNAs to induce cleavage only when they both bind to the genome with the correct spacing and orientation. As such recognition sites are expected to occur at a lower frequency than those contained within a single guide RNA, off-target effects are reduced. By combining these techniques, it is possible that the frequency of rAAV vector HR could be increased significantly.

Another challenge when utilising rAAV genome editing is the consideration of selection cassette placement. For the studies discussed in this work, the position of the selection cassette and loxP cloning scar could be designed so as not to influence annotated coding or regulatory sequences. However, in other studies, it is likely that this requirement limits the utility of rAAV genome editing. It would therefore be beneficial if vectors could be designed to not leave a cloning scar behind after selection cassette
removal. The PiggyBac system, discussed in section 3.3, offers this possibility as no cloning scar remains after excision of the DNA between the PiggyBac ITRs.

In this study, regulatory SNPs have been investigated in isolation of all other variants in the CAD-associated linkage blocks of interest. Whilst this approach does enable the direct elucidation of the functions of the individual variants of interest, the information provided is also limited. This is because we cannot examine the interaction and function of multiple variants within the locus at once. For example, a disease-associated haplotype may contain several functional SNPs that interact. These interactions may be complex and affect gene product function and expression in response to multiple signalling pathways.

In order to investigate a locus in detail, an allelic series would be required. In such a study, the genotype of every putative SNP is altered individually, and in many combinations, so as to identify the truly functional SNPs, their functions and the interactions between them. Such an approach utilising rAAV genome editing would require extensive time and financial investment; such studies are beyond our current technological capabilities.

*In vitro* disease modelling using rAAV genome-edited cell lines can also be challenging. This is because of the limitations with the cell types that can be used for the technique. Cancer cells and some stem cells are routinely used because of their proliferative capabilities, but primary cells cannot be utilised, despite being the ideal model, due to replicative senescence. The use of cancer cells may limit the information that can be obtained from some studies because of the inherent “abnormal” phenotype of cancer cells. Highly specialised cellular processes that may be of interest may not be functional to the same level, if at all, in cancer cells, limiting the utility of these cells and this technique for some projects. In addition, some lines display genomic instability in long-term culture (e.g. EA.hy926) and are therefore unsuitable. This highlights the importance of parental cell line characterisation.

Stem cells, such as iPSCs and HUES, offer a means by which differentiated cell types can be obtained after rAAV targeting of stem cells. iPSCs and HUES cells have been shown to be capable of long-term culture (Shimamoto et al, 2014 and Amit et al. 2000, respectively). However, stem cell culture can be difficult, and differentiation protocols may not induce differentiation of all cells in a population, meaning cell sorting is required. In addition, stem cell differentiation produces cells that resemble,
but may not be identical to, the primary cell type of interest. Depending upon the situation and the pathway(s) to be investigated, this may or may not be satisfactory.

In this work, it has been shown that mesenchymal stem cells, derived from Wharton’s jelly, are unsuitable for rAAV genome editing. These cells reached replicative senescence following targeting of two different loci in independent donor lines, and so are unlikely to be suitable for targeting of other loci.

The development of conditionally reprogrammed cells (CRCs) by Chapman et al. (2010) is a possible method by which the limited growth potential of primary cells can be circumvented. This method, which elicits a stem-like state in treated epithelial cells (Suprynowicz et al. 2012), requires treatment of cells with feeder cell (irradiated MEFs) conditioned media and Y-27632 (a Rho kinase inhibitor). Treatment with these two components stimulates cell division indefinitely, until the two components are removed. At this point, the cells return to a normal state. Whilst so far only created using epithelial cells, CRCs derived from other cell types may be possible and may be useful for functional studies.

6.3 Genome Editing: Technology Considerations & Learnings from this Work
Currently, genome editing can be achieved using multiple technologies, including ZFN or TALENs, CRISPRs and rAAV. Depending upon the experimental approach to be taken, the appropriate technology to use can vary. In cases where a gene knock-out is desired, ZFNs, TALENs and CRISPRs are more likely to result in the highest efficiencies due to the production of indel mutations through repair of DSBs by NHEJ. rAAV would not be highly efficient, due to the reliance upon HR which occurs at a much lower frequency than NHEJ. However, when making specific changes such as SNP alteration as performed in this work, or insertion of large amounts of DNA, homologous recombination is required. The efficiency of this repair pathway is very low irrespective of the approach taken. This means that NHEJ at target loci often occurs with ZFN, TALEN or CRISPR approaches at a much greater frequency than HR with a repair template, and in cases where HR does occur on one chromosome, the other can be affected by a NHEJ-induced indel mutation. The rAAV technique may be preferable in such cases, as although the same low efficiency of HR is observed, rAAV does not induce NHEJ repair, so indels do not occur. This means that editing of a single chromosome is possible, or two identical alterations can be made through additional targeting and screening.
Further, the rAAV approach lends itself to significantly reduced off-target effects due to the long homology arms utilised. Off-target effects with ZFNs or TALENs can occur because although a DSB is only induced upon the binding of two proteins (one on either DNA strand), the zinc finger or TALE domains do not require complete identity for binding and subsequent cleavage, allowing off-target effects due to similar sites to the target being present elsewhere in the genome. CRISPRs only utilise ~20bp in the gRNA for target recognition and require a short PAM sequence. It is therefore possible that the same or similar sequences exist at other loci in the genome which may also be cleaved. Because of this, it is possible that off-target effects with CRISPRs occur at a greater frequency than that seen with ZFNs or TALENs. Off-target effects can be predicted bioinformatically and ZFN/TALENs/CRISPRs designed so as to minimise non-specific cleavage. However, it is difficult to be certain that NHEJ has not occurred in loci not predicted bioinformatically following off-target cleavage.

There are some limitations regarding the loci that can be targeted for some technologies that may affect experimental strategies that can be used. CRISPRs require a PAM sequence to be located adjacent to the recognition sequence. If this is absent, cleavage by the Cas protein will not occur. This issue may, in most cases, be circumvented by changing the Cas protein used as each protein uses a different PAM. With rAAV technology, positioning of the vector due to selection cassette constraints, as well as subsequent removal, can reduce applicability of this technique. Careful design of homology arm placement is required, and may not be possible for some loci and editing approaches.

Another aspect of each technology that affects their usability is the ease with which targeting vectors can be produced. CRISPR expression vectors can be made by cloning the guide sequence into a vector; this is the simplest and least expensive genome editing vector to assemble. This permits the rapid generation of multiple vectors for targeting of loci of interest, enabling determination of optimal guide sequences easily. In contrast, TALEN and ZFN plasmids require cloning of individual nucleotide binding domains sequentially into the vector. Due to the repetitive and modular nature of TALENs, assembly has been simplified through the use of techniques such as FLASH (Reyon et al. 2012), which allows the cloning of several binding domains per ligation reaction. Because multiple rounds of cloning are required to produce a TALEN plasmid, and two are needed for each targeting experiment, the assembly complexity compared to CRISPRs is greater.
rAAV vectors produced by Horizon Discovery Group (UK) are not made using cloning techniques – due to the size and complexity of the vectors, rAAV production plasmids are manufactured using DNA synthesis. Following this, the plasmid is transformed into *E. coli* cells for purification of large quantities of plasmid. Virus particle assembly and purification is required in order for genome editing to be performed. The time taken from vector design to completion of virus particle purification is greater than the time and cost needed for cloning of CRISPR or TALEN/ZFN plasmids. However, for the majority of rAAV vector designs only a single vector is needed for sufficient rates of HR at the target locus – with CRISPRs and ZFN/TALENs, multiple designs may be necessary before the optimal vector is identified.

All genome editing technologies have their own merits and limitations with respect to efficiency, accuracy (frequency of off-target effects), vector production (complexity, production time and reagent cost) and target locus limitations (CRISPR PAM site selection and rAAV homology arm/selection cassette placement). The approach chosen for any genome editing study should ideally consider all of these factors in conjunction with the chosen edit(s) to be made in order to select the most appropriate technology to use.

Another important aspect of genome editing is the cell system used. Immortalised cell lines such as cancer cell lines (as used in this work), offer a relatively simple model that benefit from being (generally) easy to culture with inexpensive reagents. However, they have limited applicability due to their atypical growth behaviours and extensive genomic differences compared to primary cells. Further, no cancer cell lines are available for cardiovascular cell types (although immortalised cell lines have been developed). The use of immortalised cells (cancer cells or otherwise) is amenable for use with phenotypes that do not involve proliferation differences. However, if alterations in proliferation are to be investigated, it may be difficult to assess this in a cell line that already proliferates rapidly and continually.

Human iPSCs are an alternative to immortalised cell lines. Whilst immortal themselves, iPSCs are capable of differentiation into multiple lineages following genome editing. Because of this, iPSCs are capable of providing several cell types for analysis following a single genome editing event. Further, the differentiated cells are functionally similar to their primary equivalents, providing a highly applicable model cell type in which to explore the effects of genome editing.
From this work, we have developed a greater understanding of rAAV genome editing in terms of cell line characterisation, virus design, production and quantification, as well as ways in which to maximise efficiency in terms of plating density following both selection and selection cassette removal.

The iterative approaches taken for the two loci and cell lines highlight the differences required for editing different loci and cell types. This work shows that the locus targeted, and the cell line used, both influence efficiency and the procedures needed to obtain sufficient homologous recombination frequencies. Despite this, the efficiencies obtained were low and required PCR screening of several thousand cell samples in total for identification of the targeted cell lines obtained in this work. The approaches taken in these studies that produced the greatest efficiencies may provide a useful guide for future projects utilising this technique, though it is possible experiment-specific modifications may be required.

Considering the findings of this work, the availability of additional cell models and the development of CRISPRs as a genome editing technology (which was not available at the commencement of this study), it is possible that the work performed in this thesis could have been carried out using an alternative approach.

In order to properly examine the effects of alterations made by genome editing, a model cell system that is as similar as possible to primary cells is necessary so as to yield data that is representative of the effects expected in vivo. iPSCs are a suitable model, and would, through differentiation, produce cells suitable for the examination of the 1p13 SNP rs12740374 (hepatocytes) and the 9p21 SNPs rs10811656/rs10757278 (for example, vascular endothelial and smooth muscle cells). This is the greatest benefit of using iPSCs rather than immortalised cells. Additionally, by using iPSCs, optimisation of genome editing for only a single cell line would be necessary, and therefore should increase the rate at which edited cell lines can be made for the two projects. Similarly, parental cell line characterisation would be simpler as only one cell line would need to be examined for applicability to both the 1p13 and 9p21 projects.

Cell functions applicable to hepatocytes may be examined using 1p13 edited cells, whilst as it is unknown what the effects of the 9p21 SNPs are upon cell function, the ability to examine multiple cell types from a single targeting event may help to discern the functional effects more easily.

It is important to consider that the differentiation protocol(s) used should be optimised prior to, or alongside targeting, in order to ensure that the differentiation can
be performed and to assess the efficiency of it. This is important so as to ensure that optimal conditions are used for differentiating edited iPSCs to ensure enough differentiated cells are produced for subsequent experiments.

In order to obtain high rates of targeting, rAAV technology may be combined with CRISPRs in order to promote homologous recombination repair through induction of DSBs at the target locus. This has been demonstrated to improve efficiency up to approximately 10-fold by Kaulich et al. (2015). Careful design and thorough examination for off-target effects should minimise off-target effects induced by CRISPR-specific NHEJ repair of DSBs. However, the site targeted by the CRISPR should be considered, and, to prevent cleavage of the rAAV vector, should be positioned across the boundary between the left and right homology arms. This will ensure that genomic DNA is targeted by Cas9 for cleavage, whilst the vector DNA is not due to the presence of the selection cassette.

As the rate of HR is increased by the combination of CRISPRs and rAAV, it is likely that the generation of double knock-in (i.e. homozygous editing) is increased from a single targeting event. Further, if a double knock-in is not achieved through a single round of targeting, selection and screening, repeated targeting prior to removal of the selection cassette from the targeted chromosome is necessary when using CRISPR combined with rAAV. This ensures that the correctly targeted chromosome is not cleaved by the CRISPR as the guide RNA is prevented from binding due to the presence of the selection cassette. However, the non-targeted chromosome is cleaved and is subject to increased rates of homologous recombination as a result. The identification of double knock-in cells can be achieved through screening (the parental/wild-type PCR product will not be produced from DNA isolated from single colonies of targeted cells). The selection cassettes from both targeted chromosomes in double KI cells can then be removed simultaneously via Cre recombination.

Evidence indicates that the combined use of CRISPR and rAAV would increase the efficiency of single and double KI. Further, the time needed to produce the cell lines should be reduced, due to the need for a single Cre recombination for removal of the selection cassettes from DKI cells produced from two successive rounds of targeting.

In summary, if the two projects performed in this work were to be repeated using our current knowledge, a CRISPR/rAAV dual approach may be taken to improve efficiency. Targeting of iPSCs would provide a cell model suitable for both 1p13 and 9p21 projects through differentiation into appropriate lineages. It is expected that this
new approach would improve efficiency of targeting, allow the generation of double
knock-in lines more easily, and thereby increase the overall speed at which cell lines
can be produced. The differentiation of targeted iPSCs would allow for a thorough
examination of cell function following alteration of the 1p13 or 9p21 SNPs of interest.
This approach is also applicable to a wide variety of loci in cardioavascular disease
research (such as additional CAD GWAS regions) as well as other fields of genetic
research.

6.4 Fine Mapping of GWAS Loci and Functional Study Design

Functional analysis of disease-associated loci following GWAS is required in order to
identify the causal variant(s) in the locus and also to characterise the genes they effect,
and in turn, the pathways they influence. Selecting a locus and variants to examine can
be difficult, as genetic loci contain many variants that may be potentially functional.
Bioinformatic analysis can help to predict which variants are likely causal, but this
approach is limited due to the large number of variants in the region. In addition, only
common variants, with a relatively high minor allele frequency, can be examined, as
rare variants, with potentially large effects, remain unidentified.

Genome editing approaches may help to substantially improve our understanding of
the genetic variants that contribute to disease. These technologies enable the direct
analysis of individual variants in isolation, thereby allowing their functional effects to
be characterised. This can only be achieved, however, if causal variants are selected for
investigation.

Recently, fine-mapping approaches have been utilised in order to both refine the loci
associated with CAD and identify new variants. This is anticipated to ultimately provide
information as to which variants are potentially causal and thus require functional
examination. Two key approaches used for the fine mapping of CAD-associated loci are
exome sequencing and analysis of genotype association data following release of data
from the 1000 Genomes Project.

Exome sequencing analysis involves the sequencing of the exons of genes in either a
genome-wide or targeted manner. When performed in a large cohort of individuals, this
approach is expected to identify rare variants that may contribute to genetic risk of
disease. As such variants have a low minor allele frequency in the population,
associations with disease (such as CAD) may not be identified with GWAS due to
insufficient power to detect an association in such studies.
One recent example of successful utilisation of the exome sequencing approach was presented by Do et al. (2015). In their publication, rare variants were identified in two genes involved in lipid metabolism, \textit{APOA5} and \textit{LDLR}.

In \textit{APOA5}, 46 rare variants (frequency <1\%) were found, comprised of non-synonymous or splice site-altering single nucleotide variants, or indels. Carriers of a rare \textit{APOA5} mutation exhibited an increased risk of CAD or MI (by 2.2-fold) when compared to non-carriers. When the mutations were analysed by Do et al. (2015) using algorithms designed to predict the effects of genetic variants, mutations were defined as “deleterious” if they were predicted to be splice site, indels, nonsense or missense mutations. A total of 5 algorithms were used; mutations found to be “deleterious” by 1-4 of the algorithms were subdivided into a “deleterious (broad)” group, whilst those predicted to be deleterious by all 5 algorithms were classified as “deleterious (strict)”. Carriers of “deleterious (strict)” mutations had a 3.3-fold risk of CAD/MI compared to non-carriers.

Heritability and MI variance was calculated for the \textit{APOA5} mutations identified by Do et al. (2015). 0.14\% of total MI variance and 0.28\% heritability of MI/CAD can be explained by these rare mutations in \textit{APOA5}. The heritability estimate used by Do et al. (2015) assumes an additive model and a total genetic contribution to total variance of 50\%.

The effects of the rare mutations in \textit{APOA5} could be detected biochemically; carriers of rare, non-synonymous mutations had elevated plasma triglycerides and lower HDL than non-carriers ($p = 0.007$ for both parameters). The LDL cholesterol level in carriers and non-carriers, however, remained highly similar.

Do et al. (2015) also identified mutations in \textit{LDLR}; a total of 156 were found, of which 49\% have been previously identified and recorded in familial hypercholesterolaemia databases as mutations associated with this disease. Bioinformatic prediction of mutation effects was also performed for the mutations in \textit{LDLR}. When examining the “deleterious (strict)” set of mutations, carriers of these mutations exhibited a 4.2-fold increased risk of MI or CAD when compared to non-carriers. Carriers of “disruptive” mutations (defined as nonsense, splice site or indel mutations) had a 13-fold increased risk of disease compared to controls ($p = 9 \times 10^{-5}$).

In terms of heritability and MI total variance, \textit{LDLR} mutations have a larger contribution than \textit{APOA5} mutations. \textit{LDLR} rare variants were found to explain 0.48\% of the heritability and 0.24\% of the total variance of MI.
In contrast to APOA5 mutations, Do et al. (2015) found an effect of LDLR mutations upon LDL cholesterol levels. Carriers of non-synonymous and disruptive LDLR mutations were 4.25 times more likely to have LDL-C levels above 190mg/dl than non-carriers.

The work by Do et al. (2015) is important for many reasons. This study demonstrates the utility of the exome sequencing approach for detection of disease-associated rare variants. Also, Do et al. (2015) provide advice regarding sample sizes needed for adequate statistical power in detecting rare variants at exome-wide significance. Do et al. (2015) recommend sample sizes of at least 10,000 for 80% statistical power. This is an important consideration in future exome sequencing studies so as to ensure that data produced are accurate and capable of detecting as many rare variants as possible. Finally, this study shows that the rare variants found in just these two genes can explain up to 0.76% of the heritability of CAD/MI, suggesting that rare variants across the genome, with medium to large effects, may be an important contributor to genetic risk of CAD. Further large-scale exome (or whole genome) sequencing studies are needed to identify additional rare variants that, together, may contribute significantly to the heritable risk of CAD.

In terms of locus and variant selection for genome editing studies, this work shows that exome sequencing is capable of identifying novel rare variants that may be causal. Further, studies such as this help to provide information as to which variants should be investigated functionally by techniques such as genome editing. Bioinformatic analysis of the variants can help to identify the mutations with deleterious effects, thereby decreasing the number of variants to choose from when performing genome editing studies.

Data from the 1000 Genomes Project has also been instrumental in identifying new loci associated with CAD. Nikpay et al. (2015) performed a meta-analysis of GWAS data from approximately 185,000 CAD cases and controls. During this work, SNP genotype imputation was performed using data from the 1000 Genomes Project. In addition to being able to replicate the association with the majority of previously-identified GWAS loci, this approach was also able to identify 10 new loci associated with CAD. Of these, 8 were associated with CAD when using an additive model, and 2 were associated when using a recessive model. In these loci, candidate genes and/or ENCODE-annotated functional elements are present. This work, enabled by 1000 Genomes Project data, therefore helps to expand the list of loci associated with CAD.
that can be investigated using functional approaches, such as genome editing. The work by Nikpay et al. (2015) shows that the 1000 Genomes Project data can be used in conjunction with GWAS data to discover new disease associations.

1000 Genomes Project data has also been used by Brænne et al. (2015) as part of an extensive method to identify candidate causal variants and affected genes. In this approach, 159 SNPs identified by GWAS, and those in LD with the GWAS SNPs (imputed using 1000 Genomes Project data), were analysed bioinformatically for eQTL associations, predicted protein sequence changes and whether the SNP is in a regulatory region. Any SNPs identified by this analysis (and any associated genes) were taken forward for prioritisation according to either prior-knowledge (including evidence from previous publications, mouse phenotypes, and involvement in biochemical pathways) or data, such as eQTLs, knowledge of a protein-altering effect or being in a promoter region (where a regulatory effect is possible), as well as analysis of Bayesian networks of CAD-relevant pathways and data regarding phenotypes in human or mouse. Lists of genes and SNPs from both data-analysis pipelines were then compared to generate a list of candidate genes and variants. Using this approach, Brænne et al. (2015) identified 12 non-synonymous SNPs that are predicted to be deleterious which warrant further functional examination, such as by genome editing, to determine their effects. Further, 66 lead SNPs were associated with local gene expression, indicating a significant proportion of causal variants mediate their effects through augmentation of gene expression.

This extensive annotation methodology designed by Brænne et al. (2015) was successful for a number of the loci studied. For 32 loci, new information was found that may inform future studies, whilst for 15 loci, identical annotation to that previously made following GWAS was obtained. However, for 25 loci, there was no overlap at all between this method and previous annotations, and, for the remaining 87 loci, no functional data could be found or candidate genes defined.

By combining fine mapping approaches (e.g. exome sequencing and 1000 Genomes Project data), and bioinformatic assessment of predicted SNP functionality of variants, it is anticipated that causal variants may be predicted with greater specificity and confidence. This will inform and facilitate the design of successful future functional studies, including those based upon genome editing, for diseases such as CAD.
6.5 Concluding Remarks

GWAS provide valuable information regarding the identification of genetic loci that are associated with a particular phenotype. GWAS data, however, cannot provide information as to which variant(s) at the locus are causal. For a significant proportion of disease-associated loci, there are no obvious candidate SNPs (i.e. none that are located within exons and cause a non-synonymous change). For these loci, it is likely that the causal SNP(s) influence gene regulation, subsequently contributing to disease.

Bioinformatic analysis of putative regulatory SNPs can be performed to identify potential transcription factors affected by SNP genotype. In some cases, the gene controlled by the TF may be known, but often not. Genome editing by rAAV offers a method by which SNP alleles can be altered, keeping background variation constant and generating isogenic cell lines. These cell lines can be used for examining SNP genotype-driven functional effects.

Two coronary artery disease-associated loci (1p13 and 9p21) have been investigated in this work using rAAV genome editing. Both loci contain SNPs whose genotypes affect TF binding. These binding differences affected gene expression at 1p13, and provide evidence for a mechanism that contributes to CAD at this locus.

However, the absence of a genotype-specific effect upon IFN-γ 9p21 gene regulation for the SNPs involved indicates other, unknown, mechanisms are at work at the 9p21 locus.

This work shows that rAAV genome editing is a powerful and adaptable technique that promises to revolutionise the functional analysis of disease-associated loci in the post-GWAS era, either alone or in combination with additional genome editing approaches. The technique and methods used in this work are applicable for all diseases with a genetic component. The findings of this work are informative and provide an experimental approach upon which future studies investigating the genetics of complex disease, such as CAD, can be based.
Appendix: Conference Poster Presentation Abstracts


Presented at:
British Cardiovascular Society Annual Conference, Manchester (UK), 2nd-3rd July 2014

Abstract:
Coronary artery disease (CAD) is a leading cause of mortality worldwide and has a significant heritable component. In the last few years genome-wide association studies (GWAS) have identified 46 genome-wide significant loci associated with risk of CAD. A major post-GWAS challenge is the functional investigation of the biological and cellular mechanisms by which specific variants at CAD associated loci contribute to disease. The functional impact of putative functional variants is likely to be relatively subtle and could be masked by other genetic variation. We have therefore established genome editing techniques to create precise knock-in of single nucleotides in disease-relevant human cell lines, allowing the effects of a CAD associated variant of interest to be investigated on an isogenic genetic background.

Here, we present a proof-of-principal investigation of a CAD associated locus using isogenic human cell lines created using genome editing. Musunuru et al (Nature. 2010. vol 466 (7307). pp. 714–9) presented compelling evidence that rs12740374 is the causal variant at the 1p13 CAD locus. They show that the risk allele of rs12740374 disrupts the binding of the transcription factor CEBPα and consequently alters the cis transcriptional regulation of CELSR2, PSRC1 and SORT1 and identify the SORT1 protein as a novel mediator of LDL cholesterol metabolism. We have used recombinant adeno-associated virus (rAAV)-mediated genome editing to knock-in the non-risk allele of rs12740374 in the HepG2 (hepatic carcinoma) cell line. Using qPCR we show that rs12740374 genotype affects the expression of nearby genes, including SORT1. Further functional studies of the 1p13 locus using these isogenic cell lines are ongoing.

Our data shows that genome editing is a powerful functional genomics tool in the investigation of specific CAD-associated variants.

Presented at:
European Society of Human Genetics Conference, Glasgow (UK), 6th-9th July 2015

Introduction: Genome-wide association studies (GWAS) have identified multiple loci associated with coronary artery disease, the most significant of which is located on chromosome 9p21. Harismendy et al (Nature, 2011) suggest that the causal variants at this locus are two SNPs that disrupt STAT1 transcription factor binding to an enhancer element. This may therefore result in altered gene expression of nearby protein and RNA genes. Here, we investigated this hypothesis by utilising genome editing to produce isogenic cell lines that differ at only the two SNPs of interest. This allows any functional effects of these variants, in response to inflammatory signalling, to be determined.

Methods: Genome editing by recombinant adeno-associated virus (rAAV) was used to specifically alter the genotype of the two SNPs in HeLa-S3 cells, with all other variants remaining constant. Heterozygous lines were derived from the homozygous-protective parental line. Genotype-specific variation in STAT1 binding was assessed following interferon-γ treatment using chromatin immunoprecipitation and quantitative PCR (ChIP-qPCR). Local gene expression of treated cells was also investigated using qPCR.

Results: ChIP-qPCR revealed a reduction in STAT1 binding by approximately 50% in cells heterozygous for the risk alleles compared to the homozygous non-risk control. However, no difference in gene expression of CDKN2A, CDKN2B and several ANRIL transcripts was observed. Analysis of additional local genes is on-going.

Discussion: The observed reduction in STAT1 binding suggests that these SNPs are functional, but the absence of a gene expression difference indicates that other functional variants in the region may be required to contribute to disease.
Bibliography


Wellcome Trust Case Control Consortium. 2007. Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature*, 447(7145), 661-678.


