

UNIVERSITY COLLEGE LONDON  
UCL Institute for Liver & Digestive Health  
UCL Molecular Psychiatry Laboratory

---

---

# THE GENETICS OF ALCOHOL- RELATED LIVER DISEASE

---

---

*Michael J Way*

*Supervisors:*

*Dr Marsha Morgan (Primary)*

*Dr Andrew McQuillin (Secondary)*

2011-2016

I, Michael J Way, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated.

.....

**MICHAEL J WAY**

*“Nobody ever figures out what life is all about, and it doesn't matter. Explore the world.  
Nearly everything is really interesting if you go into it deeply enough.”*

**RICHARD P. FEYNMAN**

---

---

# TABLE OF CONTENTS

---

---

Table of Contents .....	I
Acknowledgements.....	V
Publications .....	VI
Table of Tables.....	IX
Table of Figures.....	XI
Abbreviations.....	XIV
Abstract .....	XVI
Contributions .....	XVII
CHAPTER 1 Introduction .....	1
1.1 - Overview.....	2
1.2 - Alcohol.....	2
1.2.1 - Beverage Alcohol.....	2
1.2.2 - Historical Background.....	3
1.2.3 - Alcohol Consumption.....	4
1.3 - Alcohol and Harm .....	10
1.3.1 - Alcohol Use/Misuse .....	14
1.3.2 - Alcohol-Related Liver Disease.....	20
1.4 - Alcohol and Genetics.....	35
1.4.1 - Alcohol Use/Misuse .....	35
1.4.2 - Alcohol-Related Liver Disease.....	42
1.5 - Aims Of Thesis .....	50
CHAPTER 2 Genome-wide Association Study of Alcohol-related Cirrhosis .....	51
2.1 - Overview.....	52
2.2 - Background .....	52
2.3 - Aim .....	53
2.4 - Materials and Methods.....	53
2.4.1 - Discovery Cohort .....	54
2.4.2 - Validation Cohort .....	56
2.4.3 - Discovery Analysis.....	58

2.4.4 - Validation Analysis.....	60
2.4.5 - Post-hoc Analysis .....	61
2.5 - Results .....	63
2.5.1 - Discovery Analysis.....	63
2.5.2 - Validation Analysis.....	67
2.5.3 - Post-hoc Analysis .....	69
2.6 - Discussion .....	72
CHAPTER 3 Extended Genome-wide Association Study.....	77
3.1 - Overview.....	78
3.2 - Background .....	78
3.3 - Aim .....	80
3.4 - Materials and Methods.....	80
3.4.1 - Cohorts.....	81
3.4.2 - Discovery Analysis.....	83
3.4.3 - Validation Analysis.....	86
3.5 - Results .....	87
3.5.1 - GWAS Dataset Processing.....	87
3.5.2 - Discovery Association Analysis.....	90
3.5.3 - Validation Genotyping.....	97
3.6 - Discussion .....	104
CHAPTER 4 Genetic Variation in <i>PNPLA3</i> and Alcohol-related Cirrhosis Risk .....	108
4.1 - Overview.....	109
4.2 - Background .....	109
4.2.1 - <i>PNPLA3</i> .....	109
4.2.2 - Genetic Variation in <i>PNPLA3</i> and Liver Disease.....	110
4.2.3 - <i>PNPLA3</i> in Alcohol-related Liver Disease .....	113
4.3 - Aims .....	116
4.4 - Materials And Methods .....	116
4.4.1 - Alcohol-Dependence Cohort.....	116
4.4.2 - SNP Genotyping.....	120

4.4.3 - Time-to-event Analysis .....	120
4.4.4 - Population Attributable Risk.....	123
4.5 - Results .....	124
4.5.1 - Genotyping .....	124
4.5.2 - Genetic Association Analysis.....	126
4.5.3 - Time-to-event Analysis .....	127
4.5.4 - Population Attributable Risk.....	138
Discussion .....	140
CHAPTER 5 Structural Studies of <i>PNPLA3</i> .....	146
5.1 - Background .....	147
5.1.1 - Overview.....	147
5.1.2 - The Patatin Domain .....	147
5.1.3 - PNPLA3 Structure .....	149
5.1.4 - PNPLA3 Function .....	151
5.1.5 - PNPLA3 Ile148Met: Structure and Function.....	155
5.2 - Aims .....	156
5.3 - Protein Sequence Analysis.....	157
5.3.1 - Overview.....	157
5.3.2 - Materials and Methods.....	157
5.3.1 - Results.....	161
5.3.2 - Summary .....	169
5.4 - Protein Expression Vectors.....	169
5.4.1 - Overview.....	169
5.4.2 - Materials and Methods.....	170
5.4.3 - Results.....	179
5.4.4 - Summary .....	186
5.5 - Protein Purification.....	188
5.5.1 - Overview.....	188
5.5.2 - Materials and Methods.....	188

5.5.3 - Results.....	193
5.5.4 - Summary .....	201
5.6 - Discussion .....	201
CHAPTER 6 General Discussion.....	203
6.1 - Review of Findings .....	206
6.1.1 - GWAS .....	206
6.1.2 - Extended GWAS.....	206
6.1.3 - <i>PNPLA3</i> and Alcohol-related Cirrhosis Risk.....	206
6.1.4 - <i>PNPLA3</i> : Structural and Functional Studies.....	207
6.2 - Where Are We Going.....	207
6.2.1 - Determining At Risk Groups.....	207
6.2.2 - Understanding Disease Mechanisms.....	210
6.2.3 - Developing New Treatments.....	211
References .....	213
Appendices.....	248

---

---

# ACKNOWLEDGEMENTS

---

---

So many friends and colleagues have provided support, help and guidance, but foremost are my supervisors. First, I am indebted to my primary supervisor, Dr Marsha Morgan; you have provided me so many opportunities from conferences and collaboration to funding and teaching. Your passion for the study of alcohol-related liver disease and your commitment to your students has guided me, vastly broadened my interests and resulted in some cutting edge research. Second, Dr Andrew McQuillin, you have always provided advice when I needed it. Without your guidance, I wouldn't have found the command prompt and the hours of tinkering away at bioinformatics problems that I have enjoyed since. Last, the late Prof. Hugh Gurling, who passed away mid-way through this work. His pioneering research into psychiatric genetics allowed the foundation of the Molecular Psychiatry laboratory without which none of this research would have been possible.

My gratitude goes to Dr Sally Sharp and Dr Alessia Fiorentino for showing me the basics of DNA extraction, PCR and genotyping. I would like to thank all my colleagues and ex-colleagues from the Molecular Psychiatry laboratory including: Yi, John, Niamh, Radhika, Bob, Dave, Giorgia and everyone else for listening and providing ideas at the weekly lab meetings. Next are the team in the UCL Crystallography laboratory; I am indebted to Dr Alun Coker: thank you for providing all the opportunities from trips to Diamond and the Oxford Protein Production Facility to the use of your laboratory facilities. I would like to thank my other colleagues from this laboratory including Shu-Fen, Steve, John, Rebecca, Kirsten, Jason and Raj. It was also a pleasure to supervise, and work with, the intercalating BsC students Rebecca, Adam and Ee-Teng. Most recently, I would like to thank Will and Steve for all their help with proof reading and all the other things. It is good to know that you will both be studying PNPLA3 and the genetics of alcohol-related liver disease.

Without my partner Charlotte and my parents Martin and Pat, I would not have reached this point. Your love, support and encouragement has made all the difference. Mark and Em, thanks for the final proof read and all the good times that we had together in London.

None of this work would have been possible without: first, the financial support of the Hobson Family Trust, Prof. Marsha Morgan and a UCL IMPACT award; second, our collaborators from Imperial College London, Germany and Belgium; and last, the generous donation of DNA samples from many thousands of participants.



---

---

# PUBLICATIONS

---

---

## CONFERENCE ABSTRACTS

Buch S, Stickel F, Trépo E, Way MJ, Herrmann A, Nischalke HD, Brosch M, Rosendahl J, Berg T, Fischer J, Ridinger M, Rietschel M, McQuillin A, Frank J, Kiefer F, Schreiber S, Lieb W, Soyka M, Datz C, Schmelz R, Brückner S, Wodarz N, Devière J, Clumeck N, Sarrazin C, Lammert F, Gustot T, Deltenre P, Völtzke H, Lerch MM, Mayerle J, Eyer F, Schafmayer C, Cichon S, Nöthen MM, Nothnagel M, Ellinghaus D, Franke A, Zopf S, Hellerbrand C, Moreno C, Franchimont D, Morgan MY, Hampe J. (2015) **A two-stage genome-wide association study identifies *TM6SF2* and *MBOAT7* as risk loci for alcohol-related cirrhosis.** *J Hepatol*, 62(1):S260-S261

Way MJ, Atkinson S, McQuillin A, Thursz MR, Morgan MY (2015) **A functional variant in *TM6SF2* associates with alcohol-related cirrhosis risk in a British and Irish population.** *J Hepatol* 62(1): S772

Atkinson S, Way MJ, Mellor J, McQuillin A, Morgan MY, Thursz M (2015) **Homozygosity for the Ile148Met variant in *PNPLA3* is significantly associated with reduced survival following an episode of severe alcoholic hepatitis.** *J Hepatol* 62(1): S771

Way MJ, Gordon HM, Marshall JC, McQuillin A, Morgan MY (2014) **Carriage of the Ile148Met mutation in *PNPLA3* has a detrimental effect on survival in patients with alcohol-related cirrhosis.** *Hepatology* 60(S1): 781A-782A

Way MJ, Morgan MY (2013) **The *PNPLA3* I148m mutation significantly increases the risk of developing alcohol-related cirrhosis in alcohol dependent individuals.** *Alcohol Alcohol* 48(6): 37-37

Way MJ, McQuillin A, Gurling HMD, Morgan MY (2013) **The *PNPLA3* I148M Mutation Significantly Increases The Risk Of Developing Alcohol-Related Cirrhosis In Alcohol Dependent Individuals.** *J Hepatol*, (3):S563-S564

Way MJ (2013) **Genetic variants in the alcohol dehydrogenase genes *ADH1B* and *ADH1C* independently affect susceptibility to alcohol dependence syndrome in a British and Irish Population.** *Alcohol Alcohol*. 48, 30-30

Way MJ, McQuillin AS, Gurling HM, Morgan MY (2013) **The *ZNF699* gene is not associated with alcohol dependence in a UK sample but may alter risk for alcohol-related cirrhosis.** *J Hepatol* S564

## ORIGINAL ARTICLES

Buch S, Stickel F, Trépo E, Way M, Herrmann A, Nischalke HD, Brosch M, Rosendahl J, Berg T, Ridinger M, Rietschel M, McQuillin A, Frank J, Kiefer F, Schreiber S, Lieb W, Soyka M, Semmo N, Aigner E, Datz C, Schmelz R, Brückner S, Zeissig S, Stephan AM, Wodarz N, Devière J, Clumeck N, Sarrazin C, Lammert F, Gustot T, Deltenre P, Völzke H, Lerch MM, Mayerle J, Eyer F, Schafmayer C, Cichon S, Nöthen MM, Nothnagel M, Ellinghaus D, Huse K, Franke A, Zopf S, Hellerbrand C, Moreno C, Franchimont D, Morgan MY, Hampe J. (2015) **A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis.** *Nat Genet* 47(12):1443-8

Ali MA, Way MJ, Marks M, Guerrini I, Thomson AD, Strang J, McQuillin A, Morgan MY. (2015) **Phenotypic heterogeneity in study populations may significantly confound the results of genetic association studies on alcohol dependence.** *Psychiatr Genet* 25(6):234-40

Way MJ, McQuillin A, Saini J, Ruparelia K, Lydall GJ, Guerrini I, Ball D, Smith I, Quadri G, Thomson AD, Kasiakogia-Worlley K, Cherian R, Gunwardena P, Rao H, Kottalgi G, Patel S, Hillman A, Douglas E, Qureshi SY, Reynolds G, Jauhar S, O'Kane A, Dedman A, Sharp S, Kandaswamy R, Dar K, Curtis D, Morgan MY, Gurling HMD (2015) **Genetic variants in or near ADH1B and ADH1C affect susceptibility to alcohol dependence in a British and Irish population.** *Addict Biol* 20(3): 594-604

O'Brien NL, Way MJ, Kandaswamy R, Fiorentino A, Sharp SI, Quadri G, Alex J, Anjorin A, Ball D, Cherian R, Dar K, Gormez A, Guerrini I, Heydtmann M, Hillman A, Lankappa S, Lydall G, O'Kane A, Patel S, Quested D, Smith I, Thomson AD, Bass NJ, Morgan MY, Curtis D, McQuillin A (2014) **The functional GRM3 Kozak sequence variant rs148754219 affects the risk of schizophrenia and alcohol dependence as well as bipolar disorder.** *Psychiatr Genet* 24(6): 277-278

Way MJ (2014) **Computational modelling of ALDH1B1 tetramer formation and the effect of coding variants.** *Chem Biol Interact.* 234(1): 38-44

Quadri G, McQuillin A, Guerrini I, Thomson AD, Cherian R, Saini J, Ruparelia K, Lydall GJ, Ball D, Smith I, Way MJ, Kasiakogia-Worlley K, Patel S, Kottalgi G, Gunawardena P, Rao H, Hillman A, Douglas E, Qureshi SY, Reynolds G, Jauhar S, O'Kane A, Sharp S, Kandaswamy R, Dar K, Curtis D, Morgan MY, Gurling HM (2014) **Evidence for genetic susceptibility to the alcohol dependence syndrome from the thiamine transporter 2 gene solute carrier SLC19A3.** *Psychiatr Genet* 24(3):122-123

Anderson-Schmidt H, Beltcheva O, Brandon MD, Byrne EM, Diehl EJ, Duncan L, Gonzalez SD, Hannon E, Kantojärvi K, Karagiannidis I, Kos MZ, Kotyuk E, Laufer BI, Mantha K, McGregor NW, Meier S, Nieratschker V, Spiers H, Squassina A, Thakur GA, Tiwari Y, Viswanath B, Way MJ, Wong CC, O'Shea A, DeLisi LE (2012) **Selected rapporteur summaries from the XX world congress of psychiatric genetics, Hamburg, Germany, October 14–18.** *Am J Med Genet B Neuropsychiatr Genet.* 162(2): 96-121

---

---

# TABLE OF TABLES

---

---

TABLE 1-1 THE AMOUNT OF ALCOHOL PRESENT IN TYPICAL ALCOHOLIC BEVERAGES .....	3
TABLE 1-2 THE ACUTE AND CHRONIC PHYSICAL HARM ASSOCIATED WITH ALCOHOL MISUSE .....	10
TABLE 1-3 THE EFFECTS OF INCREASING BLOOD ALCOHOL CONTENT IN A NAÏVE MALE* DRINKER .....	11
TABLE 1-4 GLOBAL ALCOHOL-ATTRIBUTABLE FRACTION FOR SELECTED CAUSES OF DEATH.....	12
TABLE 1-5 COMPARISONS BETWEEN THE ICD AND DSM DIAGNOSTIC CRITERIA FOR ALCOHOL DEPENDENCE .....	16
TABLE 1-6 LIVER FUNCTION TEST BIOMARKERS AND FACTORS THAT MAY CONFOUND THEIR INTERPRETATION .....	25
TABLE 1-7 TWIN STUDIES OF ALCOHOL-DEPENDENCE .....	37
TABLE 1-8 ADOPTION STUDIES OF ALCOHOL USE PHENOTYPES.....	38
TABLE 1-9 GENOME-WIDE LINKAGES STUDIES OF ALCOHOL USE PHENOTYPES.....	39
TABLE 1-10 GENOME-WIDE ASSOCIATION STUDIES OF ALCOHOL USE PHENOTYPES.....	41
TABLE 1-11 GENOME WIDE ASSOCIATION STUDIES OF LIVER DISEASE PHENOTYPES .....	46
TABLE 1-12 GENETIC ASSOCIATION STUDIES OF ALCOHOL-RELATED LIVER DISEASE PHENOTYPES .....	48
TABLE 1-13 META-ANALYSES OF CANDIDATE GENES STUDIES IN ALCOHOL-RELATED LIVER DISEASE ..	49
TABLE 2-1 DEMOGRAPHICS FEATURES OF THE DISCOVERY COHORT.....	54
TABLE 2-2 DEMOGRAPHICS OF THE VALIDATION COHORT .....	57
TABLE 2-3 THE GENOME-WIDE GENOTYPING BEAD-CHIP ARRAYS USED ON THE DISCOVERY COHORT...59	
TABLE 2-4 THE DEMOGRAPHICS OF THE SAMPLES IN THE ADJUSTED META-ANALYSIS .....	62
TABLE 2-5 TOP-ASSOCIATION STATISTICS IN THE UK SAMPLE.....	63
TABLE 2-6 FIXED EFFECTS META-ANALYSIS RESULTS FROM THE UK AND GERMAN DISCOVERY COHORT .....	65
TABLE 2-7 VALIDATION OF THE TOP SNPs IDENTIFIED IN THE VALIDATION COHORT .....	67
TABLE 2-8 POST-HOC ASSOCIATION ANALYSIS ADJUSTED FOR GENDER, AGE, BMI AND TYPE II DIABETES STATUS.....	69
TABLE 2-9 THE GENES IN EACH CANDIDATE PATHWAY.....	70
TABLE 2-10 POST-HOC GENE SET-BASED ASSOCIATION TEST RESULTS.....	70
TABLE 2-11 POST-HOC ANALYSIS OF CANDIDATE GENE VARIANTS .....	71
TABLE 3-1 DEMOGRAPHIC FEATURES OF THE UCL AND STOPAH DISCOVERY GROUP .....	82
TABLE 3-2 DEMOGRAPHIC FEATURES OF THE UCL AND STOPAH VALIDATION GROUP .....	82
TABLE 3-3 DEMOGRAPHICS OF THE GERMAN DISCOVERY GROUP.....	86
TABLE 3-4 THE MOST-SIGNIFICANT GENETIC ASSOCIATIONS* IN THE UCL/STOPAH DATASET .....	92
TABLE 3-5 THE MOST SIGNIFICANT ASSOCIATIONS FROM A FIXED EFFECTS META-ANALYSIS IN THE EXTENDED GWAS ANALYSIS.....	95
TABLE 3-6 POSITIONS AND FEATURES OF THE NOVEL LOCI IDENTIFIED DURING THE DISCOVERY META- ANALYSIS .....	96
TABLE 3-7 NUCLEOTIDE SEQUENCES OF THE PRIMERS USED FOR VALIDATION GENOTYPING OF THE SIX MOST SIGNIFICANTLY ASSOCIATED VARIANTS .....	98
TABLE 3-8 A CONCORDANCE COMPARISON OF DIRECT AND IMPUTED GENOTYPE DATA.....	100

TABLE 3-9 THE RESULTS OF A GENETIC ASSOCIATION ANALYSIS IN THE VALIDATION GROUP .....	102
TABLE 3-10 THE RESULTS OF A META-ANALYSIS INCLUDING THE VALIDATION GENETIC ASSOCIATION DATA .....	103
TABLE 4-1 GENOME-WIDE ASSOCIATION STUDIES WHERE <i>PNPLA3</i> HAS BEEN IDENTIFIED AS A SIGNIFICANT LOCUS.....	112
TABLE 4-2 THE ASSOCIATION BETWEEN RS738409 AND HCC RISK.....	114
TABLE 4-3 PUBLISHED CANDIDATE GENETIC ASSOCIATION STUDIES OF <i>PNPLA3</i> IN ALCOHOL-RELATED LIVER DISEASE .....	115
TABLE 4-4 PRIMERS USED FOR GENOTYPING RS738409 IN <i>PNPLA3</i> .....	120
TABLE 4-5 ALLELIC ASSOCIATION ANALYSIS OF RS738409 IN THE UCL COHORT.....	126
TABLE 4-6 BASE LINES DEMOGRAPHICS OF THE TIME-TO-EVENT COHORT BY LIVER DISEASE STATUS.....	127
TABLE 4-7 MEDIAN KAPLAN-MEIER TIME-TO-EVENT ESTIMATES FROM THE DATE OF BIRTH UNTIL THE DATE OF PRESENTATION WITH CIRRHOSIS STRATIFIED BY RS738409 GENOTYPE .....	128
TABLE 4-8 BASELINE CLINICAL AND DEMOGRAPHIC FEATURES IN PATIENTS WITH CIRRHOSIS STRATIFIED BY RS738409 GENOTYPE .....	130
TABLE 4-9 COX PROPORTIONAL HAZARDS ANALYSIS OF THE TIME FROM PRESENTATION WITH CIRRHOSIS UNTIL DEATH OR ORTHOTOPIC LIVER TRANSPLANT USING UNIVARIATE AND MULTIVARIATE MODELS .....	131
TABLE 4-10 KAPLAN-MEIER MEDIAN TIME-TO-EVENT ESTIMATES FOR SEVERAL VARIABLES ASSOCIATED WITH THE TIME TO DEATH OR TRANSPLANT FROM CIRRHOSIS PRESENTATION .....	131
TABLE 4-11 BASELINE CLINICAL DEMOGRAPHIC FEATURES OF PATIENTS WITH CIRRHOSIS AS STRATIFIED BY THE SUBSEQUENT DEVELOPMENT OF HEPATOCELLULAR CARCINOMA.....	134
TABLE 4-12 TEST FOR ASSOCIATION BETWEEN SEVERAL VARIABLES AND ALCOHOL USE RECIDIVISM IN THE ALCOHOL-RELATED CIRRHOSIS TIME-TO-EVENT DATASET .....	134
TABLE 4-13 BASELINE DEMOGRAPHICS OF THE NO-SIGNIFICANT LIVER INJURY PATIENTS STRATIFIED BY RS738409 GENOTYPE .....	135
TABLE 4-14 COX PROPORTIONAL HAZARDS TIME TO DEATH FROM ENROLMENT ANALYSIS IN THE NO-SIGNIFICANT LIVER INJURY GROUP .....	136
TABLE 5-1 THE TOP TEMPLATE STRUCTURES USED IN <i>PNPLA3</i> STRUCTURAL MODEL BUILDING.....	165
TABLE 5-2 POPIN SUITE VECTORS CHOSEN FOR HETEROLOGOUS <i>PNPLA3</i> EXPRESSION.....	170
TABLE 5-3 THE PCR AMPLIFICATION CONDITIONS USED ON THE <i>PNPLA3</i> INSERT .....	173
TABLE 5-4 <i>PNPLA3</i> EXPRESSION CONSTRUCTS CREATED AT THE OXFORD PROTEIN PRODUCTION FACILITY.....	175
TABLE 5-5 MEDIA USED FOR PROPAGATING <i>E. COLI</i> .....	189
TABLE 5-6 PLASMID SEQUENCING PRIMERS .....	189
TABLE 5-7 <i>E. COLI</i> GROWTH AND INDUCTION MEDIA .....	190
TABLE 5-8 THE CONSTITUENTS OF POLYACRYLAMIDE GELS, ELECTROPHORESIS RUNNING BUFFERS AND GEL STAINING MEDIA.....	191
TABLE 5-9 BUFFERS USED IN THE FIRST PURIFICATION .....	195
TABLE 5-10 BUFFERS USED IN THE SECOND PURIFICATION.....	198

---

---

# TABLE OF FIGURES

---

---

FIGURE 1-1 WORLDWIDE ANNUAL PER CAPITA ALCOHOL CONSUMPTION IN 2010.....	5
FIGURE 1-2 ANNUAL PER CAPITA ALCOHOL CONSUMPTION IN THE UK BETWEEN 1900-2010.....	6
FIGURE 1-3 THE SELF-REPORTED DRINKING HABITS OF ADULTS IN THE UK .....	6
FIGURE 1-4 A SCHEMATIC OF THE PRIMARY ALCOHOL METABOLIZING PATHWAYS .....	8
FIGURE 1-5 PHARMACOKINETIC PROFILE OF ALCOHOL CLEARANCE .....	9
FIGURE 1-6 ALCOHOL CONSUMPTION AND THE RISK OF PHYSICAL HARM .....	14
FIGURE 1-7 THE CHEMICAL STRUCTURES OF THE NEUROTRANSMITTERS GLUTAMATE AND GABA.....	18
FIGURE 1-8 BABYLONIAN CLAY LIVER MODEL .....	20
FIGURE 1-9 THE STAGES OF ALCOHOL-RELATED LIVER DISEASE .....	22
FIGURE 1-10 THE ALCOHOL ATTRIBUTABLE FRACTION FOR CIRRHOSIS DEATHS IN EUROPE .....	27
FIGURE 1-11 STANDARDIZED CIRRHOSIS MORTALITY RATES.....	27
FIGURE 1-12 SURVIVAL FROM CLINICAL PRESENTATION WITH ALCOHOL-RELATED CIRRHOSIS .....	29
FIGURE 1-13 MECHANISMS OF ALCOHOL-RELATED STEATOSIS .....	33
FIGURE 1-14 METABOLISM OF ALCOHOL IN THE HEPATOCYTE AND MECHANISMS OF CELL INJURY .....	34
FIGURE 1-15 THE NUMBER OF GENETIC ASSOCIATION STUDIES OF ALCOHOL DEPENDENCE .....	42
FIGURE 2-1 A SCHEMATIC OF THE TWO-STAGE GENOME-WIDE ASSOCIATION STUDY DESIGN.....	54
FIGURE 2-2 A LOCUS PLOT OF GENETIC ASSOCIATIONS IN <i>PNPLA3</i> IN THE GERMAN DISCOVERY GROUP .....	64
FIGURE 2-3 A MANHATTAN AND QUANTILE-QUANTILE PLOT OF THE COMBINED UK-GERMAN GWAS META-ANALYSIS .....	66
FIGURE 2-4 FINE-MAPPING ANALYSIS OF THE GENOME-WIDE SIGNIFICANT LOCI.....	68
FIGURE 2-5 FORMS OF PLEIOTROPY BETWEEN GENETIC VARIANTS AND PHENOTYPES AT A LOCUS .....	73
FIGURE 2-6 PHOSPHOLIPID METABOLISM IN THE LANDS CYCLE .....	74
FIGURE 3-1 THE CHEMICAL STRUCTURES OF PREDNISOLONE AND PENTOXIFYLLINE .....	78
FIGURE 3-2 THE STUDY DESIGN PLAN OF THE EXTENDED GENOME-WIDE ASSOCIATION STUDY.....	81
FIGURE 3-3 THE OVERLAP IN NO-SIGNIFICANT LIVER INJURY CONTROLS.....	84
FIGURE 3-4 A SCHEMATIC OF THE STAGES OF THE DISCOVERY ANALYSIS.....	88
FIGURE 3-5 PRINCIPAL COMPONENT PLOT IN THE UCL-STOPAH MERGED DATASET .....	89
FIGURE 3-6 QUANTILE-QUANTILE PLOT OF THE GENETIC ASSOCIATION RESULTS IN THE EXTENDED UCL/STOPAH DATASET.....	90
FIGURE 3-7 A MANHATTAN PLOT OF THE EXTENDED UCL/STOPAH ANALYSIS <i>P</i> -VALUES .....	91
FIGURE 3-8 QUANTILE-QUANTILE PLOT OF THE META-ANALYSIS <i>P</i> -VALUES .....	93
FIGURE 3-9 A MANHATTAN PLOT OF META-ANALYSIS <i>P</i> -VALUES.....	94
FIGURE 3-10 FLUORESCENCE CLUSTER PLOTS OF VALIDATION GENOTYPING EXPERIMENTS .....	99
FIGURE 3-11 LINEAR REGRESSION PLOT OF GENOTYPING CONCORDANCE VERSUS IMPUTE2 INFO SCORE .....	100
FIGURE 3-12 COMPARING THE CONCORDANCE BETWEEN IMPUTED AND DIRECT GENOTYPE DATA .....	101
FIGURE 3-13 A LOCUS PLOT OF THE GENETIC ASSOCIATION SIGNAL NEAR <i>LIPG</i> AND <i>ACAA2</i> .....	106
FIGURE 4-1 TRANSCRIPTION FACTOR BINDING SITES AND REGULATORY ELEMENTS IN <i>PNPLA3</i> .....	110

FIGURE 4-2 MAP OF RS738409 GENOTYPE FREQUENCIES .....	111
FIGURE 4-3 THE SITES AT WHICH ALCOHOL-DEPENDENT SUBJECTS WERE RECRUITED.....	117
FIGURE 4-4 VENN DIAGRAM OF THE UCL ALCOHOL-DEPENDENCE COHORT SAMPLES STRATIFIED BY LIVER DISEASE STATUS.....	118
FIGURE 4-5 THE OVERLAP IN SAMPLES BETWEEN THE GENETIC ASSOCIATION STUDY OF PNPLA3 AND THE GWAS OF ALCOHOL-RELATED CIRRHOSIS.....	119
FIGURE 4-6 THE TOTAL ENROLMENT OF PATIENTS BY YEAR FROM THE ROYAL FREE HOSPITAL .....	121
FIGURE 4-7 A KASPAR GENOTYPING CLUSTER PLOT FOR RS738409.....	124
FIGURE 4-8 CONCORDANCE BETWEEN DIRECT AND IMPUTED GENOTYPE DATA FOR RS738409 .....	125
FIGURE 4-9 THE GENOTYPE DISTRIBUTION AND ALLELE FREQUENCY OF RS738409 IN THE UCL COHORT .....	126
FIGURE 4-10 KAPLAN-MEIER CURVE IN ALL PATIENTS STRATIFIED BY LIVER STATUS.....	128
FIGURE 4-11 KAPLAN-MEIER CURVE IN ALL PATIENTS STRATIFIED BY RS738409 GENOTYPE .....	129
FIGURE 4-12 KAPLAN-MEIER SURVIVAL CURVES IN PATIENTS WITH CIRRHOSIS .....	132
FIGURE 4-13 THE CAUSES OF DEATH OR LIVER TRANSPLANT IN PATIENTS WITH CIRRHOSIS .....	133
FIGURE 4-14 KAPLAN-MEIER SURVIVAL CURVES IN PATIENTS WITH NO SIGNIFICANT LIVER INJURY ....	137
FIGURE 4-15 POPULATION ATTRIBUTABLE RISK ESTIMATES OF ALCOHOL-RELATED CIRRHOSIS BY RS738409 GENOTYPE AT DIFFERENCE INCIDENCE LEVELS.....	139
FIGURE 4-16 LEAD TIME BIAS.....	141
FIGURE 4-17 THE RELATIONSHIP BETWEEN RELATIVE RISK AND ODDS RATIOS .....	144
FIGURE 5-1 MECHANISM OF THE NUCLEOPHILIC SERINE ACTIVE-SITE RESIDUE.....	148
FIGURE 5-2 PHYLOGENETIC AND STRUCTURAL COMPARISON OF HUMAN PNPLA FAMILY MEMBERS ..	149
FIGURE 5-3 THE DOMAIN ARCHITECTURE OF PNPLA3.....	150
FIGURE 5-4 THE STAGES OF PROTEIN STRUCTURE DETERMINATION BY X-RAY CRYSTALLOGRAPHY...	150
FIGURE 5-5 A PREDICTED MODEL OF THE PATATIN DOMAIN STRUCTURE IN PNPLA3 .....	151
FIGURE 5-6 A PROPOSED FEED FORWARD LOOP NUTRITIONAL REGULATORY MECHANISM OF PNPLA3 .....	152
FIGURE 5-7 A PREDICTED MODEL OF THE PNPLA3 ACTIVE SITE .....	156
FIGURE 5-8 EXPERIMENTAL STAGES TO OBTAIN A PROTEIN STRUCTURE .....	156
FIGURE 5-9 EXPERIMENTAL STAGES TO OBTAIN A PROTEIN STRUCTURE .....	157
FIGURE 5-10 A FLOWCHART OF THE STAGES OF THE ITASSER STRUCTURAL MODELLING PIPELINE..	160
FIGURE 5-11 THE RESIDUE CONTENT OF PNPLA3.....	162
FIGURE 5-12 PLOT OF RESIDUE HYDROPHOBICITY ALONG THE SEQUENCE OF PNPLA3.....	163
FIGURE 5-13 THE PREDICTED INTRINSIC DISORDER ALONG THE SEQUENCE OF PNPLA3 .....	163
FIGURE 5-14 MULTISPECIES ALIGNMENT OF MAMMALIAN PNPLA3 ORTHOLOGUES .....	164
FIGURE 5-15 TOP ITASSER THREADING TEMPLATE SEQUENCE ALIGNMENTS .....	166
FIGURE 5-16 THE ESTIMATED LOCAL ACCURACY OF THE PNPLA3 STRUCTURAL MODEL.....	167
FIGURE 5-17 THE TOP STRUCTURAL MODEL OF FULL LENGTH PNPLA3.....	168
FIGURE 5-18 THE PATATIN DOMAIN OF THE TOP MODEL .....	168
FIGURE 5-19 EXPERIMENTAL STAGES TO OBTAIN A PROTEIN STRUCTURE .....	169
FIGURE 5-20 THE BACKBONE OF A POPINE PLASMID .....	171
FIGURE 5-21 PNPLA3 INSERTS SELECTED FOR PLASMID CONSTRUCTION.....	172

FIGURE 5-22 A SCHEMATIC OF THE STAGES OF HIGH-THROUGHPUT PLASMID CONSTRUCTION .....	173
FIGURE 5-23 STAGES OF INFUSION CLONING .....	174
FIGURE 5-24 STAGES OF <i>E. COLI</i> /PROTEIN EXPRESSION TRIALS.....	177
FIGURE 5-25 A COMPARISON OF CORRECT <i>PNPLA3</i> INSERT SIZE FOR IN-FUSION CLONING .....	179
FIGURE 5-26 PLATE LAYOUT OF POPIN-PNPLA3 CONSTRUCTS .....	180
FIGURE 5-27 SDS-PAGE GELS OF AFFINITY PURIFIED, AUTO-INDUCED ROSETTA LYSATES .....	181
FIGURE 5-28 SDS-PAGE GELS OF AFFINITY PURIFIED, IPTG-INDUCED ROSETTA LYSATES .....	182
FIGURE 5-29 SDS-PAGE GELS OF AFFINITY PURIFIED, AUTO-INDUCED LEMO21 LYSATES.....	183
FIGURE 5-30 SDS-PAGE GELS OF AFFINITY PURIFIED, IPTG-INDUCED LEMO21 LYSATES .....	184
FIGURE 5-31 PNPLA3 EXPRESSION TRIALS IN HEK293T CELLS .....	185
FIGURE 5-32 PNPLA3 EXPRESSION TRIALS IN Sf9 CELLS .....	186
FIGURE 5-33 A PLASMID MAP OF THE A4 PLASMID CONSTRUCT.....	187
FIGURE 5-34 EXPERIMENTAL STAGES TO OBTAIN A PROTEIN STRUCTURE .....	188
FIGURE 5-35 PRINCIPLE OF SIZE-EXCLUSION CHROMATOGRAPHY .....	193
FIGURE 5-36 SEQUENCING CONFIRMATION OF THE PNPLA3-MBP CONSTRUCT.....	194
FIGURE 5-37 SIZE EXCLUSION PURIFICATION ELUTION PROFILE FROM THE FIRST PURIFICATION .....	196
FIGURE 5-38 THE PROTEIN BANDS PRESENT IN THE PEAK ELUTION FRACTIONS.....	197
FIGURE 5-39 THE PROTEIN BANDS SELECTED FOR MASS-SPECTROMETRY ANALYSIS .....	197
FIGURE 5-40 FAST LIQUID PROTEIN CHROMATOGRAPHY ELUTION PROFILE FROM AN AFFINITY COLUMN .....	199
FIGURE 5-41 VISUALIZATION OF PROTEIN SAMPLES ELUTED FROM AFFINITY COLUMN .....	199
FIGURE 5-42 VISUALIZATION OF SAMPLE AT SEVERAL DILUTIONS AND DENATURATION CONDITIONS...200	200
FIGURE 5-43 SIZE EXCLUSION CHROMATOGRAM ELUTION PROFILE FROM THE SECOND PURIFICATION .....	200
FIGURE 6-1 STANDARDIZED UK MORTALITY RATE DATA .....	204
FIGURE 6-2 ALCOHOL-RELATED CAUSES OF DEATH BY AGE-GROUP IN ENGLAND AND WALES IN 2012 .....	204
FIGURE 6-3 DIRECTLY STANDARDIZED YEARS OF LIFE LOST DUE TO CHRONIC LIVER DISEASE INCLUDING CIRRHOSIS PER POPULATION BY PRIMARY CARE TRUST.....	205
FIGURE 6-4 THE POTENTIAL FOR, AND COST OF, INTERVENTION DURING THE COURSE OF LIVER DISEASE .....	208
FIGURE 6-5 THE OPPOSING EFFECTS OF THE RS58542926 IN <i>TM6SF2</i> .....	211



---

# ABBREVIATIONS

---

AAF. Alcohol Attributable Fraction  
ABV. Alcohol By Volume  
ABW. Alcohol By Weight  
ADH. Alcohol Dehydrogenase  
ALDH. Acetaldehyde Dehydrogenase  
AMP-K. Adenosine Monophosphate-activated Kinase  
BLAST. Basic Local Alignment Search Tool  
BMI. Body Mass Index  
CI. Confidence Interval  
C-score. Confidence Score  
CTP. Child-Turcotte-Pugh Score  
CYP2E1. Cytochrome P450 2E1  
DSM. Diagnostic and Statistical Manual of Mental Disorders  
GABA. Gamma-amino Butyric Acid  
GFP. Green Fluorescent Protein  
GWAS. Genome-wide Association Study  
HCC. Hepatocellular Carcinoma  
HDL. High-density Lipoprotein  
HEK. Human Embryonic Kidney Cell  
hg19. Human Genome Build 19  
HPLC. High Performance Liquid Chromatography  
HSC. Hepatic Stellate Cell  
ICD. International Classification of Diseases  
KASPAR. K-Bioscience Allele Specific PCR  
kDa. Kilodaltons  
LOD. Logarithm of Odds Ratio  
MBP. Maltose Binding Protein  
MC. Monte Carlo  
MD. Molecular Dynamics  
MELD. Model for End Stage Liver Disease Score  
MEOS. Microsomal Ethanol Oxidizing System  
mg. Milligram  
mL. Millilitre  
MMP. Matrix Metalloproteinases  
MS. Mass Spectrometry  
NAD. Nicotinamide Adenine Dinucleotide  
NADPH. Nicotinamide Adenine Dinucleotide Phosphate  
NAFLD. Non-alcohol-related Fatty Liver Disease  
NCBI. National Center for Biotechnology Information  
NMDA. N-methyl-D-aspartate  
OPPF. Oxford Protein Production Facility  
OR. Odds Ratio  
PAGE. Polyacrylamide Gel Electrophoresis

PAR. Population attributable risk  
PDB. Protein Databank  
PNPLA. Patatin-like Phospholipase Domain Containing  
PNPLA3. Patatin-like Phospholipase Domain Containing 3  
PPAR. Peroxisome Proliferator Activated Receptors  
QQ. Quantile-Quantile  
RMSD. Root Mean Squared Deviation  
RR. Relative Risk  
SDS. Sodium Dodecyl Sulphate  
SEC. Size Exclusion Chromatography  
SNP. Single Nucleotide Polymorphism  
SREBP. Sterol Regulatory Element Binding Protein  
STL. Southampton Traffic Light  
STOPAH. Steroids or Pentoxifylline for Alcoholic Hepatitis Trial  
SUMO. Small Ubiquitin Like Modifier  
TB. Terrific Broth  
TIMP. Tissue Inhibitors of Metalloproteinase  
TLR4. Toll-like Receptor 4  
TM-score. Template Modelling Score  
TNF- $\alpha$ . Tumour Necrosis Factor Alpha  
TXR. Thioredoxin Reductase  
WHO. World Health Organization

---

# ABSTRACT

---

The presence of cirrhosis is associated with significant morbidity and mortality. The prolonged heavy consumption of alcohol results in the development of cirrhosis in approximately 20% of people. This variance in risk is attributed to both genetic and environmental risk factors. To date there have been no systematic genome-wide association studies to investigate the contribution of common genetic variants to alcohol-related liver disease risk despite the success of these type of analyses in cirrhosis of other aetiologies. The work presented in the first half of this thesis describes the first genome-wide association study of alcohol-related liver disease identifying three highly replicable risk variants in the candidate loci: *PNPLA3*, *TM6SF2* and *MBOAT7*. Subsequent to this, an extended GWAS was performed including over three-hundred additional cases, resulting in the identification of several novel loci. Of all the identified susceptibility loci, genetic variation in *PNPLA3* is by far the most significant contributor to overall risk. The primary genetic variant in this locus is rs738409 and has the largest effect on cirrhosis risk of any genetic variants identified. It encodes a non-synonymous amino acid substitution (Ile148Met) in the PNPLA3 protein and is thought to be functional. This variant was genotyped in a UCL cohort of phenotypically well-characterized patients with cirrhosis and appropriate controls and investigated with regards to disease outcome and prognosis using time-to-event analysis. However, there is considerable debate about the functional significance of the PNPLA3 variant and to date little, if any, information about its structure. This information would provide insight into the pathogenesis of alcohol-related liver injury and identify possible targets for therapy. In the second half of the thesis the work undertaken to obtain the crystal structure of PNPLA3 is described. Initial computational structural modelling of the PNPLA3 protein was undertaken to characterize predicted structural features of the full-length protein to aid understanding of protein and crystallography process. A range of recombinant protein expression vectors were then developed for the purification of the protein in multiple expression systems. This led to the development of a high-yield expression protocol in *E. coli* obtaining milligram amounts of protein. Work to functionally characterize this protein is on-going.

---

---

# CONTRIBUTIONS

---

---

## GENERAL

I was involved with the extraction, purification and quantification of many hundreds of DNA samples and the management of a database for the storage of phenotypic information. Before I arrived in the laboratory, Radhika Kandaswamy, Jit Saini, Kush Ruperalia, Alex Dedman and Andrew McQuillin were responsible for the extraction, quantification and database logging of many of the early UCL population control and alcohol dependent DNA samples. Sample recruitment was coordinated by Alex Narula and Andrew McQuillin through the National Institute of Health Research funded Mental Health Research Network and other Gastroenterology/Hepatology centres.

## CHAPTER 2

The design and conceptualization for the GWAS was performed by Prof. Marsha Y Morgan, Dr Felix Stickel and Prof. Jochen Hampe. I was involved in the preparation and shipment of all UK samples for GWAS genotyping and data analysis. The primary data analysis and management of data were performed by Dr Stephan Buch of the University of Kiel in Germany. I drafted parts of the original publication and critically reviewed the manuscript. I further performed all of the replication genotyping of a number of SNPs in a separate population of UK alcohol dependent and population controls, a pathways analysis and a candidate gene analysis.

## CHAPTER 3

Prof. Marsha Y Morgan, Dr Steve Atkinson and I, performed the design and conceptualization for the extended GWAS. Dr Stephen Atkinson and Prof. Marsha Y Morgan were responsible for the selection and phenotypic characterisation of the samples that underwent genome-wide genotyping I performed all the analysis from the merging and harmonisation of genome-wide genotyping data, genome-wide imputation, primary association analysis. I further performed all of the replication genotyping of a number of SNPs in a separate population of UK alcohol dependent and population controls.

## CHAPTER 4

Prof. Marsha Morgan, Dr Andrew McQuillin and I, performed the design and conceptualization for the detailed rs738409 genetic association analysis. For this work I performed all of the experimental design, genotyping and data analysis. This work

was reliant on the accurate phenotyping of alcohol-related liver disease samples, which would not be possible without the tireless efforts of Prof. Marsha Morgan and the data collection of many other individuals including: Harriet Gordon, Jonathan Marshall Clive Jackson and Sara Montagenese.

## CHAPTER 5

I was involved in the design and conceptualization of this project with Prof. Marsha Morgan and Dr Alun Coker helping with the initial application to the Oxford Protein Production facility. I further performed much of the cloning and recombinant protein expression trials at the Oxford Protein Production facility with the help of staff at this centre and Shu-Fen Coker and Rebecca Boys. I independently performed bioinformatics analysis of the protein and structural modelling and further optimisation of protein expression trials in *E. coli*.

---

---

# CHAPTER 1 INTRODUCTION

---

---

## 1.1 - OVERVIEW

Millions of people consume alcohol without ill effects, yet for some its consumption is associated with significant adverse social, emotional and medical effects. This section provides an overview of alcohol in the context of its basic chemical properties; the history of its consumption; the current epidemiology of alcohol use/misuse; the harms associated with its consumption, especially when misused, with particular focus on the pathophysiology of alcohol-related liver injury and the interplay between environmental and genetic factors in its pathogenesis.

## 1.2 - ALCOHOL

### 1.2.1 - BEVERAGE ALCOHOL

Ethanol or ethyl alcohol ( $C_2H_5OH$ ), otherwise known as alcohol, is the primary agent in alcoholic beverages such as beer, wine and spirits. Alcohol is a small molecule (molar mass=46.07 g/mol) that is soluble in water and some non-polar solvents. Pure alcohol, known as absolute alcohol, is a volatile, flammable, colourless liquid at room temperature.

Alcoholic beverages are produced through the processes of fermentation and distillation. Fermentation is the process by which carbohydrates are converted to alcohol through anaerobic respiration in yeast. The process of fermentation is self-limiting, as concentrations of alcohol in the fermenting medium above certain concentrations will kill the yeast. The concentration of beverage alcohol may be increased by distillation through the selective evaporation of alcohol (boiling point = 78.4°C) from water (boiling point = 100°C).

Beers, ciders and wines are all produced by the fermentation of the sugar found in malted barley, apple juice or grape juice respectively. Spirits, such as vodka, whisky and gin, are produced by the distillation of the alcohol produced by the fermentation of natural sugars extracted from a variety of natural sources. There are many beverages that are processed significantly beyond the point of fermentation and distillation, such as mixers or alcopops, fortified wines and white ciders. Due to the large differences in production methods, the amount of alcohol present in different alcoholic beverages varies significantly (Table 1-1).

Table 1-1 The amount of alcohol present in typical alcoholic beverages

Drink	Typical volume of a single measure	Alcohol concentration (% ABV)	Units
Wine	250 mL	12 %	3
Beer, Lager, Cider	567 mL	5.2 %	3
Spirits	25 mL	40 %	1

Abbreviations: ABV – Alcohol By Volume; mL - Millilitre

There are many different systems in use for the quantification of the alcohol content in beverage alcohol. The alcohol by volume (ABV) is routinely used and it is defined as the number of mL of pure alcohol present in a 100 mL volume of beverage alcohol at 20°C. A related unit, the alcohol by weight (ABW) may be used to quantify the weight of alcohol present in a beverage by multiplying the ABV by the relative gravity of alcohol (0.789). Many nations, have attempted to simplify the quantification of alcohol content in beverage alcohol into ‘standard units’. In the UK, the widely used ‘unit’ equates to approximately 10 mL or 8 g of absolute alcohol.

## 1.2.2 - HISTORICAL BACKGROUND

The earliest direct archaeological evidence of purposeful alcoholic fermentation dates back to beyond 8,000 BC evidenced by chemical residues on ceramic jars used by Neolithic cultures present in East-Asia<sup>287</sup> and the Middle East<sup>292</sup>. In Europe, there is clear evidence of wine production by Southern European cultures by 3,000 BC<sup>163</sup> and the production of beer in Northern-Europe by 1,000 BC<sup>414</sup>. There is also evidence of alcohol production from the 1<sup>st</sup> century AD onwards in South America, Africa and the Indian sub-continent<sup>99</sup>. By the 17<sup>th</sup> century, only isolated cultures in North America and the Pacific islands lacked access to the knowledge, and technology, required to ferment alcohol until these techniques were introduced.

The discovery of distillation, which is the process of separating the component substances from a liquid mixture by selective evaporation and condensation, changed the nature of human alcohol consumption making it far easier to ingest large quantities of alcohol rapidly. This technique was first used from the 1<sup>st</sup> to the 4<sup>th</sup> centuries by Greek alchemists living in Alexandria, which was then the capital of Roman and later Byzantine Egypt, where the process was used to purify plant extracts. The distillation of alcohol is thought to have been developed by Arab chemists who, following the Muslim conquest of Egypt, applied techniques learned from alchemist scripts present in Alexandria. Notably, the word alcohol is thought to derive from the Arabic ‘*al-kuhl*’. The first unambiguous evidence of distillation appeared from the School of Salerno in



Italy in the 12<sup>th</sup> century<sup>128</sup>. However, early distilled alcohol was primarily reserved for medicinal uses rather than consumption. It was not until the 17<sup>th</sup> century onwards that spirits were widely consumed. A particularly notable period of historical heavy spirit consumption was the 'gin craze' in 18<sup>th</sup> century Britain where, due to a combination of increased levies in the price of beer and the popularisation of gin, following the accession of William of Orange, per capita alcohol consumption significantly increased.

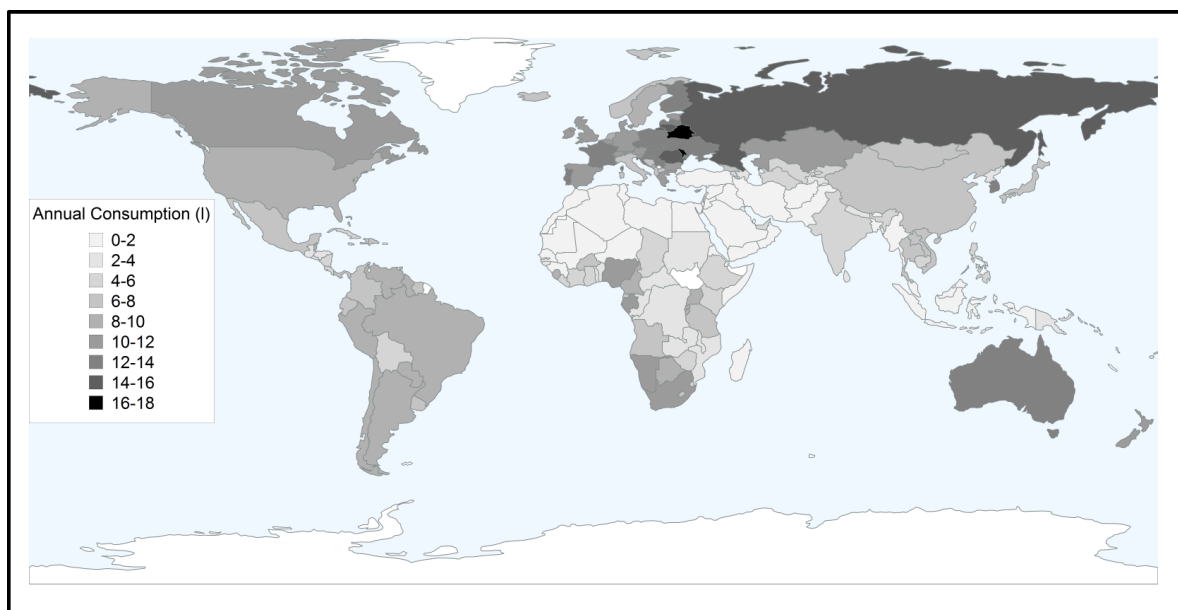
Despite producing alcohol for thousands of years, humans have had little understanding of the microbiological underpinnings of fermentation. In the 19<sup>th</sup> century, Louis Pasteur demonstrated that yeast are living micro-organisms that can convert sugars into alcohol in the absence of oxygen<sup>327</sup>. This led to the discovery, by Eduard Buchner<sup>45</sup>, that cell-free yeast extracts have the capacity to ferment carbohydrates to form ethanol. This discovery in turn played a key role in the discovery of enzymes, which would lay the foundations for the development of biochemistry.

### 1.2.3 - ALCOHOL CONSUMPTION

Alcohol consumption is ubiquitous; the worldwide average per capita alcohol consumption for adults (aged 15+ years) is 13.5 g of alcohol per day<sup>466</sup>. Most countries control aspects of alcohol production, importation and exportation, distribution, marketing and aspects of consumption through legislation. Thus, Governments can control price by imposition of excise duties; distribution by limiting the licensing and operating hours of outlets and personal consumption by imposing a minimum age-limit for purchase, and drink-driving laws. It is estimated that nearly a quarter of the alcohol consumed globally is homemade or illegally produced and therefore outside of Governmental control<sup>466</sup>.

The total per capita alcohol consumption in different nations and global regions varies significantly. For example, the nations of Europe consume over 25% of global alcohol yet comprise only 15% of the total global population. The factors that mediate the levels of alcohol consumption within countries and indeed communities are complex. In some instances there are clear explanations; thus alcohol is proscribed in certain cultures, for instance Islam, but otherwise a multitude of factors work in concert including affordability, legislation, socio-demography, the level of economic development, diverse cultural factors, and beverage preference. Typically, the lowest levels of alcohol consumption are recorded in North Africa and the Middle East, due to the predominance in these regions of the Islamic faith, whereas the highest levels of consumption are predominantly seen in Europe, North America and Russia (Figure 1-1). There are many nations where the profile of alcohol consumption levels is rapidly changing, largely in relation to income levels; most notably the emerging economies of India and China, which contain over a third of the global population, have some of the

most significant increases in levels of alcohol consumption over a five-year period 2006-2010<sup>466</sup>.



**Figure 1-1 Worldwide annual per capita alcohol consumption in 2010**

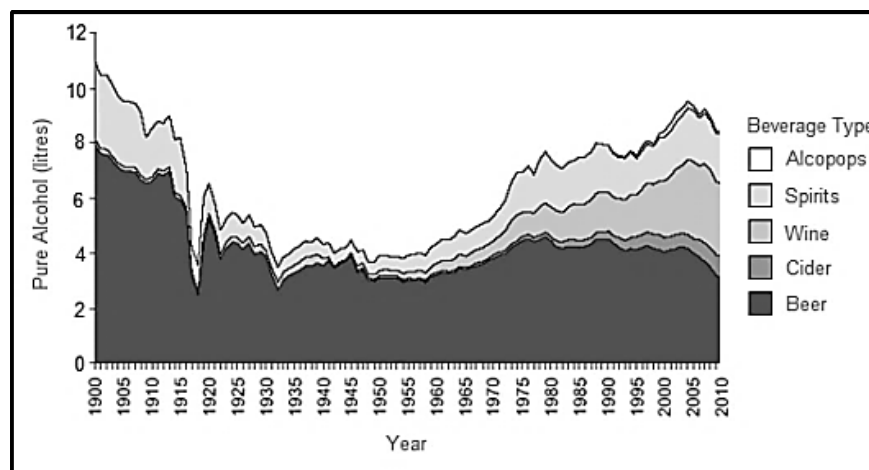
The worldwide annual consumption of alcohol varies significantly with the nations of Europe and Russia typically consuming the most. Data source: World Health Organization, 2010<sup>465</sup>

## UNITED KINGDOM

Alcohol consumption is prevalent in the UK. An analysis of harmful alcohol use by the Organisation for Economic Co-operation and Development found that the UK was the eleventh highest consumer of alcohol out of 40 nations investigated<sup>314</sup>. Another global health report by the World Health Organization (WHO) found the average alcohol consumption for adults in the UK during 2010 was 13.8 litres of pure alcohol per year<sup>466</sup>. The predominant alcoholic beverages consumed in the UK were beer (37%), wine (34%) and spirits (22%).

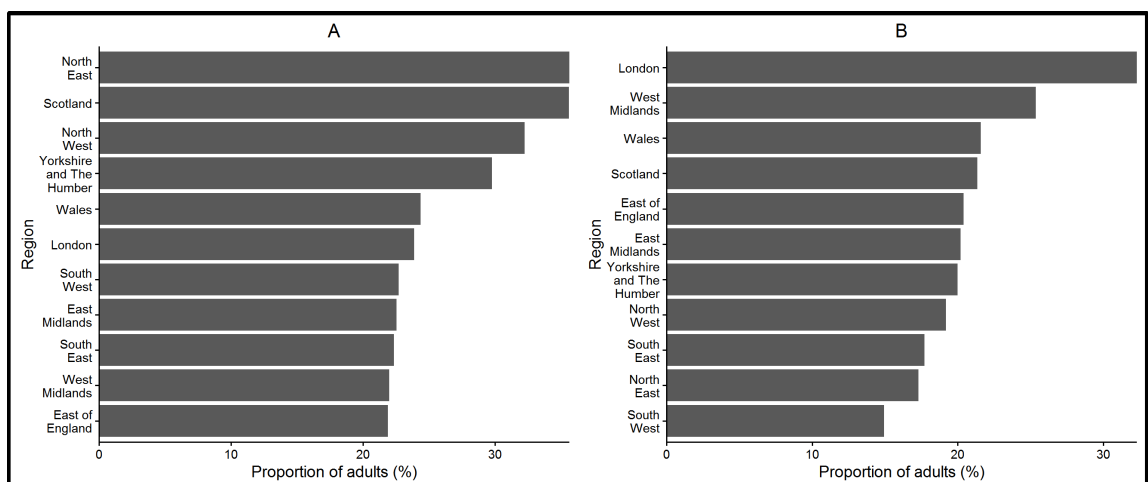
There are significant demographic differences in alcohol consumption in the UK. As in other countries, men consume on average more than double the amount of alcohol compared with women. There are also differences in the patterns and of levels of consumption by age group. In the UK-wide general lifestyle survey over a single week<sup>84</sup>, nearly one third of sixteen to twenty-four year-olds reported heavy drinking (>12 units/week) in comparison to 3% of those aged over sixty-five. In contrast, older people were more likely to drink daily where 18% of the over sixty-five year's old age group consumed an alcoholic beverage on five or more days of the week. There are significant regional differences in the patterns of consumption with adults in the North of England and Scotland more likely binge on alcohol whereas almost a third of adults in London are teetotal<sup>319</sup> (Figure 1-3).

There have been considerable shifts in the amounts and the types of alcoholic beverages consumed in the UK over the past century (Figure 1-2). Since the 1950's the average alcohol consumption has doubled reaching a peak in 2004 of 11.5 litres of pure alcohol per year per person. This increase has followed equivalent increases in the UK gross domestic product. There have also been significant changes in the types of alcoholic beverage consumed, with decreasing beer consumption and increasing wine, cider and alcopop consumption (Figure 1-2). In the past fifty years, the year with the highest yearly consumption levels was 2004 and since then there has been a moderate decline in total alcohol consumption on average. This decrease in total alcohol consumption has not resulted from equivalent declines in every member of the population, as in part this trend reflects an increasing relative number of teetotallers in the UK population, the number of whom have increased by 40% since 2005<sup>319</sup>.



**Figure 1-2 Annual per capita alcohol consumption in the UK between 1900-2010**

The type and amount of alcohol consumed annually per capita has altered significantly over the past century. Image adapted from: Health Committee report, 2012<sup>171</sup>



**Figure 1-3 The self-reported drinking habits of adults in the UK**

The showing the proportion of adults in each region that are: A – binge drinking; B – teetotal. Data source: Orchard et al., 2015<sup>319</sup>

## ALCOHOL METABOLISM

Beverage alcohol is absorbed through both the stomach and the upper small intestine<sup>162</sup>, entering the body via absorption. Both the stomach and liver are exposed to high alcohol concentrations during this initial absorption phase: the stomach via ingestion and the liver via blood received directly from the stomach and small bowel through the portal vein. Once alcohol has entered peripheral circulation, it rapidly equilibrates with water in the body and thus most tissues besides the stomach and liver are exposed to equivalent concentrations of alcohol.

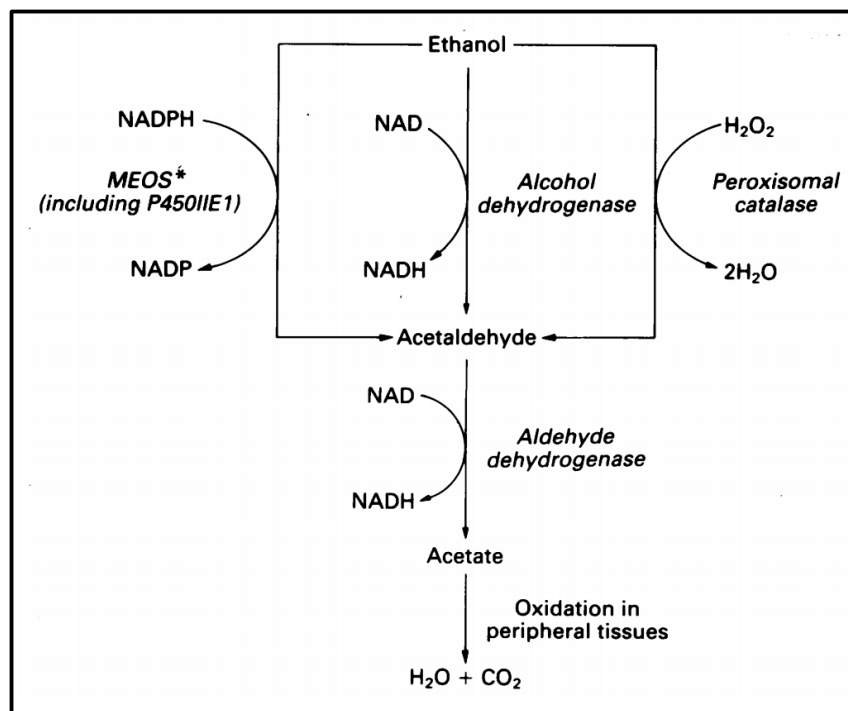
The majority (80-90%) of alcohol metabolism occurs in the liver. The predominant cell type in the liver, hepatocytes, are responsible for the metabolism of alcohol via a relatively simple biochemical pathway known as the alcohol metabolizing pathway. The primary oxidative enzyme responsible for the breakdown of alcohol is alcohol dehydrogenase (ADH). In humans there are several ADH enzymes divided into four classes. The class I ADH enzymes (ADH1A, ADH1B and ADH1C) have a high affinity for ethanol at physiological concentrations, and actively oxidize ethanol into acetaldehyde in hepatocytes<sup>474</sup>. ADH enzymes utilise a nicotinamide adenine dinucleotide (NAD) cofactor, which is reduced by the electrons released from the oxidation of alcohol (Figure 1-4). Other classes of ADH enzymes may play roles in alcohol metabolism: notably, the class IV enzyme ADH7, which is predominantly expressed in the stomach, may play a significant role in gastric tissues when exposed to high alcohol concentrations and thus contribute to first pass metabolism<sup>317</sup>.

The acetaldehyde produced through the oxidation of alcohol is highly reactive, having the capacity to form adducts with important biomolecules including DNA and proteins. Acetaldehyde is significantly more toxic than alcohol, requiring only sub-micro molar concentrations to induce an acute toxic response and it is a known mutagen<sup>93</sup>. For these reasons, the efficient clearance of acetaldehyde is essential. This process is effected by three separate acetaldehyde dehydrogenase (ALDH) enzymes: ALDH1A1, ALDH2 and ALDH1B1, which form homo- and hetero-tetramers with differing substrate affinities and kinetic profiles for acetaldehyde<sup>409</sup>.

ALDH2 is primarily responsible for the oxidation of acetaldehyde to acetate. ALDHs are oxidative enzymes with both cytosolic and mitochondrial isoforms, which utilise a reduced NAD cofactor to oxidize acetaldehyde to acetate (Figure 1-4). Acetate is non-toxic, is produced through the metabolism of many other compounds besides alcohol and may be utilised in the citric acid cycle generating carbon dioxide and water. The overall metabolism of alcohol is biochemically exergonic where every gram of alcohol contains 7.2 calories of energy. However, alcoholic beverages provide little nutritional

benefit besides raw calories although the net energy provided by alcohol metabolism is contentious<sup>250</sup>.

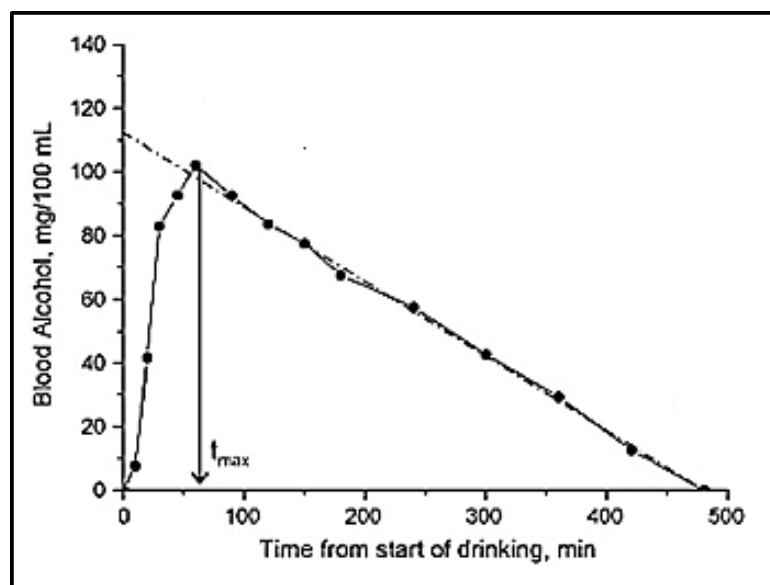
The majority of alcohol is metabolized via the ADH enzyme pathway in the liver although other metabolic pathways exist. The primary non-ADH pathway is the microsomal ethanol oxidizing system (MEOS)<sup>251</sup>. The primary enzyme of this system, cytochrome P450 2E1 (CYP2E1), metabolizes alcohol to produce acetaldehyde through oxidation. However, unlike ADH this enzymes utilises oxygen and a nicotinamide adenine dinucleotide phosphate (NADPH) cofactor by transferring electrons from NADPH to the oxygen molecule forming a superoxide molecule ( $O_2^-$ ). This superoxide molecule can react with alcohol, forming an unstable intermediate, which quickly breaks down to acetaldehyde that will be further metabolized by ALDH enzymes. CYP2E1 is a 'leaky' enzyme having the ability to create superoxide and other reactive oxygen species without the presence of substrate, contributing to oxidative stress in hepatocytes. The MEOS system is inducible, unlike the ADH enzymes, and thus contributes to the increased efficiency of alcohol metabolism observed in chronic alcohol misusers<sup>6</sup>.



**Figure 1-4 A schematic of the primary alcohol metabolizing pathways**

There are several pathways through which alcohol is metabolized including the primary pathway consisting of alcohol and aldehyde dehydrogenase enzymes but also by secondary mechanisms by the enzyme catalase and the microsomal ethanol oxidizing system. Image adapted from Day et al, 1992<sup>86</sup>

The pharmacokinetic elimination profile of alcohol from the body generally follows dose-independent or zero-order kinetics. Following alcohol consumption, the blood alcohol content rapidly rises, reaching a maximum after approximately one hour (Figure 1-5). The class I ADH enzymes have a low  $K_m$  for alcohol (2–10 milligram (mg)/100 mL) and therefore it is readily saturated resulting in a subsequent linear rate of elimination known as the alcohol elimination rate. This rate varies significantly between individuals, in a Swedish study of over 1090 (88.3% male) apprehended drink drivers the average alcohol elimination rate measured over an hour at two time-points was 19.1 mg/100mL of blood per hour<sup>206</sup>.



**Figure 1-5 Pharmacokinetic profile of alcohol clearance**

Following the consumption of a single dose of alcohol the blood alcohol content (y-axis) rapidly rises to reach a peak concentration at the time  $t_{max}$  (x-axis) followed by a linear rate of clearance. Image modified from Jones et al. 2010<sup>205</sup>

A number of factors influence the peak blood alcohol content, including the speed at which beverages are drunk, whether consumed with food, the rate of gastric emptying and body habitus. There are significant differences in the peak blood alcohol content by gender where women attain consistently higher blood alcohol concentration than men following equivalent consumption and controlling for differences in total body mass<sup>96,283</sup>. Compositional differences between the typical male and female body explain these differences as on average, the fraction of body water is lower in women (mean body water %  $\pm$  standard deviation = 48.5%  $\pm$  8.6%) than men (mean body water %  $\pm$  standard deviation = 58.3%  $\pm$  6.7%)<sup>451</sup>. A number of factors also influence the alcohol elimination rate including functional genetic variation in the genes encoding ADH and ALDH enzymes<sup>75,474</sup>, whether certain other drugs are present in peripheral circulation<sup>291</sup>, whether chronic alcohol consumption has upregulated the MEOS system and whether the liver is functionally impaired<sup>321</sup>. There are also significant, yet small, differences in the alcohol elimination rate by gender where women have a faster

alcohol elimination rate than men. This difference may result from higher mass of the liver in relation to the total water volume in women<sup>96</sup>.

### 1.3 - ALCOHOL AND HARM

The consumption of alcohol is associated with wide-ranging problems relating to physical harm, either through direct injury, or by its contribution towards other health conditions. In 2012 over three million deaths, 6% of the global total, were attributable to alcohol<sup>466</sup>. This is a proportion higher than the total number of deaths attributable to HIV/AIDS, tuberculosis or violence. The harm attributed to alcohol may be characterized as either acute or chronic (Table 1-2) relating to whether the effect results from a single drinking event or from long-term alcohol consumption.

Table 1-2 The acute and chronic physical harm associated with alcohol misuse

<b>Acute Harm</b>	
Accidents and injury	Pancreatitis
Acute alcohol poisoning	Cardiac arrhythmias
Aspiration pneumonia	Cerebrovascular accidents
Oesophagitis	Neuropraxia
Mallory-Weiss syndrome	Myopathy/rhabdomyolysis
Gastritis	Hypoglycaemia
<b>Chronic Harm</b>	
Accidents and injury	Brain damage:
Oesophagitis	Dementia
Gastritis	Wernicke-Korsakoff syndrome
Malabsorption	Cerebellar atrophy
Malnutrition	Marchifava-Bignami syndrome
Pancreatitis	Central pontine myelinosis
Liver damage:	Peripheral neuropathy
Fatty change	Myopathy
Hepatitis	Osteoporosis
Cirrhosis	Skin disorders
Systemic hypertension	Malignancies
Cardiomyopathy	Sexual dysfunction
Coronary heart disease	Infertility
Cerebrovascular accidents	Foetal damage

Table adapted from Morgan et al. 2010<sup>296</sup>

## ACUTE HARM

The risk of acute harm associated with alcohol consumption is strongly related to the blood alcohol content attained and the effects of this on the individual (Table 1-3). Depending on the country in which it is used, blood alcohol content is units of mass of alcohol per volume of blood or mass of alcohol per mass of blood. In the UK, blood alcohol content is measured as mg of alcohol present in 100 mL of blood.

Table 1-3 The effects of increasing blood alcohol content in a naïve male\* drinker

Amount imbibed (units)	BAC (mg/100mL)	Effects
2	30	Increased accident risk
3	50	Increased mood; judgement impaired; loss of inhibitions
5	80	Increased risk-taking behaviour; jocularity; drink-drive limit
10	150	Loss of self-control; exuberance; slurring; quarrelsomeness
12	200	Staggering; diplopia; memory loss
25	400	Oblivion; sleepiness; coma
30	500	Death possible
38	600	Death certain

Abbreviation: BAC – blood alcohol content

\*BAC is approximately one-third higher in women drinking the same amount of alcohol as a man of the same body weight.

Table adapted from Morgan et al., 2010<sup>296</sup>

A primary cause of acute harm relating to alcohol consumption is either intentional or unintentional injury. Intentional injuries include those to self or others as a result of alcohol-related violence and the deliberate self-harm associated with parasuicide. Unintentional injuries, such as accidents in the home, at work and on the roads, may affect not only the health of the individual but also that of others. The risk of such injuries increases exponentially with increasing blood alcohol content<sup>422</sup> and are attributed to the neurobiological effects of alcohol including behavioural disinhibition, loss of motor-coordination and impaired judgement. Alcohol-related physical harms account for over a quarter of alcohol-attributable deaths and occur following both chronic and acute alcohol consumption.

Another harm associated with acute alcohol consumption is alcohol poisoning. At moderate consumption levels (< 5 units), there is little risk of alcohol poisoning whereas there is moderate risk of accidental death (Table 1-3). More severe intoxication results in the significantly increased risk of alcohol poisoning, which may



result in death. This occurs when alcohol suppresses vital bodily processes such as breathing, heart rate and temperature homeostasis. There are many factors that influence the acute effects of alcohol and hence risk of alcohol poisoning or physical harm such as gender, whether chronic alcohol consumption has upregulated the MEOS system, the pattern of alcohol consumption, whether food was consumed with alcohol, liver function and the body habitus.

## CHRONIC HARM

The harm caused by chronic alcohol consumption is a component cause of over two-hundred health conditions<sup>463</sup>. The alcohol attributable risk fraction (AAF) may be used to quantify the relative contribution of alcohol consumption to an outcome of a health condition. Out of the many conditions in which alcohol consumption is a contributory factor, cirrhosis has the highest AAF for death<sup>360</sup> (Table 1-4). Other conditions with a high AAF for death include pancreatitis and certain cancers of the mouth and gullet. The link between alcohol consumption and disease risk is often complex where at moderate levels may be protective whereas heavy consumption significantly increases risk<sup>361,369</sup>.

Table 1-4 Global alcohol-attributable fraction for selected causes of death

Condition	Alcohol Attributable Fraction (%)
Cirrhosis	50
Pancreatitis	25
Oesophageal cancer	22
Interpersonal violence	22
Self-harm & parasuicide	22
Falls	16
Poisoning	18
Traffic injuries	15
Hepatocellular Carcinoma	12
Ischemic heart disease	7

Data from World Health Organisation, 2014<sup>466</sup>

## AT-RISK GROUPS

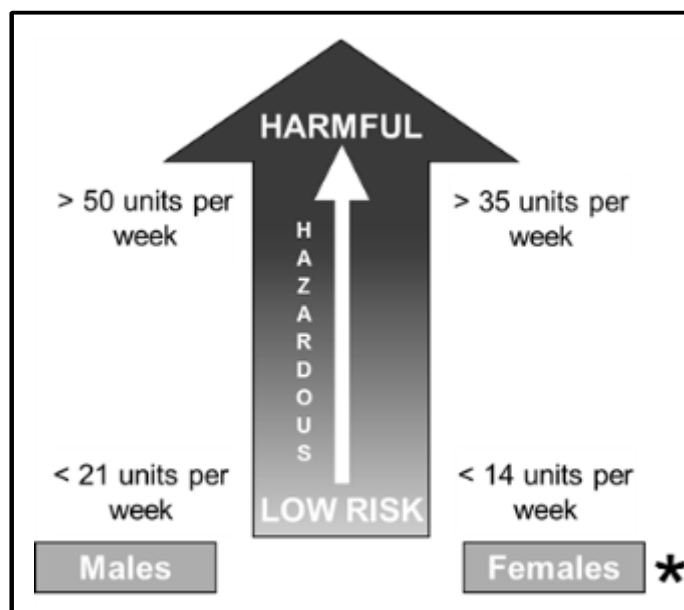
The chronic and acute harm attributable to alcohol consumption is associated with significant morbidity and mortality. Many factors mediate alcohol consumption per se, however, the harm attributable to alcohol is not equal between individuals for a given level of drinking. There is significant variety in risk between countries: higher income

countries typically consume greater amounts of alcohol per capita yet lower income countries the generally have a greater alcohol mortality for a given level of drinking<sup>466</sup>, suggesting that socio-economic factors are important. Consistent with this, in most countries, lower educational levels are associated with greater harm in both men and women<sup>154</sup>. Individuals of lower social-economic status may be at greater risk of alcohol-related harm due to less choice in the drinking environment, less access to healthcare and a less extensive support network. Individuals in certain age brackets appear to have increased risk of alcohol-related harm for a given volume of alcohol, in particular adolescents seem to be at greater risk. In the UK, factors such as younger age, socio-economic deprivation and urbanicity are all associated with increased alcohol-related mortality risk<sup>117</sup>. The factors that mediate these health inequalities are complex yet the identification of at-risk groups may be key to the minimisation of alcohol attributable harm via targeted interventions.

## **GOVERNMENTAL LEGISLATION**

Many Governments provide guidance on drinking levels with the aim of minimizing harm associated with alcohol consumption<sup>425</sup>. Beyond regulating the price or availability of alcohol directly, Governments may attempt to minimize harm by educating and informing alcohol consumers, primarily through increasing awareness of personal consumption levels and its associated harms. The concept of a standard drink, which represents a measure of an alcoholic beverage containing a fixed amount of alcohol, has been a cornerstone of public alcohol advice in several countries despite concerns that it is overly simplistic. In the UK, a standard drink contains one unit of alcohol, which roughly equates to a small glass of wine, a pint of beer or a shot of spirits.

From the 1980's onwards three bodies have advised the UK Government on recommended drinking levels associated with different levels of risk (Figure 1-6). For two decades, these recommendations stated that women should consume less than fourteen units per week and men less than twenty-one units per week, with two to three drink free day in between. In addition, daily limits of two to three units are recommended for women and three to four units are recommended for men. Finally it recommended that no more than 6 or 8 units of alcohol should be consumed at any one sitting by men and women respectively<sup>95</sup>. Recently, the UK department of health altered these guidelines following a review of current literature by a panel of experts. It now recommends that consuming no more than fourteen units of alcohol weekly spread over three days, is associated with low risk of harm in both men and women. These recommendations do not account for the known different risk profiles and metabolism profiles between men and women.



**Figure 1-6 Alcohol consumption and the risk of physical harm**

The previous recommended guidelines for low risk, hazardous and harmful drinking in the UK.

\*Pregnant women should abstain in the first trimester and then consume no more than 1 to 2 units once or twice weekly. Image adapted from Morgan et al, 2010<sup>296</sup>

### 1.3.1 - ALCOHOL USE/MISUSE

Chronic alcohol misuse typically results from the alcohol use disorders, diagnosed as alcohol misuse and alcohol-dependence. In the UK, these are the most common alcohol-related diagnoses in the National Health Service<sup>303</sup>.

#### HISTORICAL BACKGROUND

Before the late 18<sup>th</sup> century there was no concept in Western medical literature of alcohol misuse as an involuntary behaviour and instead 'drunkenness' was seen as a matter of willpower. One of the earliest proponents of a non-voluntary aspect to alcohol misuse was the American physician Benjamin Rush (1745-1813) who led a campaign in the USA warning the public of the ill-health resulting from alcohol<sup>243</sup>. The first book-length characterization of alcohol as an addiction was made by the British physician Thomas Trotter (1760-1832) in his 1804 *An Essay, Medical, Philosophical, and Chemical, on Drunkenness and its Effects on the Human Body*<sup>434</sup>. Despite this early characterization, the impact of Trotters publication on his contemporary physicians and the medical-profession was limited<sup>110</sup>. In contrast, the German-Russian physician Brühl-Cramer (?-1821)<sup>222</sup> published a popular book characterizing alcohol addiction as a disease of the mind. This work promoted the term *dipsomania*, a Greek cognate meaning *thirst-mania* that was used to define alcohol addiction well into the 19<sup>th</sup> century.

Despite the recognition by some of alcohol dependence as a disease, many more still remained of the opinion that it was a weakness and a feature of lowmindedness. The

late 19<sup>th</sup> and early 20<sup>th</sup> century saw the birth of the temperance movement as the extent of the ill effects of alcohol misuse became even more apparent. A notable outcome of this movement was periods where all alcohol consumption was illegalized, known as prohibition, most notably in the USA during 1920-1933 but in many other countries as well. During the late 19<sup>th</sup> century, scientific theories regarding the causes of alcohol dependence began to develop such as the proto-evolutionary theory of degeneration. This theory was used as a medical explanation for alcohol problems in society, which although fundamentally incorrect, laid the foundations for future studies of inheritance. However, this concept also resulted in the mass sterilization of alcohol-dependent individuals in Nazi Germany<sup>274</sup>.

After the Second World War, the modern disease conceptualization of alcohol dependence developed, following the clinical characterization of alcohol withdrawal phenomena such as delirium tremens<sup>197</sup> and the acknowledgment of this condition by large international bodies such as the WHO. This culminated in a seminal paper by Edwards and Gross<sup>111</sup> in which the modern idea of alcohol dependence was proposed as a clinical entity defined by simple uniform diagnostic criteria. Despite the wide acceptance of the disease model of alcohol dependence a few individuals in the medical profession remain critical of this model<sup>454</sup>.

## DIAGNOSIS

The diagnosis of alcohol use disorders is based on two separate but similar systems. The WHO publishes the *International Classification of Diseases (ICD)* and the American Psychiatric Association publishes the *Diagnostic and Statistical Manual of Mental Disorders (DSM)*<sup>10</sup>. These criteria are subject to revisions and there are several versions of both the DSM and ICD diagnostic criteria. Currently the 4<sup>th</sup> edition of the DSM (DSM-IV) of 10<sup>th</sup> edition of the ICD (ICD-10) are the most widely used in clinical practice and medical research.

In both the DSM-IV and ICD-10 diagnostic models, the more severe alcohol dependence criteria is separated from less severe alcohol use criteria. The less severe alcohol use criteria in the ICD-10 is called 'harmful use' whereas in the DSM-IV it is called 'alcohol abuse'. In both, this is defined as a pattern of alcohol use that causes damage to mental and/or physical health. Alcohol dependence is further defined by cravings for alcohol, tolerance to its neurobiological effects, a preoccupation with alcohol and continued alcohol consumption despite harmful consequences. The major difference is that the DSM-IV criteria have a greater emphasis on alcohol use (Table 1-5). The UK's National Institute for Health and Care Excellence guidelines state: "*Alcohol dependence is characterised by craving, tolerance, a preoccupation with alcohol and continued drinking in spite of harmful consequences (for example, liver*

*disease or depression caused by drinking*)". The recently published 5<sup>th</sup> edition of the DSM (DSM-V) has integrated alcohol misuse and alcohol dependence into a single disorder, known as alcohol use disorder, with mild, moderate and severe sub-classifications. The 11<sup>th</sup> edition of the ICD is currently undergoing revision will likely alter its criteria to become closer to those of the DSM-V<sup>438</sup>.

The accurate definition alcohol use disorders is also important for genetic and epidemiological research. The ICD-10 and DSM-IV diagnostic criteria for alcohol dependence largely overlap and have a high concordance in diagnosis<sup>152</sup>. Thus, these diagnostic criteria are interchangeable for research purposes. However, the less severe diagnosis of alcohol misuse has notable discordances between criteria<sup>168</sup>. Therefore, alcohol misuse diagnoses from the DSM-IV and ICD-10 criteria are not interchangeable for research purposes.

Table 1-5 Comparisons between the ICD and DSM diagnostic criteria for alcohol dependence

Criterion Number	ICD-10*	Criterion Number	DSM-IV*
1	Strong desire or sense of compulsion to use the substance	NA	NA
2	Impaired capacity to control use as evidenced by the substance often being taken in larger amounts or over a longer period than intended or by a persistent desire or unsuccessful efforts to control use	3	Persistent desire or one or more unsuccessful efforts to cut down or control drinking
		4	Drinking in larger amounts or over a longer period than the person intended
3	Physiological withdrawal	2	Physiological withdrawal
4	Tolerance	1	Tolerance
5	Preoccupation with substance use as manifested by important interests being given up or reduced or a great deal of time spent in activities necessary to obtain, take, or recover from the effects of the substance	5	Important social, occupational, or recreational activities given up or reduced because of drinking
		6	A great deal of time spent in activities necessary to obtain, to use or to recover from the effects of drinking
6	Persistent substance use despite clear evidence of harmful consequences	7	Continued drinking despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to be caused or exacerbated by drinking

\*In both the DSM-IV and the ICD-10 criteria, a diagnosis of alcohol dependence is made if 3 or more of the criteria are present together at some time during the previous 12 months  
Abbreviations: DSM-IV- Diagnostic and Statistical Manual of Mental Disorders, 4th Edition; ICD-10 – International Classification of Diseases, 10<sup>th</sup> Edition

## PATHOPHYSIOLOGY

It has long been realised that reinforcement, tolerance and withdrawal are key physiological mechanisms that result in substance dependence<sup>255</sup>. The consumption of alcohol results in reinforcing effects and in the longer term chronic consumption results in neuro-adaptive effects that result in tolerance and withdrawal symptoms. Unlike many other addictive drugs, alcohol does not have a single specific neurotransmitter binding partner, instead binding to a diversity of receptors in the central nervous system.

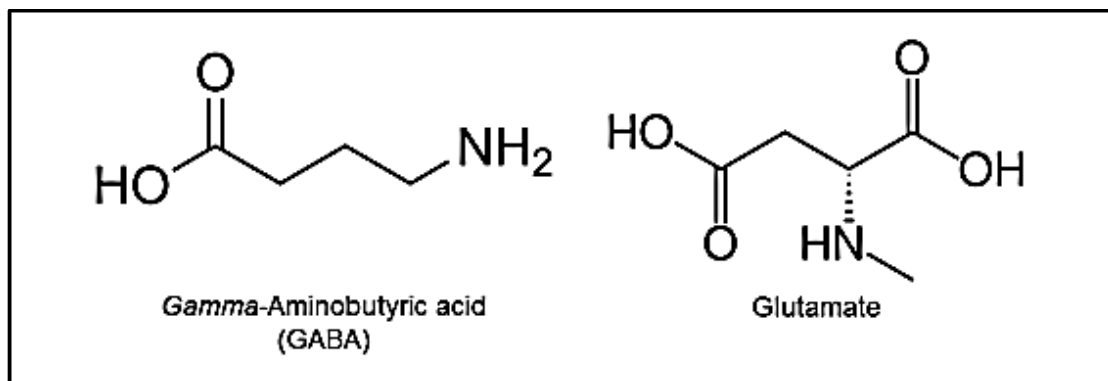
### Reinforcement

Reinforcement occurs when a stimulus induces a rewarding state such as euphoria or relieves an unpleasant state such as anxiety. Alcohol has both euphoric and anxiolytic effects that follow its consumption and these effects are likely essential for the development of alcohol dependence. Alcohol causes these effects through interactions with several neurotransmitter receptors present in the central nervous system. This diversity in interaction partners include the *gamma*-amino butyric acid (GABA) system, the glutamate system and the endogenous opioid system.

Alcohol causes many psychoactive effects through its interactions with GABA systems in the brain and central nervous system. GABA (C<sub>4</sub>H<sub>9</sub>NO<sub>2</sub>) (Figure 1-7) is the chief inhibitory neurotransmitter of the central nervous system and binds to, and activates, two classes of multi-subunit receptors. These are broadly classified as ligand-gated ion channels, known as ionotropic receptors, and G-protein-coupled receptors known as metabotropic receptors. These GABA receptors may occur in many isoforms with differing subunit constituents and thus they are pharmacologically diverse<sup>395</sup>. The sedative effects of alcohol are mediated through agonism of GABA receptors resulting in channel activation<sup>293</sup> and the influx of chloride ions into neurons, which leads to the inhibition of neuronal action potentials. A primary effect of this is sedation, which is shared with other GABA receptor agonists such as the benzodiazepines. The short-term sedative effects of alcohol consumption may reinforce alcohol drinking through relief from anxiety via interactions with GABA receptors in certain regions of the brain<sup>208</sup>.

The N-methyl-D-aspartate (NMDA) receptor is also a target of alcohol. This receptor is an ion channel protein formed of GluN1 and GluN2 hetero-tetramers, which may have multiple different isoforms. These receptors are activated by the neurotransmitter glutamate (C<sub>5</sub>H<sub>9</sub>NO<sub>4</sub>), which is the major excitatory neurotransmitter in the brain (Figure 1-7). NMDA receptors are predominantly located in the central nervous system and play key roles influencing synaptic plasticity and memory formation. Alcohol is an

antagonist of the NMDA receptor<sup>265</sup>: its binding to the NMDA receptor inhibits the binding of glutamate, in consequence preventing the activation of the receptor resulting in a decrease in the transmission of nerve impulses. The result of NMDA receptor antagonism is sedation and memory loss, a feature of acute alcohol consumption that is shared with other drugs that antagonise the NMDA receptors such as ketamine.



**Figure 1-7 The chemical structures of the neurotransmitters glutamate and GABA**

A primary reinforcing effect of alcohol consumption is the feeling of well-being or mild euphoria<sup>267</sup> that follows its immediate consumption. This effect in part results from the release of endogenous opioids in the brain where it has been shown that alcohol consumption directly promotes their release in brain regions associated with reward and decision making<sup>294</sup>. Endogenous opioids activate a physiologically important type of G-protein-coupled receptor known as opioid receptors, which play a key role in pain modulation and the regulation of behaviour. Many other addictive substances, such as morphine, directly activate opioid receptors.

### Tolerance and Withdrawal

Tolerance to the psychoactive effects of alcohol and withdrawal symptoms upon abstinence from alcohol are distinguishing features of alcohol dependence<sup>111</sup>. These phenomena occur due to neuro-adaptive processes that occur in the brain in several key neurotransmitter systems following chronic exposure to alcohol.

Chronic antagonism of the NMDA receptor by alcohol results in a neuro-adaptive and compensatory process that upregulates NMDA receptor mediated functions. Chronic alcohol consumption increases the number of NMDA receptors present in the brain<sup>132</sup>, partly counteracting the sedative effects of alcohol resulting in some tolerance to the psychological effects of alcohol. However, the increased number of receptors may be responsible for withdrawal effects, as without systemic alcohol in play the increased receptor numbers result in neuronal over-excitation that can result in seizures, hallucinations<sup>284</sup> and excitotoxicity mediated neuronal cell death<sup>193</sup>. This process may

also contribute to the neurodegeneration observed in some alcohol-dependent individuals.

## MANAGEMENT

Psychosocial intervention is key to the management of alcohol misuse and alcohol dependence and can be delivered by a variety of health care professionals in a variety of settings<sup>302</sup>. In addition and perhaps more germane to the concept of a genetic component to alcohol use is the use of pharmacotherapy to treat the symptoms of withdrawal and aid the maintenance of abstinence.

The GABA<sub>A</sub> receptor agonists, benzodiazepines are a routinely prescribed and efficacious treatment for alcohol withdrawal symptoms including seizures and delirium tremens<sup>9</sup>. Long-term benzodiazepine use may result in dependence and therefore these compounds are generally prescribed on a short-term basis with tapering doses (three to seven days). Long-acting benzodiazepines such as chlordiazepoxide or diazepam are most effective at treating alcohol withdrawal symptoms. In patients with significant liver impairment, long-acting benzodiazepines should be used with caution due to impaired metabolism and increased risk of toxicity. Another GABA<sub>A</sub> receptor agonist that is not a benzodiazepine, called baclofen, may reduce alcohol withdrawal symptoms<sup>256</sup> and have use in patients with liver disease. The opiate receptor antagonist naltrexone (naltrexone hydrochloride) is prescribed to aid the maintenance of abstinence and has been shown to have a significant but modest effect on drinking behaviour in carefully selected patients admitted to intensive treatment programmes<sup>64</sup>. There is evidence that treatment outcome relate, at least in part, to the presence of a functional variant allele in the  $\mu$ -opioid receptor gene *OPRM1*<sup>54</sup>. The functional glutamate antagonist acamprosate (calcium acetylhomotaurinate), which may be co-prescribed with naltrexone, also aids the maintenance of abstinence<sup>275</sup>. The mechanism of action of acamprosate is controversial and it may interact as either a weak agonist or antagonist of NMDA and GABA. The irreversible ALDH antagonist disulfiram (Antabuse) is prescribed to aid the maintenance of abstinence<sup>63</sup>. Antagonism of ALDH results in the build-up of toxic acetaldehyde molecules following alcohol consumption, resulting in unpleasant effects including flushing, headache, nausea, dizziness and an irregular heartbeat and thus acts to maintain abstinence via negative feedback. Disulfiram often results in over-sensitivity to other common sources of environmental alcohol.



### 1.3.2 - ALCOHOL-RELATED LIVER DISEASE

Alcohol-related liver disease covers a spectrum of changes to the liver that may result in significant liver dysfunction and potentially death. The stages of alcohol-related liver disease are not necessarily distinct and may occur simultaneously in a given individual, including steatosis, steatohepatitis, cirrhosis and hepatocellular carcinoma.

#### HISTORICAL BACKGROUND

Many ancient cultures attributed importance to the liver as evidenced in ancient mythical and religious texts. The ancient Babylonians for example, characterized the basic anatomical structures of sheep livers as a result of their divination practice of hepatoscopy and thus they are credited with characterization of the liver's basic anatomy (Figure 1-8).



**Figure 1-8 Babylonian clay liver model(circa 1900 BC)**

Clay liver models were used by specialist priests, and were used as a reference source for the interpretation of individual markings on a sheep liver when predicting the future. Image obtained from British Museum<sup>2</sup>

From the 5<sup>th</sup> century BC until the fall of the Roman Empire in the 5<sup>th</sup> century AD, many Greek and Roman scholars studied the liver as epitomised by the prominent Greek physician, surgeon and philosopher Claudius Gaenus, better known as Galen of Pergamon (AD129-C200/216). Galen, was a prominent figure in the Roman Empire and characterized the anatomy of the liver, where he distinguished the hepatic veins and arteries, and provided theories on how bile and blood travelled through the liver. Despite clear evidence of liver diseases existing in the ancient world<sup>481</sup>, knowledge of liver pathology was limited<sup>341</sup>; at best, Hippocrates (460 -370 BC) recognised ascites, jaundice and cysts of the liver<sup>408</sup>.

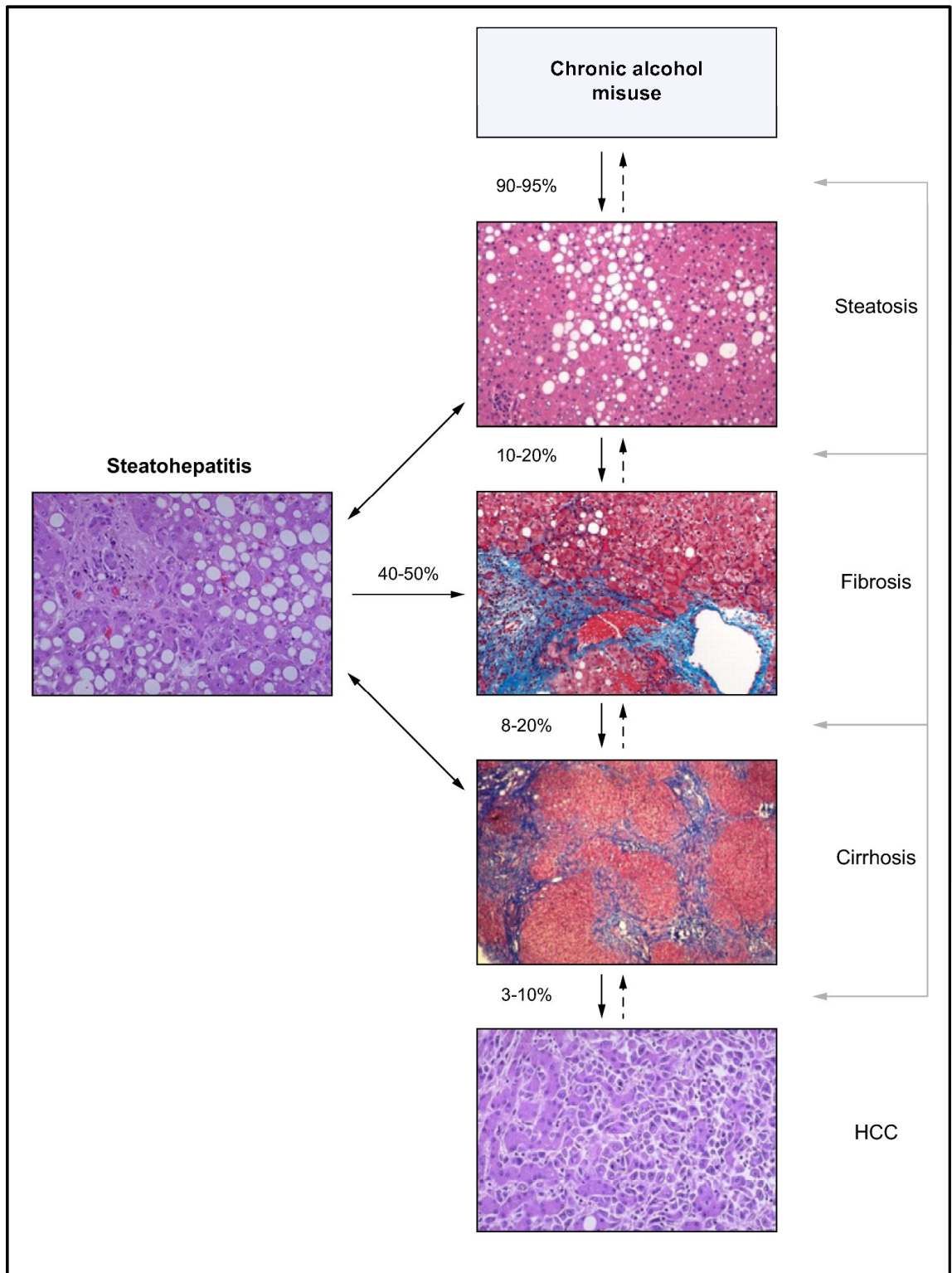
The invention of the microscope in the 14<sup>th</sup> century resulted in a considerable expansion of the understanding of the detailed pathology and pathophysiology of the liver. Many of the findings from this period are summarized in the detailed work

*Anatomis hepatis* by the English physician Francis Glissons (1599-1677)<sup>144</sup>. For example during this period, the disrupted architecture of the liver caused by fibrous bands of tissues was discovered. The French physician René Laennec (1781 –1826) coined the term cirrhosis, which was used to describe this pattern of fibrotic tissue deposition. It is a neologism derived from the Greek *kirrhós* meaning "yellowish, tawny" referring to the orange-yellow colour of the diseased liver and the suffix *-osis*, meaning condition. Later, during the 19<sup>th</sup> century, many of the individual cell types in the liver were distinguished including the hepatocytes, Kupffer cells and hepatic stellate cells.

It was not until the 19<sup>th</sup> century that the pathological consequences of excess alcohol consumption on the liver were first identified. The English physician Matthew Baillie (1761-1823)<sup>18</sup> noted "*the process ... [of cirrhosis]... is commonly produced by a long habit of drinking spirituous liquors*". Another English physician, Thomas Addison (1793-1860) noted the frequent coincidence of heavy alcohol consumption and the deposition of fat in the liver<sup>3</sup>. In the early 20th century the American physician Frank Burr Mallory (1862-1941)<sup>271</sup> described the inclusion bodies that bear his name, which are a feature of alcoholic steatohepatitis. Hepatology as a speciality in its own right developed during the 20th century but was not recognised as a subspecialty of medicine until the mid-20th century and at that point, more interest was focused on the role of alcohol consumption in liver disease.

## STAGES

Alcohol-related liver disease is a broad term for a spectrum of abnormalities, which may occur following the consumption of alcohol at harmful levels for prolonged periods of time. The spectrum includes steatosis characterized by the accumulation of lipid droplets in the cytosol of hepatocytes; steatohepatitis characterized by inflammation and cell death; cirrhosis characterized by disruption of the architecture by fibrous bands and the presence of regenerative nodules; and hepatocellular carcinoma, which develops within the cirrhotic liver. The occurrence of these liver lesions is not mutually exclusive and they may co-exist. Not everyone who drinks heavily develops significant liver injury and its evolution is significantly modified by subsequent drinking behaviour (Figure 1-9).



**Figure 1-9 The stages of alcohol-related liver disease**

Each image represents typical microscopic features seen on biopsy. The arrows represent an estimate of the number of individuals that will progress through each stage of alcohol-related liver disease to more advanced forms of the disease. Image adapted from the EASL clinical practice guidelines, 2012<sup>285</sup>

## Steatosis

A normal, healthy liver has a lipid content of ~0.5-1.5%. Alcohol consumption results in the deposition of lipids in the liver and the vast majority of individuals that consume more than 60 g of alcohol daily will develop steatosis<sup>74,109</sup> (Figure 1-9). In severe steatosis the liver's fat content may be as high as 50%. It has been demonstrated that steatosis is more prevalent in individuals who are obese with alcohol-use disorders, compared to non-obese individuals with alcohol-use disorders<sup>27</sup>. Steatosis may rapidly reverse when alcohol consumption is reduced. However, the presence of steatosis influences the progression to more severe forms of liver disease<sup>404</sup>.

## Steatohepatitis and Fibrosis

In a subset of individuals, steatosis may be associated with inflammation (Figure 1-9). This is identified as the histological lesion of alcoholic steatohepatitis, which is characterized by the ballooning degeneration, necrosis and apoptosis of hepatocytes in association with neutrophilic infiltration, cholestasis and the formation of Mallory bodies. Ballooning degeneration occurs when lipids, proteins and water are retained in hepatocytes. Mallory bodies are proteinaceous deposits, which may be histologically evident in the damaged hepatocytes. The recurrent generation and resolution of inflammation leads to the accumulation of fibrous tissue and over time, this may result in the development of cirrhosis.

## Cirrhosis

Cirrhosis is histologically defined by characteristic features including the disruption of the liver's normal architecture and its replacement with fibrous bands of collagen and the presence of regenerative nodules. At a histological level, these features are graded into three types based on the size and homogeneity of the nodules: micronodular, macronodular and mixed-nodular cirrhosis.

## Hepatocellular carcinoma

The development of hepatocellular carcinoma (HCC) is considered part of the natural history of alcohol-related liver injury. However, in alcohol-related liver disease it generally only develops in patients with long-standing alcohol-related cirrhosis and in this respect differs from other liver disease such as that caused by hepatitis B and non-alcohol-related fatty liver disease (NAFLD) where tumours may arise in non-cirrhotic livers. The tumours arise in discrete nodules most likely centred around dysplastic rests and most tumours are likely to be multifocal<sup>227</sup>.

## CLINICAL FEATURES

The early stages of alcohol-related liver injury have few clinically evident manifestations and are therefore asymptomatic. If liver injury is detected at this stage, it is invariably because patients have presented either because of problems associated with the alcohol consumption per se or because of the incidental finding of abnormal liver function tests when they present with another condition. Liver function tests are a battery of biochemical markers that although they may be abnormal in the presence of liver injury, are non-specific (Table 1-6). Many factors contribute to variability in liver enzyme levels, which may confound the accuracy of these tests<sup>238</sup> and therefore more specific testing usually follows up abnormal readings. Some liver function tests may be normal even in the presence of advanced alcohol-related liver disease.

The clinical manifestations of alcohol-related liver disease become apparent after the liver has been significantly injured: these stages clinically present as severe alcoholic hepatitis and cirrhosis.

The clinical symptoms of cirrhosis are largely dependent on the degree to which the liver is functionally impaired. All of these symptoms relate to the development of (i) hepatocellular failure, which is associated with a gradual decline in excretory and synthetic function of hepatocytes, and (ii) the development of portal hypertension as the blood draining into the liver from the splanchnic circulation is increasingly impeded because of presence of significant amounts of fibrotic tissue in the liver. Hepatocellular failure results in the development of jaundice, disturbed blood clotting, low serum albumin and the failure to detoxify the blood adequately. Portal hypertension results in an enlarged spleen, low haemoglobin levels, low white blood cell and platelet counts and the development of collateral blood vessels most notably in the stomach and oesophagus, which can rupture and bleed. The combination of hepatocellular failure and portal hypertension results in the development of fluid retention with peripheral oedema and ascites and the development of the neuropsychiatric syndrome of hepatic encephalopathy.

Alcoholic hepatitis is a severe, if less common, cause for presentation with alcohol-related liver disease. This term is applied to the rapid deterioration in liver function, in either recently abstinent or the ongoing alcohol misuser resulting from inflammatory processes in the liver. The clinical symptoms of alcoholic hepatitis include characteristic jaundice as well as other features of hepatic failure such as bruising, gastrointestinal bleeding, ascites and hepatic encephalopathy. Alcoholic hepatitis generally has a rapid onset, and therefore the misnomer 'acute' is often applied to this clinical syndrome as a prefix, although the rapid onset invariably represents the clinical presentation of an already established, albeit undiagnosed, alcohol-related cirrhosis<sup>340</sup>.

The histological features of severe alcoholic hepatitis are often the same as those present in the histological picture known as steatohepatitis. However, only a minority of alcohol misusers with histological steatohepatitis<sup>304</sup> present with clinically syndromic alcoholic hepatitis and therefore these terms are not synonymous.

Table 1-6 Liver function test biomarkers and factors that may confound their interpretation

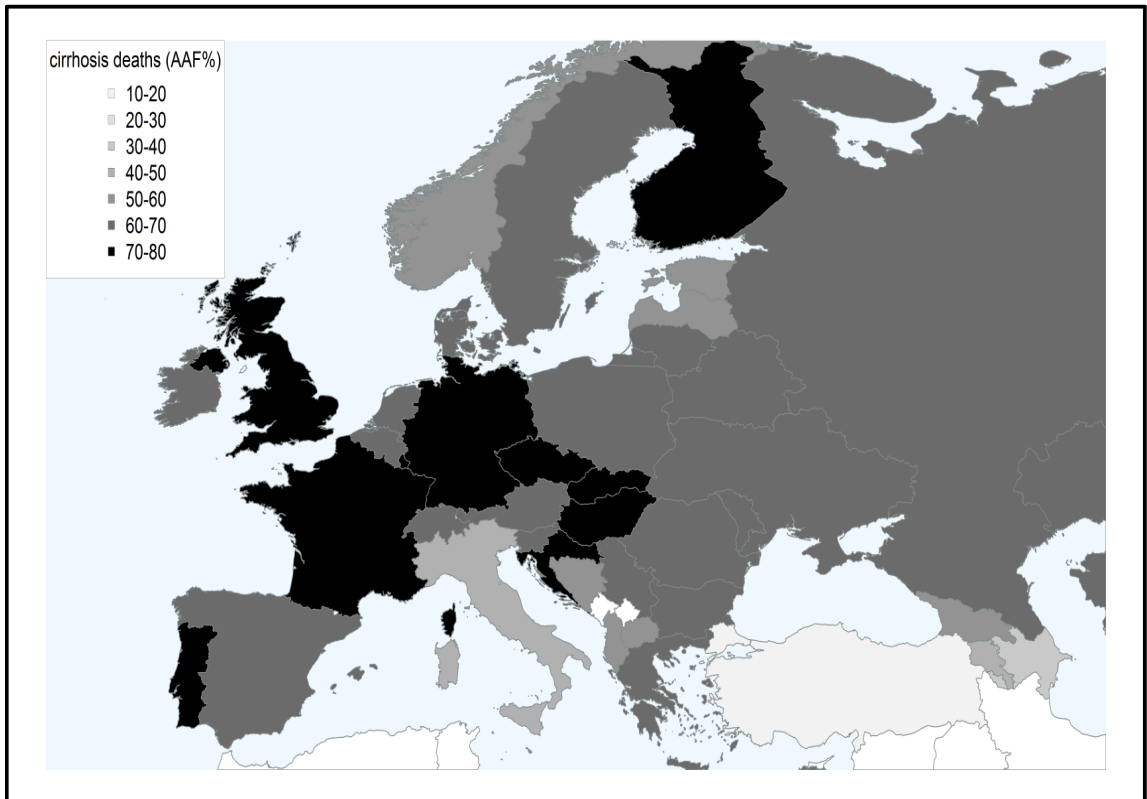
Marker	Relating to Alcohol-related Liver Disease	Confounders
Gamma-glutamyl transferase	A key metabolic enzyme found predominantly in liver tissue. Elevated levels indicate prolonged heavy drinking but not necessarily alcohol-related liver disease	Numerous – other causes of liver damage any cause of intra or extra-hepatic biliary obstruction and numerous enzyme- inducing drugs
Aspartate amino transferase	A key metabolic enzyme found in hepatocytes and many other cell types. Elevated levels are associated with cell death but in alcohol misusers are indicative of excessive alcohol consumption and not necessarily alcohol-related liver injury as they may reflect alcohol-related muscle damage	Cell necrosis in cardiac, muscle, pancreatic and renal tissues also cause elevated levels
Alanine amino transferase	A key metabolic enzyme found predominantly in hepatocytes. Elevated levels in the blood indicates liver cell necrosis	A good marker for liver injury – but not specific to alcohol
Bilirubin (direct and indirect)	A breakdown product of haemoglobin. Circulating levels can increase for several reason including increased loading following a gastrointestinal bleed; failure of hepatic clearance in decompensated disease and haemolysis in the presence of liver failure and portal hypertension	Not specific for alcohol-related liver disease. Unconjugated hyperbilirubinaemia is observed in approximately 10% of otherwise healthy men with <i>Gilberts Syndrome</i> *
Albumin	A plasma protein produced and secreted by the liver. Low levels are observed in decompensated alcohol-related cirrhosis reflecting a failure of synthetic function	Hypalbuminaemia may develop with cirrhosis of any aetiology and is not specific for alcohol –related disease; Hypoalbuminaemia is also a feature of malnutrition and other conditions characterized by fluid retention such as heart and renal failure
Prothrombin time & International Normalised Ratio	Vitamin - K dependent clotting factors are made in the liver. Synthetic failure may arise in decompensated alcohol-related cirrhosis resulting in elevation of the prothrombin time and prolongation of the international normalised ratio	Any form of severe acute liver disease or cirrhosis; anticoagulant drugs e.g. warfarin, dietary vitamin k deficiency and certain genetic disorders
Alkaline phosphatase	A key enzyme produced by cells lining the biliary ducts of the liver as well as salivary gland, intestinal , bone and placental tissue. Elevated levels can occur in alcohol-related liver disease when there is a degree of intra-hepatic cholestasis, but also as a result of parotid enlargement and alcohol-related bone disease - isoenzyme fractionation will help determine the source	Can be elevated in other forms of liver disease particularly cholestatic conditions; elevation is also seen in bone disease; certain intestinal disorders and during pregnancy

\*Gilbert's syndrome results from a defect in the uridine diphosphate glucuronosyltransferase (UGT1A1) gene, which results in a reduction in the processing of bilirubin for excretion from the body

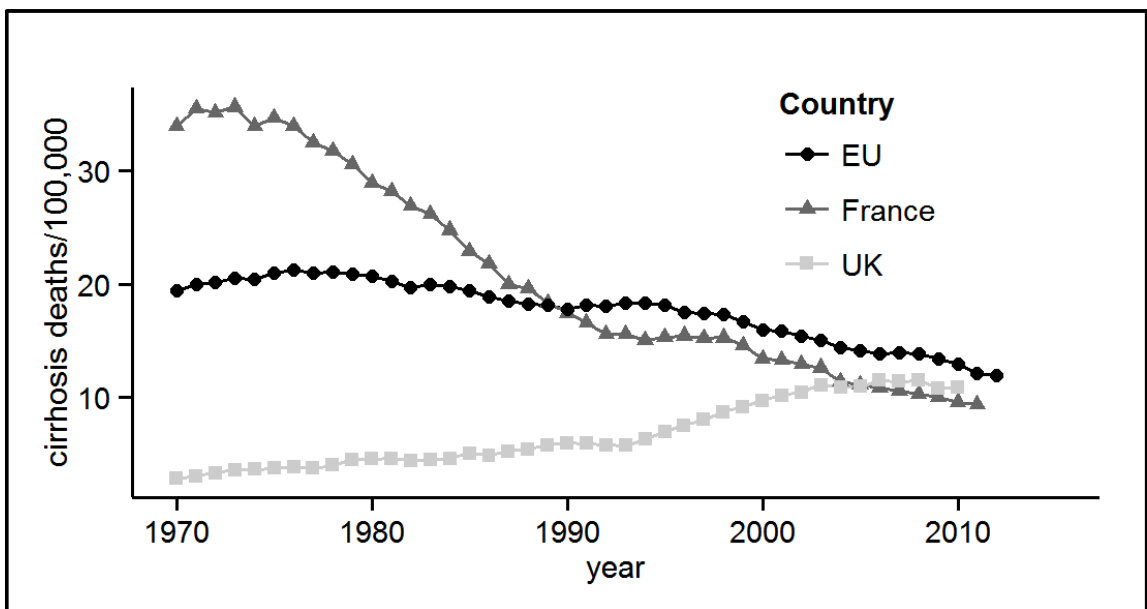
## EPIDEMIOLOGY

Steatosis can develop after drinking significant amounts of alcohol over a few days<sup>253</sup>, and hence is near universally present in active alcohol misusers. Steatosis alone is asymptomatic, and for this reason is largely unapparent in many epidemiological studies of alcohol-related liver disease. A retrospective study<sup>423</sup> has shown that 37% of alcohol misusers with histologically confirmed steatosis develop cirrhosis/fibrosis within approximately 10 years if drinking (>40 units/week) continues. Hepatic decompensation resulting from cirrhosis is the most common reason for alcohol-related liver disease to become symptomatic, and accordingly is the primary morbidity, or primary cause of mortality, recorded by epidemiological studies. The incidence rate of cirrhosis in alcohol misusers varies between studies<sup>239,241</sup> although its incidence is probably no more than 20%. The development of cirrhosis substantially increases the risk of developing hepatocellular carcinoma, and approximately 15% of with patients with cirrhosis go on to develop it<sup>379</sup> at a yearly incidence of 1.4%<sup>123</sup>. The clinical syndrome of alcoholic hepatitis is an uncommon cause for alcohol-related liver disease presentation and there is little information regarding its incidence. At clinical presentation approximately half will have cirrhosis<sup>313</sup>. The histological features steatohepatitis are more common, although in most instances are asymptomatic.

Alcohol-related cirrhosis is a major cause of global mortality accounting for around 1% of the total deaths, nearly half a million deaths, in the year of 2010<sup>361</sup>. Deaths attributable to alcohol-related cirrhosis account for nearly half of the total deaths resulting from cirrhosis of all aetiologies<sup>360</sup>. Alcohol is the major cause of cirrhosis in most parts of the world, with the exception of North Africa and the Middle East. In Europe, alcohol is the main cause of cirrhosis and total cirrhosis-related deaths (Figure 1-10). There are large national differences within Europe, with the highest cirrhosis death rates seen in Central and Eastern Europe. In the majority of Western European countries death rates from cirrhosis are falling, yet in the United Kingdom (Figure 1-11) cirrhosis death rates are on the increase. Within the UK, there are regional differences in cirrhosis death rates within the UK where the standardized cirrhosis death rates in Scotland are almost twice as high as those in both Wales and England<sup>242</sup>.



**Figure 1-10 The alcohol attributable fraction for cirrhosis deaths in Europe**  
 Data source: World Health Organisation, 2010<sup>465</sup>



**Figure 1-11 Standardized cirrhosis mortality rates**  
 In contrast to other nations in Europe, and in particular France, which is relatively equivalent to the UK in terms of gross domestic product, levels of development and healthcare, the annual cirrhosis death rates in the UK have increased in comparison to decreases elsewhere. This trend may reflect increasing alcohol consumption in the UK. Abbreviations: EU – European Union; UK – United Kingdom. Data source: World Health Organisation, 2005<sup>464</sup>



## PROGNOSIS

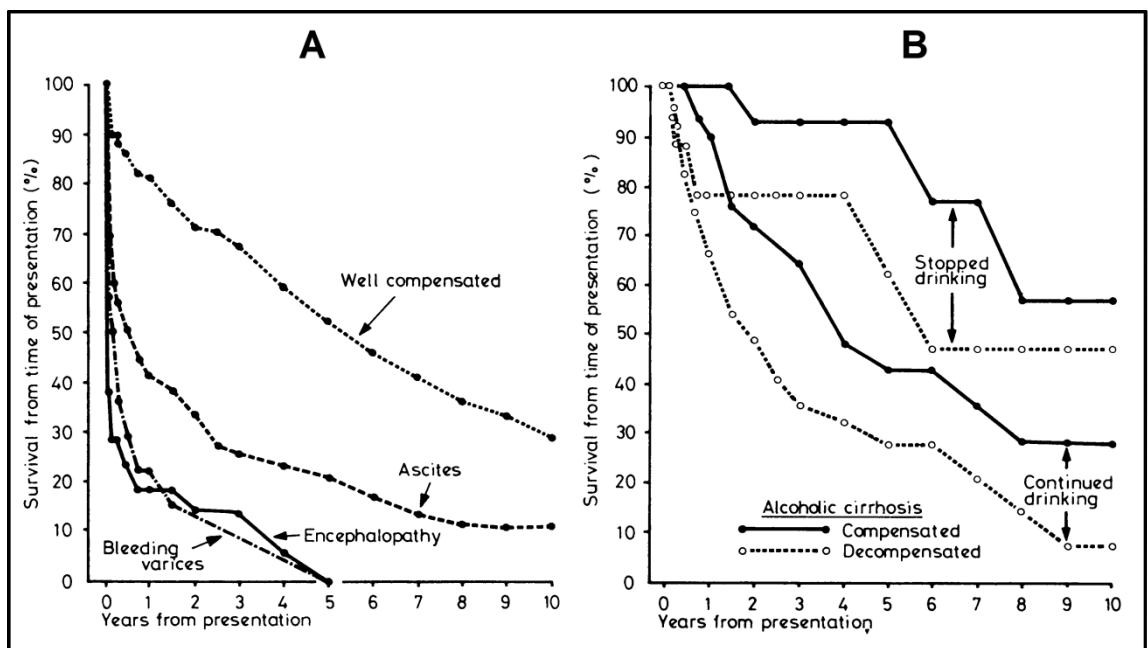
The prognosis of alcohol-related liver disease is largely dependent on the disease stage, the level of liver dysfunction and whether abstinence is maintained subsequent to diagnosis.

Following the development of steatosis, ten year survival rates range from 46% to 72%<sup>82</sup>. The development of steatosis is an indicator of liver damage and a significant proportion of individuals with steatosis will go on to develop more severe forms of alcohol-related liver disease such as cirrhosis. A Danish study following 106 patients who were diagnosed with alcohol-related steatosis found that 22 (20.7%) developed cirrhosis<sup>80</sup> within 11 years and that mortality was significantly increased in comparison to a reference population; out of the 79 deaths that occurred in the patients with steatosis, 20 (25.3%) were liver related. With abstinence, steatosis is reversible and the risk of developing further stages of alcohol-related liver disease is diminished.

Alcoholic hepatitis presents rapidly and has a poor short-term prognosis, being associated with significant mortality. As a cause of alcohol-related liver disease, it is less common than cirrhosis but its true prevalence is unknown. The histological features of steatohepatitis may be present in 10-35% of hospitalised alcohol-dependent patients<sup>66</sup>. In a large cohort (n=1092) of clinically diagnosed alcoholic hepatitis patients from the UK that were recruited for the steroids or pentoxifylline for alcoholic hepatitis trial, from the date of presentation, after twenty-eight days 16% of patients had died; after ninety days 29% of the patients had died; and, after one year 56% of patients had died<sup>427</sup>. Common causes of death, or other serious adverse events, in the STOPAH cohort included gastrointestinal disorders (13% of patients) such as upper gastrointestinal haemorrhage (4% of patients) or ascites (3% of patients), liver failure (9% of patients), infections (15% of patients) and renal disorders (4% of patients). Several prognostic models have been derived specifically to predict short-term survival in patients with alcoholic hepatitis<sup>340</sup>. The most widely used of these models is the modified Maddrey's discriminant function<sup>268</sup>. Identified predictors of mortality include histological features such as the degree of parenchymal neutrophil infiltration, biochemical features such as serum bilirubin levels and blood clotting activity, and host factors such as age and the presence of symptoms of significant liver dysfunction. In patients that survive the short term, abstinence is the key predictor of disease progress and mortality; this is more likely to happen in women and in individuals in whom the initial histological findings upon presentation with alcoholic hepatitis were classified as severe<sup>324</sup>.

The prognosis in alcohol-related cirrhosis is variable and dependent on the degree of hepatic decompensation and whether abstinence is maintained post-diagnosis (Figure

1-12). Consequently, the presentation with cirrhosis, without compromised liver function, and with the maintenance of abstinence post diagnosis is relatively favourable with ten year survival rates of 47%<sup>142</sup>. However, the presentation with cirrhosis with clinical complications of liver dysfunction such as ascites, hepatic encephalopathy and bleeding oesophageal varices has a considerably less favourable prognosis (Figure 1-12). In a UK based study containing 100 patients with biopsy-determined alcohol-related cirrhosis, the seven year survival rates from a post-cirrhosis diagnosis were 72% in abstainers and 44% for those that continued drinking<sup>444</sup>. A primary treatment option for chronic liver failure resulting from cirrhosis is orthotopic liver transplantation. When used as a treatment option for liver failure resulting from alcohol-related cirrhosis, the prognosis is favourable and is equivalent to survival in patients that have undergone transplantation due to liver failure resulting from different aetiologies; the 1-year post-transplantation survival rates following liver failure resulting from alcohol-related cirrhosis are 66%<sup>35</sup>. Major causes of death in alcohol-related cirrhosis include gastrointestinal haemorrhage, renal failure, liver failure and HCC<sup>25,379</sup>.



**Figure 1-12 Survival from clinical presentation with alcohol-related cirrhosis**

The cumulative survival in a cohort of patients with alcohol-related cirrhosis is mediated by their clinical features at presentation (image A) and whether abstinence is maintained subsequent to a diagnosis of cirrhosis (image B). Image modified from Saunders et al., 1981<sup>379</sup>

## Prognostic Scoring Systems

Hepatic decompensation may be quantified by several scoring algorithms such as the Child-Turcotte-Pugh score<sup>65,348</sup> (CTP) and the model for end-stage liver disease score<sup>213</sup> (MELD); these are commonly used in clinical practice and medical research.

The CTP scoring system determines the level of hepatic decompensation based on a five point system involving two clinical and three laboratory variables, namely the presence and severity of ascites and hepatic encephalopathy and the plasma total bilirubin, and serum albumin concentrations and the prothrombin time or INR. Each variable is scored from 1 (least) to 3 (maximum) severity. Pugh's scores of 5 and 6 are classified as Child's Grade A or compensated; 7 to 9 as Child's Grade B or mildly decompensated and 10 to 15 as Child's Grade C or severely decompensated. For purposes of statistical analysis patients with Pugh's score 5 to 7 are classified as compensated while patients with Pugh's scores of 7 to 15 are classified as decompensated. The primary limitation of the CTP scoring system is the reliance on clinical assessment resulting in some inconsistency in its scoring.

The MELD scoring system overcomes the clinical assessment limitation of the CTP scoring system and only requires information from laboratory tests, namely, the serum bilirubin levels, the international normalised ratio and serum creatinine levels and the aetiology of liver disease. The score generated by the MELD algorithm typically ranges from  $\leq 5$  to  $\geq 40$  where a higher score denotes liver dysfunction. Typically a MELD score  $\leq 15$  in cirrhosis indicates compensation whereas a score  $>15$  indicates decompensation<sup>285</sup>.

## **PATHOPHYSIOLOGY**

The pathogenesis of alcohol-related liver disease is multifactorial where complex interactions between the host and environment mediate liver injury. Alcohol is clearly the causative agent driving liver injury yet despite this knowledge, the molecular mechanisms driving disease pathogenesis remain unclear. The identified mechanisms are largely based on those pathways that have been characterized in animal or cellular models.

Chronic alcohol consumption results in the deposition of lipids in hepatocytes. Alcohol may cause this effect through several mechanisms (Figure 1-13). First, the metabolism of alcohol in the liver is known to significantly alter the ratio of NADH to its oxidized equivalent<sup>231</sup>. This alteration in the redox state of the liver likely promotes lipid accumulation via increasing lipid synthesis and inhibiting lipid oxidation. Second, alcohol inhibits the key regulator of cellular metabolism, adenosine monophosphate-activated kinase (AMP-K)<sup>186</sup>. Inhibition of AMP-K results in the upregulation of

peroxisome proliferator activated receptors (PPAR) and the downregulation of sterol regulatory element binding proteins (SREBP), which also results in hepatic lipid accumulation. There are many other potential mechanisms of hepatic lipid accumulation in alcohol-related liver disease all relating to impairments in key lipid processing pathways<sup>350</sup> (Figure 1-13).

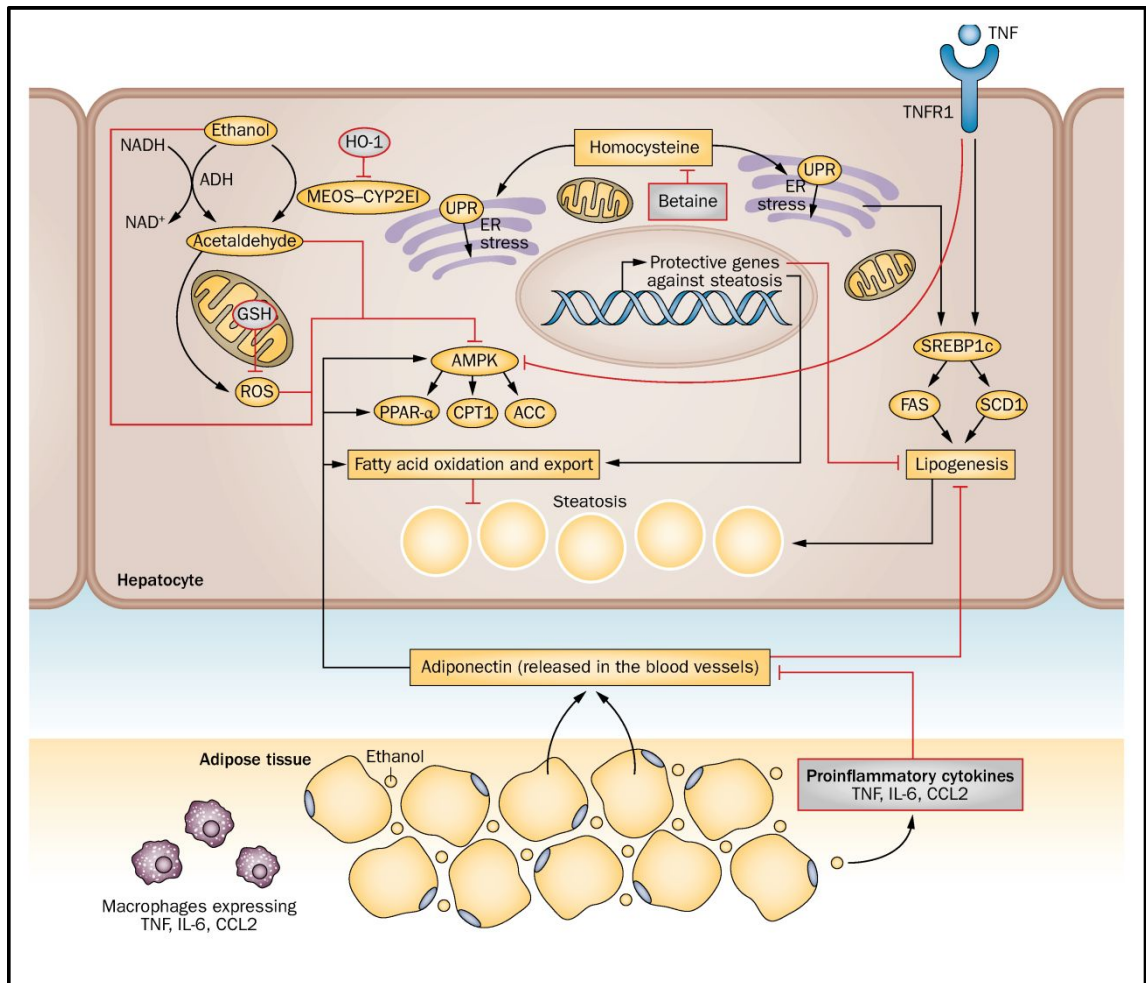
Further stages of alcohol-related liver disease involve direct injury to the liver via an upregulation of immune responses through several pathways (Figure 1-14). The metabolism of alcohol itself may cause inflammation through the generation of acetaldehyde and reactive oxygen species. Such compounds may react with proteins forming antigenic adducts upregulating an immune response. Alcohol consumption also upregulates immune responses both by increasing gut permeability and by altering the growth and diversity of bacteria present in the gut, allowing small bacterial cell wall components, known as endotoxins, to enter peripheral circulation via the portal venous system<sup>354</sup>. It has been demonstrated that the levels of endotoxins in peripheral circulation are elevated following heavy alcohol consumption<sup>38</sup>. Endotoxins are strong activators of Küppfer cells, a specialised macrophage type found in the liver. Physiologically, Küppfer cells eliminate bacteria and foreign material in the blood. However, interactions between endotoxins and the Toll-like receptor 4 (TLR4) found on the surface of Küppfer cells results in cellular activation and inflammation. In animal models, the loss of *TLR-4* results in insensitivity to endotoxin and also alcohol-induced liver injury<sup>439</sup>. A primary effect of Küppfer cell activation is the secretion of cytokines such as tumour necrosis factor alpha (TNF- $\alpha$ ). Cytokines promote acute inflammation and activate apoptosis, resulting in a signalling cascade causing the recruitment of neutrophils into the liver. Neutrophils increase liver tissue damage by inducing the necrotic death of hepatocytes.

Inflammation and tissue death frequently resolve, in part, due to liver regeneration. The liver is characterized by its high capacity for regeneration, although repeated tissue injury results in a reduced regenerative capacity. In the later stages of alcohol-related liver disease the loss of hepatocyte proliferative capacity is associated with a worsening disease<sup>92</sup>; this process may involve cellular senescence and telomere shortening<sup>453</sup>.

Cirrhosis is characterized by the significant functional impairment of the liver due to the presence of scar tissue or fibrosis and a loss of hepatocyte regenerative capability. The primary cell-type responsible for the deposition of fibrotic tissue is the hepatic stellate cell (HSC). Physiologically, HSCs function in the storage of fat soluble vitamins, such as vitamin A, and extra-cellular matrix homeostasis via the secretion of matrix metalloproteinases (MMP). These enzymes degrade extra-cellular matrix

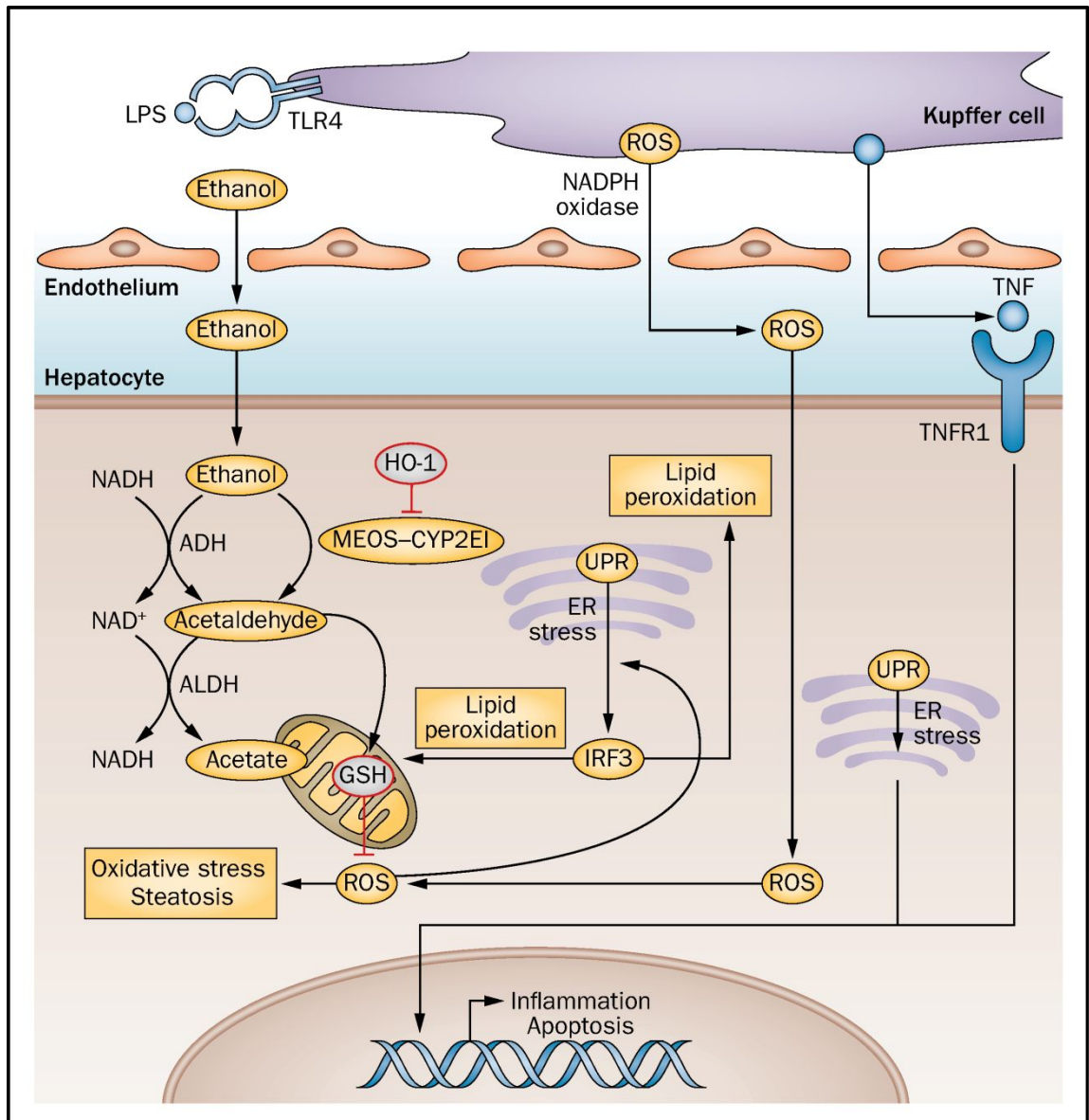
proteins, and are thus closely regulated by tissue inhibitors of metalloproteinases (TIMP) under physiological conditions. Liver injury promotes the transition of HSCs into a myofibroblast-like state, where collagen synthesis is upregulated and, matrix degradation by MMPs is downregulated resulting in fibrosis. Several other changes also occur following HSC activation including HSC proliferation and the increased migration of HSCs to the site of liver injury.

Cirrhosis significantly predisposes to the development of HCC<sup>403</sup>. HCC, like all cancers requires, in part, genomic alteration in key oncogenes that result in the uncontrolled cell cycles. There are several potential mechanisms for hepatocellular carcinogenesis in the cirrhotic liver<sup>113</sup>. First, the metabolism of alcohol produces acetaldehyde and other reactive oxygen species; these are known mutagens and thus may cause somatic mutations resulting in hepatocyte carcinogenesis. Second, liver damage induces regeneration via cell replication; the additional cell cycles induced by regeneration causes telomere shortening, which may increase chromosomal instability, allowing the genomic alterations necessary for carcinogenesis. Third, hepatocyte proliferation is impaired in cirrhosis; this may allow for loss of replicative competition in the liver by healthy hepatocytes, thus promoting tumour formation. Finally, the altered micro and macro environment of the cirrhotic liver may be directly carcinogenic in itself as it may contain a combination of unique growth factors and other promoters of tumour formation.



**Figure 1-13 Mechanisms of alcohol-related steatosis**

Chronic alcohol consumption leads to steatosis via generation of acetaldehyde, ROS and ER stress. The consequences are blockade of PPAR $\alpha$  and of AMPK, which is responsible for fatty acid oxidation and export via ACC and CPT-1. In addition, chronic alcohol consumption induces SREBP1c activation, which is responsible for fatty acid synthesis through FAS and fatty acid desaturation through SCD1. TNF also leads to steatosis by activating SREBP1c. Betaine reduces homocysteine levels, which thus enhances ER stress. Chronic alcohol consumption also induces adipose tissue inflammation, which decreases the release of the protective adiponectin, thus favouring steatosis. All these mechanisms lead to disruption of hepatic lipid metabolism by increasing lipogenesis and decreasing fatty acid oxidation and export. Abbreviations: ACC - acetyl CoA carboxylase; AMPK - adenosine monophosphate-activated protein kinase; CCL2 - CC-chemokine ligand 2; CPT1 - carnitine palmitoyltransferase 1; ER - endoplasmic reticulum; FAS - fatty acid synthetase; PPAR- $\alpha$  - peroxisome proliferator-activated receptor  $\alpha$ ; ROS - reactive oxygen species; SCD1 - stearoyl-CoA desaturase 1; SREBP1c - sterol regulatory element-binding protein 1c; TNFR1 - tumour necrosis factor receptor 1. Figure adapted from Louvet et al., 2015<sup>264</sup>



**Figure 1-14 Metabolism of alcohol in the hepatocyte and mechanisms of cell injury**

Alcohol dehydrogenase and MEOS mainly metabolize alcohol into acetaldehyde, which is responsible for the generation of ROS. ROS cause oxidative stress, ER stress and steatosis. ROS are also generated through activation of NADPH oxidase in Kupffer cells. All these changes in hepatocyte metabolism lead to inflammation and apoptosis. Abbreviations: ADH - alcohol dehydrogenase; ALDH - aldehyde dehydrogenase; CYP2E1 - cytochrome P450 2E1; ER - endoplasmic reticulum; GSH - reduced glutathione; HO-1 - haem oxygenase 1; IRF3 - interferon regulatory factor 3; LPS - lipopolysaccharide; MEOS - microsomal ethanol oxidation system; NAD - nicotinamide adenine dinucleotide; ROS - reactive oxygen species; TNFR1 - tumour necrosis factor receptor 1; TLR-4 - Toll-like Receptor 4; UPR - unfolded protein response. Figure adapted from Louvet et al., 2015<sup>264</sup>

## 1.4 - ALCOHOL AND GENETICS

### 1.4.1 - ALCOHOL USE/MISUSE

As a phenotype for genetic study, alcohol use/misuse may be investigated under a variety of phenotypic models. The diagnosis of alcohol dependence is the most widely studied qualitative phenotype as it has uniform diagnostic criteria and is the most severe alcohol use phenotype. Alcohol consumption is the most widely studied quantitative alcohol-use phenotype as this information is routinely collected in large epidemiological datasets.

#### **ENVIRONMENTAL AND HOST-MEDIATED RISK FACTORS**

The consumption of alcohol from early adolescence until early adulthood seems to be a crucial risk period for the development of later alcohol dependence. Beginning alcohol consumption in early adolescence (11-14) may be an important risk factor for the development of alcohol dependence as there is a strong correlation between the age of alcohol consumption onset and the risk of developing alcohol dependence within 10 years<sup>98</sup>. During this period of life, the drinking behaviour of others including the familial group, friends and peers is an important mediator of alcohol consumption levels and hence risk of developing alcohol dependence. Parental alcohol problems and high trait anxiety are significant risk factors for alcohol dependence during this period<sup>339</sup>.

The factors that mediate differences in alcohol-dependence risk by gender, are complex. It is generally agreed on that constitutional differences and hence differences in alcohol metabolism between males and females are one of the major factors. There are also social factors that may influence gender differences in alcohol dependence risk. Role theory suggests that traditional male gender role encourages alcohol consumption whereas alcohol consumption is discouraged as part of the traditional female gender role<sup>365</sup>. This effect appears to diminish when traditional gender roles are displaced. It is clear that women who are more educated or have higher socio-economic status are more likely to consume alcohol at hazardous levels<sup>314</sup>. It is thought that this effect is partly due to increased affluence and hence the relative affordability of alcohol but it may be the case that women with higher economic and educational status are less likely to underreport drinking behaviour.

There are many other environmentally mediated risk factors such as negative life events, chronic ill health, spouse stressors and an increased number of friends who approve of drinking<sup>42</sup>. Identifying risk factors for alcohol misuse and dependence is confounded by the complex interplay between the host and environment.



## HERITABILITY AND GENETIC RISK FACTORS

The earliest genetic studies of alcohol use quantified the contribution of environmental and genetic factors by measuring heritability and familiarity of alcohol-use phenotypes. Heritability is the proportion of the total variation of a phenotype between individuals in a population that is due to non-environmental genetic variation within a population, whereas familiarity is a measure of how often a phenotype tends to occur within families. There are three types of study design that have been applied to the heritability and familiarity of alcohol-use phenotypes: twin studies, family studies and adoption studies.

Twin studies involve comparisons between dizygotic, and monozygotic twins where, based on the laws of inheritance, it is assumed that additive genetic risk is completely shared between monozygotic twins while, in contrast, dizygotic twins will only share half of their additive genetic risk. Using this information, concordance rates of a phenotype such as alcohol dependence can be calculated by twin-type and statistically compared. A higher concordance of the phenotype in monozygotic twins than dizygotic twins suggests that genetic variation is contributing to the phenotype and may be used to calculate a heritability estimate. Several large twin studies of alcohol dependence and other alcohol-use phenotypes have been performed, with heritability estimates for alcohol dependence ranging from 30% to 70%<sup>5</sup> (Table 1-7). Such studies have demonstrated differences in heritability by gender, although this finding may relate to the smaller on average sample sizes of female twin-study cohorts. A potential confounder of the twin study design is the assumption of equal environments<sup>320</sup>.

Family studies statistically quantify whether a phenotype is present in related members of a family more than would be expected by chance. Alcohol dependence is clearly a familial phenotype as the siblings of alcohol-dependent individuals have a higher likelihood of alcohol-dependence than would be expected from a random population<sup>363,461</sup>. For example, in the USA the siblings of an individual with alcohol dependence have a lifetime chance of becoming dependent themselves of 49.7% for male siblings and 22.4% for female siblings<sup>32</sup>; this rate is significantly higher than in control siblings. By design, family studies measure the familiarity of a phenotype and provide no information regarding the heritability of a phenotype as family members are typically exposed to the same environmental factors.

Adoption studies are able to exclude the confounding shared environment by comparing the concordance between the phenotypes of adopted away offspring with the phenotypes of biological and adoptive parents. Heritability may be estimated if the concordance in phenotype between adopted away offspring and biological parents is higher than would be expected by chance. There are no adoption studies that solely

focus on the alcohol dependence phenotype per se instead focusing on less specific alcohol misuse phenotypes. The three major adoption studies of alcohol-use phenotypes are the Danish Adoption Cohort, the Stockholm Adoption Cohort and the Iowa Adoption Cohort (Table 1-8) with the majority of analyses demonstrating moderate heritability estimates.

Twin, family and adoption studies provide considerable evidence for significant familiarity and heritability of the alcohol dependence and alcohol misuse phenotypes. Despite this, many of the individual studies are non-comparable due to differences in phenotypic classification or analytical methodology between studies. Heritability estimates often differ by gender and are typically larger in cohorts that only contain males. A large meta-analysis of over fifty twin, family and adoption studies of alcohol misuse phenotypes indicated that heritability estimates of alcohol dependence exceeding 40% to 60% may be inflated and that there is significant heterogeneity in the results of most studies with the true heritability estimate closer to 30% to 36%<sup>446</sup>. It is clear that both genetic and environmental risk factors, and their interplay, have a substantial role in alcohol-dependence and alcohol-misuse phenotypes.

Table 1-7 Twin studies of alcohol-dependence

Location	Diagnostic criteria	Monozygotic twins			Dizygotic twins			HE (%)	Source
		n	Sex	C (%)	N	Sex	C (%)		
Sweden	Chronic alcoholism	27	M	71	60	M	32	72	Kaij, 1960 <sup>210</sup>
Finland	Alcoholism	172	M	26	557	M	12	30	Partanen, 1966 <sup>326</sup>
UK	Alcoholism	15	M	33	20	M	30	8	Gurling, 1981 <sup>160</sup>
		13	F	8	8	F	13	-16	
USA	Alcoholism	271	M	26.3	444	M	11.9	36	Hrubec, 1981 <sup>185</sup>
Finland	Alcoholism	69	M	13	175	M	5.7	24	Koskenvuo, 1984 <sup>228</sup>
		7	F	0	20	F	0	0	
Sweden	Alcoholism	95	M	12.6	187	M	9.1	14	Allgulander, 1991 <sup>8</sup>
USA	DSM-III AD	39	M	59	47	M	36.2	46	Pickens, 1991 <sup>333</sup>
		24	F	25	20	F	5	54	
USA	Alcohol dependence	203	F	26.2	154	F	11.9	34	Kendler, 1994 <sup>218</sup>
USA	DSM-IV AD	378	M	31.7	436	M	19.3	28	Prescott, 1999 <sup>342</sup>
Australia	Alcohol dependence	396	M	38.9	231	M	19.9	40	Heath, 1997 <sup>172</sup>
		932	F	20.9	534	F	9.2	22	
USA	Alcoholism	710	M	53.2	588	M	43.2	20	True, 1996 <sup>435</sup>
USA	Alcohol dependence	28	M	40	26	M	13	48	Prescott, 2005 <sup>343</sup>
		48	F	17	58	F	24	10	

Abbreviations: n – number, C – concordance, HE – heritability estimate, M – male, F – female, UK – United Kingdom, USA – United States of America, AD – alcohol dependence, DSM-III – diagnostic and statistical manual of mental disorders third edition.

Table modified from Walters, 2002<sup>446</sup>

Table 1-8 Adoption studies of alcohol use phenotypes

Cohort	Diagnostic Criteria	Proband subjects			Control subjects			HE (%)	Year	Source
		<i>n</i>	Sex	Outcome	<i>n</i>	Sex	Outcome			
Danish Adoption Cohort	Alcoholism	55	M	18.2	78	M	5.1	42	1973	Goodwin, 1973 <sup>147</sup>
	Problem drinking	55	M	9.1	78	M	14.1	-16	1973	
	Alcoholism	6	F	33.3	90	F	52.2	-18	1977	Goodwin, 1977 <sup>148</sup>
Stockholm Adoption Cohort	Alcohol abuse	89	M	39.4	892	M	13.1	42	1978	Bohman, 1978 <sup>40</sup>
	Alcohol abuse	172	F	7	741	F	2.6	20	1981	Bohman, 1981 <sup>41</sup>
	Severe alcohol abuse	307	M	7.8	555	M	4.9	12	1981	Cloninger, 1981 <sup>67</sup>
	Alcohol abuse	108	M	24.1	469	M	12.8	24	1996	Sigvarsson, 1996 <sup>396</sup>
	Alcohol abuse	114	F	0.9	546	F	1.3	-2	1996	
Iowa Adoption Cohort	Alcoholism	23	M	13	69	M	1.4	52	1980	Cadoret, 1980 <sup>47</sup>
	Alcohol abuse	39	B	48.7	404	B	13.9	5	1986	Cadoret, 1986 <sup>49</sup>
	Alcohol abuse	49	B	70.6	34	B	55.1	32	1994	Cadoret, 1994 <sup>48</sup>

Abbreviations: *n* – number, HE – heritability estimate, M – male, F – female, B – both.

Table modified from Walters, 2002<sup>446</sup>

## GENOME-WIDE STUDIES

There are two main approaches for the identification of risk loci/genes from the human genome: linkage studies and genome-wide association studies.

The earliest technique to be developed was the linkage study, which is typically performed in phenotypically well-characterized individuals from a single or several families in which the phenotype of interest is familial. This technique relies on the genotyping of hundreds of variants dispersed across the genome to capture the patterns of crossing-over during the recombination events of meiosis. Linkage occurs when the alleles of two more variants co-occur (i.e. are non-randomly inherited) in members of the family that have the phenotype of interest. Linkage is statistically quantified by calculating the logarithm of odds ratio score (LOD), a statistical estimate of whether two genes, or a gene and a disease, are linked to one another. A LOD score greater than 3 is indicative of linkage between a genomic region and a phenotype<sup>297</sup>. For phenotypes with Mendelian patterns of inheritance, this technique allows the accurate identification of causal genes. For complex disease phenotypes,

this technique has limitations, especially when the phenotype of interest is of adult onset such as alcohol dependence.

Several genome-wide linkage scans have been performed when studying alcohol use and alcohol dependence (Table 1-9). The genomic regions implicated by studies in alcohol dependence have been inconsistent with modest LOD scores suggesting that risk variants for alcohol dependence are of small effect size. However, the genomic locations with the modest linkage evidence contain plausible candidate loci such as the alcohol-dehydrogenase gene cluster, and the GABA receptor gene cluster.

Table 1-9 Genome-wide linkages studies of alcohol use phenotypes

Phenotype	Families	Regions*	Candidate Genes in region	Ethnicity	Cohorts	Country	Source
Alcohol dependence	713	—	—	EUR	UCSF	USA	Gizer, 2011 <sup>143</sup>
Alcohol dependence symptoms	1690	—	—	EUR	—	Australia	Hansell, 2010 <sup>165</sup>
Alcohol dependence	238	10q23.3–24.1	—	AFR	—	USA	Gelernter, 2009 <sup>138</sup>
Alcohol dependence; Alcohol Use	474	4q22-32	—	EUR	IASPSAD	Ireland	Prescott, 2006 <sup>344</sup>
Alcohol dependence	158	—	—	EUR	SMOFAM	USA	Wilhelmsen, 2005 <sup>456</sup>
Consumption	330	—	—	EUR	Framingham Heart Study	USA	Wyszynski, 2003 <sup>468</sup>
Alcohol dependence	100	~chr4	ADH1B	NA	—	USA	Ehlers, 2004 <sup>112</sup>
Alcohol dependence	172	chr11p, chr4p	ADH cluster, TH	NA	—	USA	Long, 1998 <sup>262</sup>
Alcohol dependence	105	—	—	EUR	—	USA	Reich, 1998 <sup>364</sup>

Abbreviations: SMOFAM – Smoking in Families Study; IASPSAD – Irish Affected Sib-Pair Study of Alcohol Dependence, UCSF – University of California San Francisco, EUR – European, AFR – African, NA – Native American, USA – United States of America. \*LOD score > 3

Genome-wide association studies (GWAS) are the paradigm successors of linkage studies. GWAS involve the extensive parallel genotyping of hundreds of thousands of genomic markers, typically single nucleotide polymorphisms (SNPs), which cover the majority of common genetic variation across the human genome. GWAS are performed in large, generally unrelated, populations in which either qualitative or quantitative phenotypic data have been collected. These data are analysed for significant differences in allele or genotype frequencies resulting in genetic association when an allele is associated with a phenotype at a specified significance threshold.

A typical GWAS involves the statistical testing of hundreds of thousands of partially independent genetic variants for association with a phenotype. This multiple testing increases the likelihood of false positive association signals. To account for this, a variant is typically deemed significantly associated with a phenotype when the significance value is less than a certain threshold ( $P < 5 \times 10^{-8}$ ), known as the genome-wide significance threshold. The genome wide significance threshold equates to there being 1 million independent tests when accounting for common genetic variation and its linkage disequilibrium patterns<sup>105</sup>. Before a variant is deemed associated with a phenotype or trait it generally requires independent verification through replication analysis in a separate population or cohort<sup>59</sup>. It follows that the numbers of samples needed for an effective GWAS are very large.

Several GWAS of alcohol-use and alcohol-dependence phenotypes have been undertaken (Table 1-10). Many of these have been collaborative studies, predominantly from the USA, in which large cohorts of alcohol-dependent cases and population controls have been collected. In most instances the cohorts recruited in the USA have been phenotypically heterogeneous, containing samples of multiple different, or mixed, ancestries with a high proportion of the cohort having comorbidities such as drug dependence<sup>32</sup>. Many of the smaller GWAS of alcohol dependence have failed to identify genome-wide significant associations. Meta-analyses and studies in populations with greater phenotypic surety have identified genome-wide significant associations between variants in genes responsible for alcohol metabolism, for example *ALDH2* and *ADH1B* in East Asian ancestry populations and *ADH1B* and *ADH1C* in European, African and East Asian ancestry populations.

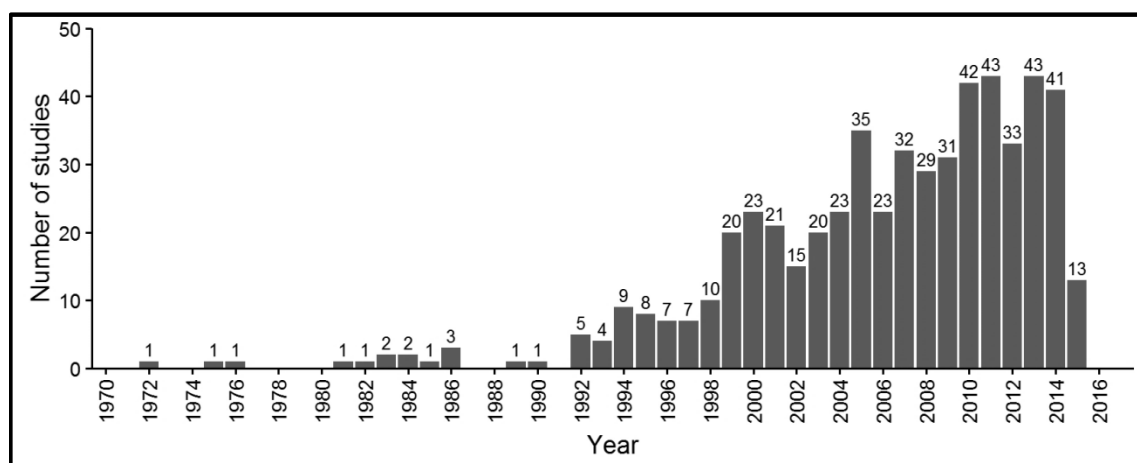
Table 1-10 Genome-wide association studies of alcohol use phenotypes

Pheno	Cases		Controls		Genes*	Ethnicity	Cohorts	Country	Source
	<i>n</i>	Sex	<i>n</i>	Sex					
AD	7,677	B	6,992	B	<i>ADH1B, ADH1C, METAP, PDLIM5</i>	EUR, AFR	GCD, SAGE	USA	Gelernter, 2014 <sup>139</sup>
AD	621	B	750	B	<i>ADH1B, ALDH2</i>	EA	-	South Korea	Park, 2013 <sup>325</sup>
AD	1,333	M	2,168	M	<i>ADH1B-ADH1C</i>	EUR	-	Germany	Frank, 2012 <sup>130</sup>
AD	2,090	B	2,026	B	<i>KIAA0040</i>	EUR, AFR	SAGE, COGA	USA	Zuo, 2012 <sup>483</sup>
CONS	2,834	M	-	-	<i>ALDH2</i>	EA	-	South Korea	Baik, 2011 <sup>17</sup>
CONS	47,501	B	-	-	<i>AUTS2</i>	EUR	AlcGen	Several EU Nations	Schumann, 2011 <sup>384</sup>
AUD	2,062	B	3,393	B	-	EUR	OZALC	Australia	Heath, 2011 <sup>173</sup>
AD/ND	1,823	B	2,763	B	-	EUR	NESDA, OZALC	Holland, Australia	Lind, 2010 <sup>254</sup>
AD	1,192	B	692	B	-	EUR, AFR	COGA	USA	Edenberg, 2010 <sup>108</sup>
AD	1,897	B	1,932	B	-	EUR, AFR	SAGE	USA, Germany	Bierut, 2010 <sup>34</sup>
AD	1,511	M	2,354	M	<i>PECR</i>	EUR	-	Germany	Treutlein, 2009 <sup>433</sup>

Abbreviations: Pheno – phenotype, AD – alcohol dependence, CONS – alcohol consumption, AUD – alcohol-use disorder, N – number, M – Male, B – Both males and females, EUR – European, AFR – African, EA – East Asian, COGA – Collaborative study on the genetics of alcoholism, SAGE - Study of Addiction: Genetics and Environment, OZALC - Australian twin-family study of alcohol use disorder, NESDA – Netherlands Study of Depression and Anxiety, AlcGen – Alcohol-GWAS consortium, GCD – GWAS discovery samples, USA – United States of America. \*P value  $\leq 5 \times 10^{-8}$

## CANDIDATE GENE STUDIES

To date there have been over three-hundred candidate genetic association studies of alcohol-dependence and related phenotypes (Figure 1-15). The majority of candidate gene studies have failed to identify statistically significant genetic association or have been inconsistent<sup>7</sup> on replication analysis; this failure to replicate is a notable feature of the candidate gene approach<sup>179</sup>. Candidate gene studies have identified some replicable associations predominantly between functional variants in gene encoding enzymes responsible for alcohol metabolism such as rs1229984 (Arg48His) in *ADH1B*<sup>245</sup> in European<sup>33</sup>, African<sup>33</sup> and East Asian<sup>245</sup> ancestry populations and rs671 (Glu504Lys) in *ALDH2* in East-Asian<sup>246</sup> ancestry populations. The combination of genetic and functional analyses has proven complementary for the study of candidate genes, as is the case for the several GABA genes, which encode subunits of GABA receptor. The likely importance of variation in GABA genes is highlighted by positive genetic associations in *GABRA2*<sup>107</sup> and the strong phenotypic effects on alcohol consumption resulting from functional mutations in the murine orthologue *Gabrb1*<sup>13</sup>.



**Figure 1-15 The number of genetic association studies of alcohol dependence**

The number of journal articles for a given year identified in the PubMed database<sup>305</sup> searching for genetic association studies of alcohol dependence in humans excluding GWAS and meta-analyses. Pubmed search term: “genetic association study Alcohol dependence NOT (GWAS OR Family OR Linkage OR Meta-analysis)”

### 1.4.2 - ALCOHOL-RELATED LIVER DISEASE

Alcohol-related liver disease is precipitated by alcohol misuse<sup>241</sup>. However, there is considerable variation in individual risk likely relating to a combination of genetic and environmental factors. Case/control genetic association studies of alcohol-related cirrhosis must avoid confounding by comparing groups that have experienced the same environmental exposure (i.e. long-term harmful alcohol consumption) and either

have established alcohol-related cirrhosis or else have no liver disease. Both groups are difficult to identify with certainty, as this would usually require examination of liver histology, which is uncommon in clinical practice. This may partially explain why the study of alcohol-related liver disease genetics has received less attention than the study of genetics in other forms of liver disease.

## ENVIRONMENTAL FACTORS

Increasing alcohol consumption is the principle contributory epidemiologic factor in alcohol-related cirrhosis<sup>241</sup>, although there are equivocal findings regarding the degree to which the pattern of consumption and the amount of alcohol consumed contribute to risk. Leibel's seminal work demonstrated a dose-response relationship between alcohol consumption and cirrhosis risk<sup>240</sup>. A number of large prospective cohort studies<sup>24,28,214,226,404</sup> and a systematic review<sup>362</sup>, have since validated this relationship from which it seems likely that above a threshold alcohol dose there is a dose-dependent increase in risk. Findings between studies are highly variable, which may relate to the difficulties in determining accurate alcohol consumption levels and differences in the ascertainment of alcohol-related cirrhosis (e.g. clinical diagnosis, biopsy, cause of death). There is a greater degree of contention as to whether the pattern of consumption (e.g. regular daily drinking or irregular bingeing) influences risk. Animal models would suggest that irregular bingeing is more harmful to the liver<sup>394</sup> and some epidemiological studies demonstrate findings that are in agreement with this<sup>85</sup>. However, many more epidemiological studies find the opposite of this<sup>15,240,404</sup> (i.e. daily consumption is associated with increased risk of cirrhosis). Studies of drinking patterns may be confounded by failing to account the total amount of alcohol consumed when directly comparing regular and irregular drinkers<sup>359</sup>, as is the case for all but one<sup>15</sup> study.

There are many other potential environmental factors, which relate either directly or indirectly to alcohol consumption. Two large prospective studies have identified that the type of alcoholic beverage, namely the consumption of wine, is associated with reduced cirrhosis risk in comparison to the consumption of beer<sup>15,23</sup>. The consumption of alcohol with food may also reduce the risk for developing cirrhosis<sup>28</sup> and a high fat and a low carbohydrate and protein diet may increase risk<sup>373</sup>. Several studies have suggested a link between coffee consumption and protection from the development of cirrhosis<sup>219</sup>. Smoking also seems to be an independent risk factor for the development of alcohol-related cirrhosis<sup>81,226</sup>. Many risk factors have been associated, although determining causality is more challenging due to the potential for confounding.

Constitutional differences influence the risk of developing alcohol-related cirrhosis. It is well established that women are at an increased risk of developing alcohol-related



cirrhosis<sup>295</sup>, even when differences in alcohol consumption levels are accounted for<sup>437</sup>. This gender difference relates to the on average proportionally higher body fat content composition of women. Body-fat content in both genders, is associated with an increased risk of developing alcohol-related cirrhosis<sup>304</sup>; this is potentially due to a synergistic interaction between alcohol consumption and weight<sup>167</sup>. It seems likely that the features of the metabolic syndrome that drive the progression of NAFLD, may also contribute to alcohol-related liver disease pathogenesis<sup>357</sup>. A similar feature may also occur in other forms of liver disease, where, in combination with alcohol consumption infection with hepatitis C or hepatitis B virus is associated with a synergistic interaction to increase risk<sup>298</sup>.

## HERITABILITY AND GENETIC RISK FACTORS

Epidemiological studies of alcohol dependence heritability and familiarity are limited to a few studies. A single cohort has been used to estimate the heritability of alcohol-related cirrhosis in study of a population of 15,924 male twin pairs<sup>185,194</sup> from which it was determined that the concordance of alcohol-related cirrhosis was three times higher in monozygotic twins (14.6%) than dizygotic twins (5.4%). A subsequent analysis of this large twin dataset data, with over a decade of additional clinical follow-up has replicated this finding<sup>358</sup>. Between both analyses, heritability estimates ranged from 21% to 67%. Otherwise, there is a single familial study of alcohol-related cirrhosis<sup>452</sup> in which patients with alcohol-related cirrhosis were more than twice as likely to self-report that their father had liver disease and alcohol problems than control alcohol misusers without significant liver injury.

There are marked differences in alcohol-related cirrhosis risk based on ethnicity. In the United Kingdom non-Muslim men of South Asian ancestry present with alcohol-related cirrhosis at a younger age and at a higher than expected frequency in comparison to their white British counterparts living in the same area<sup>102</sup>. In the United States, a similar observation has been observed where individuals of Hispanic ancestry present with alcohol-related cirrhosis up to 10 years before white/Caucasian populations when controlling for confounding issues<sup>244</sup>.

Perhaps the best existing paradigm for the role of a genetic variation in the context of alcohol-related cirrhosis risk is for the variant rs671, which encodes a non-synonymous substitution (Glu504Lys) in *ALDH2*. The 504Lys allele of rs671 results in a catalytically inactive ALDH2 enzyme that has a limited capacity to clear acetaldehyde, resulting in the so-called 'alcohol flushing reaction'. Carriers of this allele may become averse to alcohol consumption and hence the possession of this allele is highly protective against the development of alcohol-related cirrhosis. In a meta-analysis of all published studies a significant and robust association was identified between possession of this

variant and the development of several alcohol-related harms including cirrhosis ( $P_{META} = 6 \times 10^{-19}$ , Odds Ratio (OR) = 0.25, 95% confidence interval (CI) [0.19–0.34])<sup>246</sup>. Although this variant is associated with protection from alcohol-related cirrhosis there are several notable features of this association that are salient to the complexities of this phenotype: first, this variant only occurs in individuals of East Asian ancestry and thus only influences risk in these populations; second, this variant mediates alcohol consumption per se rather than alcohol-related liver disease and thus indirectly protects against alcohol-related liver disease; and finally, the variant also influences the risk of other types of injury related to harmful alcohol use. This strong genetic association at ALDH2 highlights the potential for confounding in genetic association studies of alcohol-related liver disease where a genetic variant, which directly influences alcohol consumption, may indirectly associate with the alcohol-related liver disease.

Based on the wide variety in inter-individual risk for developing alcohol-related cirrhosis and the heritability of this phenotype shown in twins, it is widely held that alcohol-related cirrhosis risk is modulated through polygenic and complex inheritance in the presence of environmental risk factors<sup>86,411</sup>.

## GENOME-WIDE STUDIES

There have not been any genome wide linkage studies or GWAS of alcohol-related liver cirrhosis or related phenotypes. This contrasts with all other major types of liver disease, in which over thirty GWAS have been performed sometimes with different aspects of the disease studied (Table 1-11). From these studies of other forms of liver disease, it is clear that genetic variation plays a substantial role and may impact clinically relevant aspects of disease progression. For example, a variant in *IL28B* has large effect on antiviral treatment response and may have clinical utility in determining treatment options<sup>136,355,417,421</sup>. Similarly, in NAFLD<sup>12</sup> a major finding has been the robust association between the variant rs738409 in *PNPLA3* with several aspects of the disease phenotype including steatosis, fibrogenesis and the development of HCC. In most cases, the loci identified via GWAS have been followed by replication genotyping in independent cohorts and functional studies to understand mechanisms of pathogenesis.

Table 1-11 Genome wide association studies of liver disease phenotypes

References	Phenotype	Summary
Romeo et al., 2008 <sup>370</sup> Chalasani et al., 2010 <sup>52</sup> Speliotes et al., 2011 <sup>405</sup> Kawaguchi et al., 2012 <sup>217</sup> Kitamoto et al., 2013 <sup>225</sup> Feitosa et al., 2013 <sup>124</sup> Kozlitina et al., 2014 <sup>229</sup>	Non-alcohol-related fatty liver disease	Several GWAS have examined features of NAFLD including liver fat content; histological NAFLD and other clinically determined phenotypes. The variant rs738409 in <i>PNPLA3</i> has been robustly associated with NAFLD and related phenotypes. Several other loci have been identified with varying degrees of support. The functional variant rs58542926 in <i>TM6SF2</i> from an extended LD block has been validated as a risk variant in several studies
Mells et al., 2011 <sup>289</sup> Liu et al., 2010 <sup>260</sup> Nakamura, 2012 <sup>301</sup> Liu et al., 2012 <sup>258</sup>	Primary biliary cirrhosis	These studies have been performed in large European and East-Asian ancestry cohorts. The majority of identified loci are involved in innate immunity, including: HLA region, <i>STAT4</i> , <i>DENND1B</i> , <i>CD80</i> , <i>IL7R</i> , <i>CXCR5</i> , <i>TNFRSF1A</i> , <i>CLEC16A</i> and <i>NFKB1</i>
Kamatani et al., 2009 <sup>212</sup> Mbarek et al., 2011 <sup>286</sup> Nishida et al., 2012 <sup>310</sup> Zhang et al., 2010 <sup>477</sup> Jiang et al., 2013 <sup>199</sup> Li et al., 2012 <sup>248</sup> Chan et al., 2011 <sup>57</sup> Chen et al., 2013 <sup>62</sup> Qu et al., 2016 <sup>351</sup> Png et al., 2011 <sup>338</sup>	Hepatitis B	A number of features of hepatitis B infection have been investigated via GWAS in primarily East-Asian and European ancestry cohorts, including: chronic infection, hepatitis B related HCC, and vaccine response. Studies of chronic infection are limited to a single GWAS in Asian ancestry populations, identifying risk and protective haplotypes at the <i>HLA-DP</i> locus. A greater number of GWAS have investigated hepatitis B related HCC in in diverse global populations validating loci including 1p36.22, <i>STAT4</i> , <i>HLA-DG</i> genes. In an Indonesian cohort, the genes implicated in vaccine response are largely found in the gene rich HLA Class III interval
Ellinghaus et al., 2013 <sup>114</sup> Folseraas et al., 2012 <sup>127</sup> Melum et al., 2011 <sup>290</sup> Liu et al., 2013 <sup>260</sup>	Primary sclerosing cholangitis	Primarily in European ancestry cohorts, several primary GWAS and re-analyses have confirmed over 10 established risk loci including the HLA region, 3p21, 2q35, <i>IL21</i> , <i>CARD9</i> , <i>REL</i> , <i>GPR35</i> and <i>TCF4</i> . These analysis reveal a significant shared genetic risk with the comorbidity ulcerative colitis
de Boer et al., 2012 <sup>88</sup>	Autoimmune hepatitis	Only a single GWAS has been performed in this less common form of liver disease. The several identified loci: <i>HLADRB1*0301</i> , <i>HLA-DRB1*0401</i> , <i>SH2B3</i> , <i>CARD10</i> , overlap with those identified in other forms of autoimmune liver disease
Rauch et al., 2010 <sup>355</sup> Kumar et al., 2011 <sup>232</sup> Hoshida et al., 2012 <sup>183</sup> Patin et al., 2012 <sup>328</sup> Miki et al., 2013 <sup>441</sup> Ge et al., 2009 <sup>136</sup> Tanaka et al., 2009 <sup>421</sup> Suppiah et al., 2009 <sup>417</sup>	Hepatitis C	Several features of hepatitis C infection have been investigated in primarily East-Asian and European ancestry cohorts, including: HCV induced cirrhosis, HCV induced HCC, viral clearance in chronic HCV and antiviral treatment response. There is a strong role of genetic variants in the HLA region. Variants in <i>MICA</i> may influence HCV progression to HCC. Variants in <i>MERTK/TULP1</i> associate with fibrogenesis and variants in <i>SAG</i> fibrosis score. The variant rs12979860 <i>IL28B</i> strongly predicts antiviral treatment response and may have utility in clinical decision-making.

Abbreviations: HCC – Hepatocellular Carcinoma; NAFLD – Non-alcohol-related Fatty Liver Disease LD – Linkage Disequilibrium; HLA – Human Leukocyte Antigen; HCV – Hepatitis C Virus

## CANDIDATE GENE STUDIES

A number of candidate gene studies have been performed in cohorts of patients with alcohol-related liver disease (Table 1-12). The selection of candidate genes has most often resulted from hypotheses based on known biological mechanisms of liver injury and almost invariably studies have genotyped protein altering variants (e.g. non-synonymous or promoter variants). These include variants in genes implicated in: alcohol metabolism<sup>482</sup> such as alcohol dehydrogenase (*ADH1B*, *ADH1C*), aldehyde dehydrogenase (*ALDH2*), or cytochrome P450 (*CYP2E1*); fibrogenesis such as matrix metalloproteinases<sup>412</sup> (*MMP3*); the deposition of lipids such as apolipoprotein  $\epsilon$ <sup>175</sup> (*APOE*), phosphatidylethanolamine N-methyltransferase<sup>207</sup> (*PEMT*) or microsomal triglyceride transfer protein (*MTTP*); DNA damage and carcinogenesis such as X-ray repair cross-complementing protein 1<sup>372</sup> (*XRCC1*); the accumulation of iron in the liver such as the human hemochromatosis protein<sup>158</sup> (*HFE*); immune response such as cluster of differentiation 14<sup>195</sup> (*CD14*), tumour necrosis factor alpha<sup>156</sup> (*TNF $\alpha$* ), cytotoxic T-lymphocyte-associated protein 4<sup>442</sup> (*CTLA4*) or the nuclear factor kappa-B subunit<sup>280</sup> (*NFKB1*); and, oxidative stress such as mitochondrial superoxide dismutase 2<sup>90</sup> (*SOD2*) or glutathione S-transferase<sup>282</sup> (*GSTM1*, *GSTT1*, *GSTP1*).

Most of the genetic association studies of alcohol-related liver disease have proven negative, or negative following meta-analysis (Table 1-13). At least three genes that have undergone meta-analysis may contain genuine risk variants: the promoter variant rs361525 (-238) in *TNF $\alpha$* , which is associated with alcohol-related liver disease when compared population controls<sup>281</sup>; the *GSTM1* null allele, which is associated with alcohol-related liver disease when compared with no-liver disease controls<sup>282</sup>; and, heterozygosity for the variant rs1800562 (Cys282Tyr) in *HFE* and risk of developing HCC in alcohol-related cirrhosis when compared with alcohol-related cirrhosis patients with no HCC<sup>200</sup>. Several factors may explain the inconsistency: first, the small sample sizes of early studies are likely to have limited statistical power; second, the differences in phenotypic criteria (e.g. fibrosis, cirrhosis, liver fat, inflammation) and different comparisons (e.g. population controls or no liver disease controls) between studies; and finally confounding resulting from genetic associations with alcohol-use phenotypes<sup>279,482</sup>.

The candidate gene approach, has proven most successful with the identification and validation of the robust associations between the variant rs738409 (Ile148Met) in *PNPLA3* and alcohol-related cirrhosis<sup>377,428</sup> and alcohol-related HCC. This variant, which was originally identified by a GWAS of NAFLD<sup>370</sup>, will be discussed in detail in later sections of this work.

Table 1-12 Genetic association studies of alcohol-related liver disease phenotypes

Source	Case/control comparison	Patients	Cases	Controls	Gene	Variant	Risk allele	P-value
Järveläinen et al., 2001 <sup>195</sup>	Cirrhosis vs. CON	Finnish men	48	178	<i>CD14</i>	rs2569190 (-159)	T	<0.005
Grove et al., 1997 <sup>156</sup>	Cirrhosis vs. CON	British	118	145	<i>TNF-<math>\alpha</math></i>	rs361525 (-238)	NA	NS
	Steatohepatitis vs. CON		58	145			A	<0.05
	Cirrhosis vs. CON		118	145			NA	NS
Grove et al., 2000 <sup>157</sup>	ALD vs. CON	British	287	227	<i>IL-10</i>	rs1800872 (-627)	A	<0.0001
						rs1800896 (-1117)	A	<0.05
Savolainen et al., 1996 <sup>380</sup>	Cirrhosis vs. NLI	Finnish men	72	54	<i>GSTM1</i>	Null allele	Null allele	<0.05
Oliver et al., 2005 <sup>316</sup>	ALD vs. CON	Spanish	165	185	<i>TGF-<math>\beta</math>1</i>	rs1800470 (+869)	NA	NS
						rs1800471 (+915)	NA	NS
						rs1800469 (-509)	NA	NS
Hernández-Nazará et al., 2008 <sup>175</sup>	Cirrhosis vs. CON	Mexican	86	133	<i>APOE</i>	rs7412 (Arg112Cys) rs429358 (Cys158Arg) APOE*2: Cys112, Cys158 APOE*3: Cys112 Arg158 APOE*4: Arg112, Arg158	APOE*2	<0.01
Rossit et al., 2002 <sup>372</sup>	Cirrhosis vs. CON	South-eastern Brazil	97	96	<i>XRCC1</i>	rs1799782 (Arg194Trp)	NA	NS
						rs25487 (Arg399Gln)	Gln	<0.01
Degoul et al., 2001 <sup>90</sup>	Cirrhosis vs. CON	French	13	79	<i>SOD2</i>	rs4880 (Ala16Val)	Ala	<0.001
Marcos et al., 2009 <sup>280</sup>	Cirrhosis vs. NLI	Spain	96	161	<i>NFKB1</i>	rs28720239 (94ins/delATTG)	NA	NS
					<i>NFKBIA</i>	rs696	NA	NS
					<i>PPARG</i>	rs1801282 (Pro12Ala)	NA	NS
Grove et al., 1998 <sup>158</sup>	ALD vs. Population controls	British	242	117	<i>HFE</i>	rs1799945 (His63Asp)	NA	NS
						rs1800562 (Cys282Tyr)	NA	NS
Jun et al., 2009 <sup>207</sup>	ALD	Korean	82	113	<i>PEMT</i>	rs7946 (Val175Met)	NA	NS
			82	95	<i>MTTP</i>	rs3816873 (Ile128Thr)	T	<0.05
Stickel et al., 2008 <sup>412</sup>	Cirrhosis vs. NLI	British; German	689	334	<i>MMP3</i>	rs35068180 (5A/6A)	NA	NS
Valenti et al., 2004 <sup>442</sup>	ALD vs. NLI	Italian	183	43	<i>CTLA-4</i>	rs231775 (Thr17Ala)	Ala	<0.05
Takamatsu et al., 2000 <sup>419</sup>	Cirrhosis vs. NLI	Japanese	77	30	<i>IL-1B</i>	rs1143634 (+3953)	NA	NS
						rs16944 (-511)	NA	<0.005
	Cirrhosis vs. CON	Japanese	77	218		rs1143634 (+3953)	NA	NS
						rs16944 (-511)	NA	<0.05
Takamatsu et al., 1998 <sup>418</sup>	ALD vs. CON	Japanese men	102	46	<i>IL1RN</i>	rs2234663 (86bp VNTR)	NA	p<0.001

Abbreviations: ALD – Alcohol-related liver disease; CON – Population controls; NLI- Alcohol misusing no liver injury controls; NA – Non-available; NS – Non-significant

Table 1-13 Meta-analyses of candidate genes studies in alcohol-related liver disease

Source	Phenotype	Ethnicity	Cases	Number		Effects Model	Gene	Variant	Risk allele	P-value
				Controls	Studies					
Wong et al., 2000 <sup>162</sup>	Alcohol-related liver disease (variety of control comparisons)	Caucasian	782	1372	9	Random effects	CYP2E1	rs2031920 (Rsa-I)	NA	NS
Marcos et al., 2011 <sup>282</sup>	Alcohol-related liver disease vs. no liver disease controls	Majority Caucasian; minority of Indian, Chinese and mixed African-European ancestry	1015	590	8		GSTM1	Null	Null	<0.001
			583	506	6	Fixed effects	GSTT1	Null	NA	NS
			879	749	6		GSTP1	rs1695 (Ile105Val)	NA	NS
Marcos et al., 2011	Alcohol-related liver disease vs. no liver disease controls	Caucasian	96	161	3	Random effects	IL10	rs1800872 (-592)	NA	NS
Jin et al., 2010 <sup>200</sup>	Alcohol-related cirrhosis patients with: HCC vs. non-HCC	European	224	380	4	Random effects	HFE	rs1799945 (His63Asp)	NA	NS
						Fixed effects		rs1800562 (Cys282Tyr)	Tyr	<0.0005
Zintzaras et al., 2006 <sup>482</sup>	Alcohol-related liver disease (variety of control comparisons)	East-Asians	200	155	3		ALDH2	rs671 (Glu504Lys)	NA	NS
			595	427	8		CYPE2E1	rs3813867 (Pst-I), rs2031920 (Rsa-I)	NA	NS
		Caucasians; East-Asians	684	595	10	Fixed effects	ADH1B	rs1229984 (Arg48His)	Arg	<0.05
Marcos et al., 2009 <sup>281</sup>	Alcohol-related liver disease vs. no liver disease controls Cirrhosis vs. population controls		490	464	6		ADH1C	rs698 (Ile350Val) rs1693482 (Arg272Gln)	NA	NS
			566	473	5			rs361525 (-238)	NA	NS
		Caucasian	449	400	4	Random effects	TNFA	rs1800629 (-308)	NA	NS
		474	812	5			rs361525 (-238)	A	<0.05	
		374	632	4			rs1800629 (-308)	NA	NS	

Abbreviations: NS – Non-significant; NA – Non-available; HCC – Hepatocellular carcinoma

## 1.5 - AIMS OF THESIS

The overriding aim of this thesis was to determine the contribution of genetics factors to alcohol-related cirrhosis risk. To address this aim, the focus of this research was to contribute to a hypothesis generating genome-wide association study, with the subsequent identification, validation and characterization of identified loci. In particular, this work focuses on the functional and genetic characterization of *PNPLA3* through a detailed study of variants in this gene in the UCL cohort and functional characterization of the PNPLA3 protein through structural studies.

The specific aims of this thesis are:

1. To perform the first genome-wide association study of alcohol-related cirrhosis, in collaboration with several research groups, focusing on the identification and validation of novel loci.
2. To identify novel alcohol-related cirrhosis risk loci through an extended genome-wide association analysis including data from additional cohorts.
3. To refine the role of rs738409 in *PNPLA3* on alcohol-related cirrhosis risk beyond genetic association per se.
4. To characterise the PNPLA3 protein and the functional effects of the non-synonymous genetic variant rs738409 (Ile148Met) through structural and functional analysis.

---

---

**CHAPTER 2 GENOME-WIDE  
ASSOCIATION STUDY OF ALCOHOL-  
RELATED CIRRHOSIS**

---

---



## 2.1 - OVERVIEW

This chapter details the first genome-wide association study of alcohol-related cirrhosis demonstrating genome-wide significant associations in three loci: *patatin-like domain containing 3 (PNPLA3)*, *transmembrane 6 superfamily member 2 (TM6SF2)* and *membrane-bound O-acyltransferase domain-containing protein 7 (MBOAT7)*. These data are summarized in a recent publication<sup>44</sup>.

## 2.2 - BACKGROUND

A GWAS requires the genotyping of hundreds of thousands of genetic variants, typically SNPs, in cohorts that often contain many thousands of samples. Consequently, there are many potential sources of error. The commonest types of error are sample mishandling errors, population stratification and the presence of low quality genotypes and DNA samples in a dataset<sup>11</sup>. A quality control process is typically effected before association analysis during which low quality genotypes and samples are removed from the data.

Novel risk loci are identified in a GWAS by statistical tests where the frequencies of alleles or genotypes are compared by a phenotype or trait of interest. Because of the stringent significance levels required in a GWAS, statistical power is a key issue in the study and analytical design<sup>120</sup>. Many variables influence the statistical power of GWAS such as the genetic architecture of a phenotype, the inheritance model that is tested (dominant, recessive, allelic, genotypic) and the number of variants genotyped and the presence of confounding factors in the study population. The techniques of imputation and meta-analysis can maximise statistical power in a GWAS<sup>389</sup>.

Genotype imputation is the process through which genotypes are inferred through known linkage disequilibrium patterns in the human genome<sup>184</sup>. The genetic variants present on most GWAS bead-chips are, by-design, in linkage disequilibrium with the majority of other un-genotyped genetic variants that are present in the human genome. This 'coverage' of the human genome increases with a greater number of variants that are genotyped by a GWAS bead-chip and is further increased by using imputation. The increased variant coverage provided by imputation increases the statistical power of a GWAS, where simulations have demonstrated that imputation can boost power by around 10%<sup>406</sup>. Another prime use for imputation is the comparison of distinct GWAS datasets, which often have discordant variant genotypes. Imputation can allow datasets to be consolidated either through meta-analysis or the harmonisation and merging of raw genome-wide genotyping data.

Imputed genotypes are probabilistic and consequently contain a degree of uncertainty, which if not accounted for may lead to error. When comparing the concordance between imputed and direct genotypes error rates vary<sup>331</sup>. The uncertainty of imputed genotypes can be accounted for by using allele dosages rather than the binary allele coding used in association analysis, which is applied by the software SNPTEST<sup>278</sup>. The quality of imputation is largely dependent on a high-quality reference dataset such as the Haplotype Map project data<sup>192</sup> or the 1000 genomes project data<sup>1</sup>.

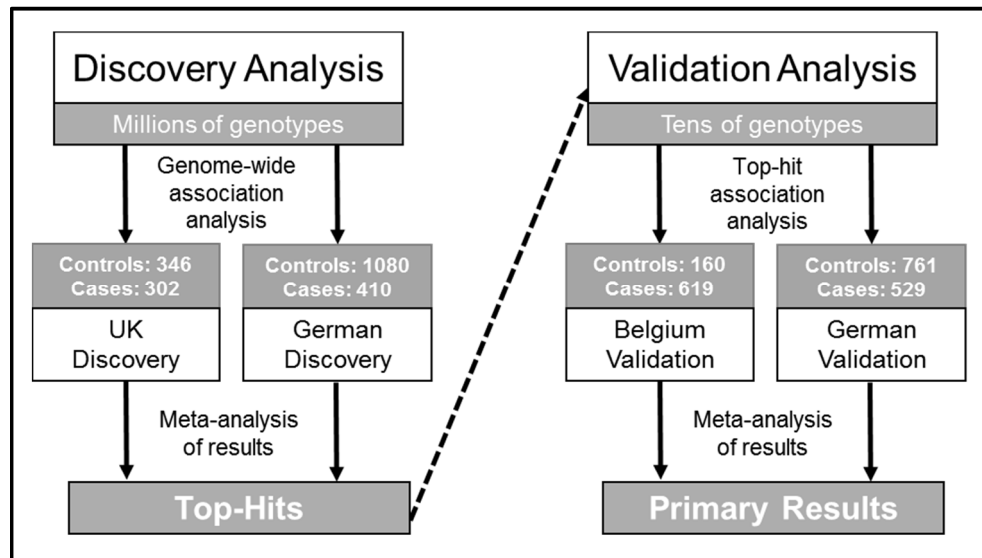
Meta-analysis is a statistical technique in which the results from multiple studies are analysed together<sup>119</sup> and is often applied in GWAS of the same phenotype or trait. It has particular utility when there are different GWAS datasets that cannot be harmonised and merged due to substantial population differences. Both fixed effects and random effects meta-analyses may be applied. Sometimes statistical heterogeneity, or the differences between study results that exceeds expected variance, occurs between studies for a number of variant associations. Heterogeneity may arise in a GWAS due to differences in ancestry or phenotypic classification between cohorts or differences in the genotyping subsequent processing of data. Despite the possibility of heterogeneity, the increased number of samples undergoing analysis increases statistical power.

## **2.3 - AIM**

To perform the first genome-wide association study of alcohol-related cirrhosis.

## **2.4 - MATERIALS AND METHODS**

This study was by design, split into two stages: (i) a discovery stage in which two primary cohorts underwent genome-wide association analysis identifying candidate loci; and, (ii) a validation stage in which the top-hit variants from the discovery stage were directly genotyped in two independent validation cohorts (Figure 2-1). Following the primary discovery and validation analyses, post-hoc analyses were performed.



**Figure 2-1 A schematic of the two-stage genome-wide association study design**

### 2.4.1 - DISCOVERY COHORT

The discovery cohort was comprised of two groups containing patients recruited in the UK and Germany (Table 2-1). These groups contained the DNA samples that underwent genome-wide association analysis for the identification of significantly associated variants for replication in the validation cohort.

Table 2-1 Demographics features of the discovery cohort

	Discovery Cohort			
	Germany		United Kingdom	
	Cases (n=410)	Controls (n=1080)	Cases (n=302)	Controls (n=346)
Median age years (IQR)	53 (47-61)	42 (36-48)	53 (47-60)	49 (42-56)
Gender (% Male)	71	100	68	77
Median BMI (IQR)*	26.2 (22.8-29.3)	24.8 (22.7-27.5)	24.8 (22.8-26.8)	24.6 (22.8-26.6)
Diabetes Type II (%)*	24	4.0	0**	0**

Abbreviations: *n* – number; IQR – inter-quartile range; BMI – Body mass index. \*\*Diagnosis of type II diabetes was an exclusion criteria in the UK cohort

### GERMAN DISCOVERY SAMPLES

The patients from Germany were recruited from 2000 until 2013 according to uniform criteria across all centres. Alcohol consumption data were self-reported via interview by a trained professional. All patients underwent careful clinical examination, standard laboratory testing and abdominal ultrasound. Patients were excluded from the study if they had any other cause of chronic liver disease, such as chronic viral hepatitis,

autoimmune liver disease or genetic haemochromatosis. Genomic DNA was extracted from peripheral blood samples according to standard procedures and quantified using the PicoGreen dsDNA Quantitation Kit (Invitrogen Corporation, Carlsbad, California).

### Cases with alcohol-related cirrhosis

Alcohol-related cirrhosis status was determined by either liver biopsy or unequivocal clinical and laboratory evidence. The clinical parameters for determining the alcohol-related cirrhosis included abnormal liver function test results, the confirmation of cirrhosis-related complications including encephalopathy, ascites or oesophageal varices and sonographic and/or radiological evidence for the presence of cirrhosis. These patients were recruited from Gastroenterology and Hepatology Departments in University affiliated hospitals in Germany, Austria and Switzerland.

### Controls with no significant liver injury

The samples classified as no-significant liver injury controls were defined as alcohol-dependent patients without cirrhosis as determined by biopsy or the clinical parameters. These patients were recruited in Psychiatric facilities in Germany and Switzerland.

## **UK DISCOVERY SAMPLES**

Subjects with self-reported English, Scottish, Welsh or Irish descent were recruited from the Centre for Hepatology at the Royal Free Hospital, London. All patients had a history of prolonged alcohol misuse and fulfilled DSM-IV criteria for a diagnosis of alcohol dependence. All patients were examined by two experienced, senior clinicians for signs liver injury. Blood was screened (antibodies to hepatitis A, B, C, D and E, cytomegalovirus, Epstein-Barr virus, herpes simplex and varicella; mitochondrial, nuclear, smooth muscle and liver kidney autoantibodies; iron, total iron binding capacity, ferritin; copper, caeruloplasmin; alpha one antitrypsin and tissue transglutaminase) and patients were excluded if they had any other potential causes of liver injury such as chronic viral hepatitis, autoimmune liver disease, genetic hemochromatosis, Wilson's disease,  $\alpha$ 1 antitrypsin deficiency or celiac disease. Patients were also excluded if they had a body mass index (BMI) >30 and were diabetic. All patients underwent abdominal ultrasound and/or abdominal X-ray computed tomography/magnetic resonance imaging; all underwent routine upper gastro-intestinal endoscopy; histological examination was undertaken whenever possible.

### Alcohol-related cirrhosis

The UK alcohol-related cirrhosis cases comprised of 302 patients. In 224 (74%), the diagnosis of cirrhosis was made on the basis of a sustained history of prolonged hazardous drinking; the presence of alcohol-dependence and histological examination of liver tissue. In the remaining 78 (26%) cirrhosis status was determined on the basis of a sustained history of prolonged hazardous drinking; the presence of alcohol-dependence and compatible historical, clinical, laboratory, radiological and endoscopic features.

### No significant liver injury

The UK no-significant liver injury controls comprised of 346 patients. In 122 (35%), the absence of significant liver injury was confirmed on biopsy. The remainder had no historical or clinical features suggestive of significant liver injury at presentation or during prolonged follow-up. Notably, 26 (11%) had isolated hyperbilirubinaemia, most likely reflecting the presence of Gilbert's syndrome. The remainder had normal serum bilirubin levels and all had normal plasma albumin concentrations and clotting profiles. The serum GGT activity was raised in the majority. The serum ALT activity was within the laboratory reference range in the majority and when elevated rarely exceeded twice the upper laboratory reference level; none had evidence of parenchymal liver injury or portal hypertension on imaging; upper gastrointestinal endoscopy was normal in those in whom it was performed.

## **2.4.2 - VALIDATION COHORT**

The validation cohorts contained those samples that underwent replication genotyping of the top-hits identified in the discovery phase of the GWAS. Genome-wide genotyping was not performed in any of these samples. This was comprised of two cohorts containing patients recruited from Belgium and Germany (Table 2-2).

Table 2-2 Demographics of the validation cohort

	Validation Cohort			
	Germany		Belgium	
	Cases (n=529)	Controls (n=761)	Cases (n=619)	Controls (n=161)
Median age years (IQR)	54 (47-62)	46 (39-53)	55 (49-61)	47 (41-55)
Gender (% male)	72	83	70	69
Median BMI (IQR)*	26.0 (23.0-29.2)	24.3 (21.7-27.0)	25.8 (22.7-29.8)	22.8 (20.6-25.7)
Diabetes Type II (%)*	18.1	11.3	18.2	2

Abbreviations: *n* – number; IQR – inter-quartile range; BMI – Body mass index

\*\*Diagnosis of type II diabetes was an exclusion criteria in the UK cohort

## BELGIAN VALIDATION COHORT

The 780 Belgium samples comprised patients with a long-term history of alcohol misuse recruited between 2002 and 2014 in the cities of Brussels and Haine-Saint-Paul. The number of alcohol-related cases and the number of males was disproportionately high in this cohort.

### Alcohol-related Cirrhosis

The Belgian alcohol-related cirrhosis cases comprised 619 patients with alcohol-related cirrhosis determined by either liver biopsy or unequivocal clinical and laboratory evidence for the presence of cirrhosis. The clinical parameters for determining alcohol-related cirrhosis included abnormal liver function test results, the confirmation of cirrhosis-related complications including encephalopathy, ascites or oesophageal varices and sonographic and/or radiological evidence for the presence of cirrhosis.

### No-significant Liver Injury

The Belgian no-significant liver injury controls were heavy drinkers with excessive alcohol intake ( $\geq 60$  g/day) and without clinically evident cirrhosis. They all received a diagnosis of alcohol-dependence based on DSM-IV criteria. These controls were screened with transient elastography and/or liver biopsy to exclude the presence of alcohol-related cirrhosis.

## GERMAN VALIDATION COHORT

The German validation alcohol-related cirrhosis cases and alcohol-misusers with no-significant liver injury were recruited by identical criteria as for the German discovery cohort. This cohort was predominantly male (Table 2-2).

## **ETHICAL APPROVAL**

In the discovery and validation cohorts none of the cases or controls had been used in any previous GWAS studies of alcohol-related cirrhosis. The protocols used for recruitment in the UK, German and Belgian cohorts in each cohort were approved by institutional review boards and all included subjects consented to inclusion into the study.

### **2.4.3 - DISCOVERY ANALYSIS**

#### **DISCOVERY GWAS ARRAY BASED GENOTYPING**

DNA samples underwent overnight whole genome amplification, followed by fragmentation by nuclease digestion and then purification via isopropanol precipitation. DNA was hybridized on to an Illumina BeadChip, which contains many thousands of covalently joined locus specific DNA sequences. Following hybridisation, the bead-chip underwent single base extension, using fluorescent labelled nucleotides allowing genotype detection<sup>190</sup>.

#### **DISCOVERY GWAS DATA PROCESSING**

The German and UK discovery cohorts underwent genome-wide genotyping using the same quality control and data processing protocols. All samples included in the discovery cohorts were genotyped on Illumina Bead Chips (Illumina, San Diego, USA) (Table 2-3). Genotype probabilities were determined from the raw fluorescence intensity data using the software BEAGLECALL<sup>43</sup>. The genotype data were converted into the PLINK<sup>349</sup> compatible binary PED file format using fcGENE<sup>371</sup>. As the German case and control populations were genotyped on three different platforms this has a potential to result in systematic differences between datasets and hence intersection strategy was applied to these genotype whereby only the genotypes that were present on all of the genotyping platforms were utilised for subsequent analysis<sup>202</sup>. For this merging and harmonisation process, all variants in these datasets were aligned to the positive strand of the human reference genome build 19 (Genome Reference Consortium GRCh37).

Table 2-3 The genome-wide genotyping bead-chip arrays used on the discovery cohort

Group	Number	Array
German Cases	410	Illumina OmniExpress array (Version 12v1_1)
German Controls	329	Illumina Human610Quad
	407	Illumina HumanHap550 BeadChip
	383	Illumina Human660w Quad BeadChip arrays
UK Cases and Controls	648	Illumina OmniExpress (version 24v1-0_a)

Following the initial stages of quality control samples underwent further quality control<sup>11</sup> performed using PLINK<sup>349</sup> (version 1.07) and the R<sup>189</sup> programming language. Samples were excluded from further analysis based on: (i) genotyping success of less than 97%; (ii) outlying autosomal heterozygosity of more or less 3 standard deviations from the mean 3 standard deviations from mean; (iii) a kinship coefficient greater than 18.5% between two samples; and, (iv) discordance between self-reported gender and sex-chromosome inferred gender.

Samples were also excluded if cryptic non-European ancestry was identified from the genome-wide genotyping data. To perform this, the UK and German datasets were merged with the HapMap phase III dataset<sup>192</sup>, which contains the ancestrally informative genotypes for several populations that are representative of the major human ancestral groups (European, East Asian, South Asian, African). The merged HapMap and UK/German data underwent multidimensional scaling analysis in PLINK<sup>349</sup>. These sources of variation were plotted in R<sup>352</sup>, in which samples that deviated from the average similarity (median  $\pm$  3 standard deviations) of the HapMap reference European ancestral group were excluded.

### Discovery Cohort Imputation

The genome-wide genotyping data for both the UK and German discovery cohorts underwent imputation. Both the UK and German discovery cohort genotypes were converted into the Oxford GENS format using fcGENE<sup>371</sup>. Imputation was performed using IMPUTE2<sup>184</sup> (version 2.3.1). The 1000 genomes phase 3 autosomal dataset and the 1000 genomes phase 1 maternal X chromosome dataset were used as the source of reference haplotypes for imputation<sup>1</sup>. Following the imputation any genotypes with high imputation uncertainty (IMPUTE2 info-score < 0.8), or low minor allele frequency



(< 1%) or by deviation of the allele frequencies from the Hardy-Weinberg equilibrium ( $P < 1 \times 10^{-6}$ ) were removed from the data and not included in association analysis.

## **DISCOVERY GWAS ASSOCIATION ANALYSIS**

The quality controlled and imputed UK and German genome-wide genotyping data underwent association testing using the software SNPTEST (v2.5) using an additive frequentist model<sup>277</sup>. In both of these dataset, the genomic inflation factor was estimated from the  $P$ -values. If population stratification was evident ( $\lambda_{GC} > 1.1$ ) the analysis was adjusted by calculating the top three principal components of genetic variation using EIGENSTRAT<sup>345</sup>. These top two principal components were included as covariates. After this, the genomic inflation factor was re-estimated to verify that  $P$ -value inflation was resolved. The association results were visualized using Manhattan plots were created using the qqman package<sup>436</sup> and Quantile-Quantile plots (QQ plots) were created in the Haplin package in R<sup>352</sup>. Regional plots of the association results in the context of genomic location, inter-variant linkage disequilibrium and local-recombination rates were created using the Locuszoom server<sup>347</sup>.

Association results from the independent UK and German discovery dataset underwent fixed-effect meta-analysis using the software META (v1.5.0). The analysis was restricted to markers present on both datasets. The combined results were analysed for genomic inflation and any deviation from the null was adjusted using genomic control for each dataset in META<sup>97</sup>. The heterogeneity between studies was quantified using the  $I^2$  statistic<sup>176</sup>. The results data were processed to identify the 10 unique 'top-hits' that were defined as the strongest association signals present in independent loci (separated by at least a 500 kB genomic distance).

### **2.4.4 - VALIDATION ANALYSIS**

#### **GENOTYPING**

The top-hits from the discovery analysis were directly genotyped in the German and Belgian validation cohorts. In the German and Belgian validation and population control cohorts genotyping was performed using Taqman genotyping (Applied Biosystems, Foster City, Ca, USA). The TaqMan assay fluorescence data was read using a 7900 HT TaqMan sequence detector system (Applied Biosystems).

#### **ASSOCIATION ANALYSIS**

The top-hit variants identified in the discovery stage underwent single marker genotyping in the German and Belgian validation cohorts. The raw genotype data were converted into binary PED format and underwent logistic regression in PLINK<sup>349</sup>. The genetic association results for the 10 top-hits from the discovery stage of analysis

underwent fixed-effect meta-analysis with the genetic association results from the validation stage using the software META (version 1.5.0).

### **2.4.5 - POST-HOC ANALYSIS**

Three separate analyses were performed post-hoc. The first was a genetic association analysis of genome-wide significant associated variants adjusted for clinical and demographic variables. The second was a selective pathways analysis to determine whether pathways annotated to the genes containing genome-wide significant associations with alcohol-related cirrhosis, contain a greater number of genetic associations than would be expected by chance. The third was an analysis of genetic associations in the GWAS discovery dataset of variants in candidate genes that have already been tested for association of alcohol-related cirrhosis in independent cohorts.

#### **ADJUSTED ANALYSIS**

Age, sex, BMI and type II diabetes status data were available in the majority of patients in the discovery validation cohorts (Table 2-4). In the German discovery cohort, 71% of cases and 70% of controls had BMI data and 97% of cases and 74% had known type II diabetes status. In the UK discovery cohort 80% of cases and 53% of controls had BMI data and none had type II diabetes (this was a cause for exclusion). In the German validation cohort, 50% of cases and 40% of controls had BMI information and 76% of cases and 64% had known type II diabetes status. In the Belgian validation cohort, 85% of cases and 91% of controls had available BMI information and 93% of cases and 99% of controls had known type II diabetes status. In these sample, logistic regression association was performed conditioning on these variables as covariates in SNPTEST and PLINK. Otherwise, associate analyses were the same as those performed in the primary analyses albeit on a smaller sample size.

Table 2-4 The demographics of the samples in the adjusted meta-analysis

Variable	Discovery Cohorts			
	Germany		United Kingdom	
	Cases (n=213)	Controls (n=691)	Cases (n=243)	Controls (n=182)
Median age years (IQR)	55 (49-61)	41 (35-48)	56 (48-61)	50 (43-48)
Gender (% Male)	100	100	66	76
Median BMI (IQR)	26.9 (23.6-29.4)	24.9 (22.8-27.5)	24.8 (22.8-26.8)	24.6 (22.8-26.6)
Diabetes Type II (%)*	26.3	4.9	0	0
Variable	Validation Cohorts			
	Germany		Belgium	
	Cases (n=272)	Controls (n=417)	Cases (n=533)	Controls (n=144)
Median age years (IQR)	54 (48-61)	47 (41-53)	55 (49-61)	47 (41-56)
Gender (% male)	77	81	70	58
Median BMI (IQR)*	26.1 (23.0-29.3)	24.3 (21.7-27.0)	25.8 (22.7-29.8)	22.8 (20.6-25.7)
Diabetes Type II (%)*	22.1	14.7	18.9	2

Abbreviations: *n* – number; IQR – inter-quartile range; BMI – Body mass index

\* A diagnosis with type II diabetes was an exclusion criteria in the UK discovery cohort

## PATHWAYS ANALYSIS

Top-hits identified through the primary analysis underwent a biological literature review to determine the gene and the biological context of the genes function. A top biological pathway for each gene was annotated. All other genes present in these annotated pathways were selected from the REACTOME biological pathways database<sup>76</sup>. These genes were used to select out all of the variants present in the UK GWAS dataset that occur within 20 kB of the genes boundaries that occur in these biological pathways. Using these data, a set of SNPs was selected for each biological pathway and these SNPs underwent adaptive permutation based set test analysis on the primary logistic regression dataset (default settings) (PLINK version 1.9)<sup>349</sup> to determine an empirical *P* value to test whether each pathways has a greater number of genetic associations than would be expected through chance.

## CANDIDATE GENE ANALYSIS

Variants that have been genotyped in previous candidate gene studies (Table 1-13; Table 1-12) were extracted from the GWAS Discovery meta-analysis results data and tabulated.

## 2.5 - RESULTS

### 2.5.1 - DISCOVERY ANALYSIS

#### UK DISCOVERY SAMPLE

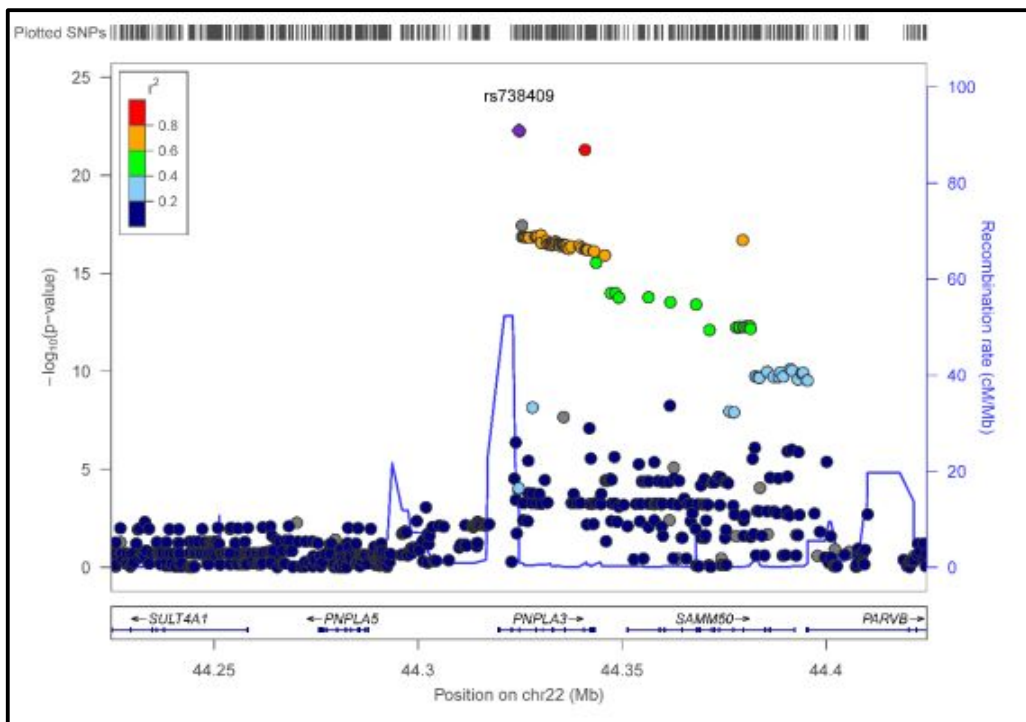
In the UK discovery cohort, raw genotype data was available for 716,503 SNPs in 672 individuals. The number of which was reduced to 609,879 SNPs in 648 individuals after quality control. This dataset of directly genotyped variants underwent imputation generating a final dataset with 7,871,013 imputed and directly genotyped variants in 648 individuals. There was no evidence of population stratification ( $\lambda_{GC} = 1.01$ ). The variants rs738409 and rs738408 associate with alcohol-related cirrhosis below genome-wide significance ( $P < 5 \times 10^{-8}$ ). Several variants associate at suggestive levels of significance ( $P < 5 \times 10^{-5}$ ) including rs2294915, rs12585483, rs4823173, rs6885856, rs1324922, rs11149248, rs1021831, rs12483959, rs13099969, rs35096230, rs12495649, rs12485518, rs12638570 and rs12484700. Of these, nine are located in the *PNPLA3* locus and the remainder are in five independent genomic regions on chromosomes 3, 5, 12 and 13 (Table 2-5).

Table 2-5 Top-association statistics in the UK sample

Chromosome	Gene/Locus	Variant	Position	P-value	Odds Ratio	95% Confidence Interval
22	<i>PNPLA3</i>	rs738409	44324727	$2.51 \times 10^{-8}$	2.12	[1.63-2.77]
22	<i>PNPLA3</i>	rs738408	44324730	$2.51 \times 10^{-8}$	2.12	[1.63-2.77]
22	<i>PNPLA3</i>	rs2294915	44340904	$2.19 \times 10^{-7}$	1.98	[1.53-2.56]
13	Intergenic	rs12585483	83298475	$7.05 \times 10^{-6}$	0.59	[0.47-0.74]
22	<i>PNPLA3</i>	rs4823173	44328730	$7.33 \times 10^{-6}$	1.93	[1.45-2.57]
5	Intergenic	rs6885856	34587844	$7.42 \times 10^{-6}$	1.70	[1.35-2.14]
13	Intergenic	rs1324922	83306586	$7.57 \times 10^{-6}$	0.59	[0.47-0.74]
13	Intergenic	rs11149248	83304898	$7.90 \times 10^{-6}$	0.59	[0.47-0.74]
12	Intergenic	rs1021831	128141516	$7.96 \times 10^{-6}$	0.53	[0.4-0.7]
22	<i>PNPLA3</i>	rs12483959	44325996	$8.42 \times 10^{-6}$	1.93	[1.45-2.58]
3	Intergenic	rs13099969	79881007	$8.70 \times 10^{-6}$	1.74	[1.36-2.22]
3	Intergenic	rs35096230	79881660	$8.70 \times 10^{-6}$	1.74	[1.36-2.22]
3	Intergenic	rs12495649	79890801	$8.70 \times 10^{-6}$	1.74	[1.36-2.22]
3	Intergenic	rs12485518	79893872	$8.70 \times 10^{-6}$	1.74	[1.36-2.22]
3	Intergenic	rs12638570	79879521	$9.78 \times 10^{-6}$	1.73	[1.36-2.22]
22	<i>PNPLA3</i>	rs12484700	44327273	$9.81 \times 10^{-6}$	1.92	[1.44-2.56]

## GERMAN DISCOVERY SAMPLE

In the German discovery cohort, genotype data was available for 298,405 variants in 1,490 individuals and after imputation this increased to 6,866,424 variants. The genomic inflation factor ( $\lambda_{GC} = 1.15$ ) suggests the presence of moderate population stratification. To counter this, all logistic regression based association analyses included the top principal component as a covariate. A total of 126 variants associate with the alcohol-related cirrhosis at genome-wide significance ( $P < 5 \times 10^{-8}$ ); nearly all of these associations occur within a 500 kB window surrounding the *PNPLA3*. An additional 130 variants associated at suggestive levels ( $P < 5 \times 10^{-5}$ ) in five independent genomic regions. When association results for the *PNPLA3* locus are plotted in the context of local recombination rates and nearby genes (Figure 2-2) the association signal peaks near a region of high recombination in the third exon of *PNPLA3* and extends over a region of high linkage disequilibrium encompassing both the *PNPLA3* and the neighbouring *SAMM50*.



**Figure 2-2 A locus plot of genetic associations in *PNPLA3* in the German discovery group**

This locus plot of the shows the associations at *PNPLA3* shows  $-\log_{10}$  of the P values for each variant plotted against their genomic position (human genome build 19). SNPs are coloured to reflect correlation with the most significant SNP, with red denoting the highest linkage disequilibrium, ( $r^2 > 0.8$ ) with the lead SNP. Estimated recombination rates from the 1000 Genomes Project<sup>1</sup> European populations are plotted in blue to reflect the local linkage disequilibrium structure. Gene annotations were obtained from the UCSC Genome Browser and the plot was generated using LocusZoom<sup>347</sup>

## DISCOVERY META-ANALYSIS

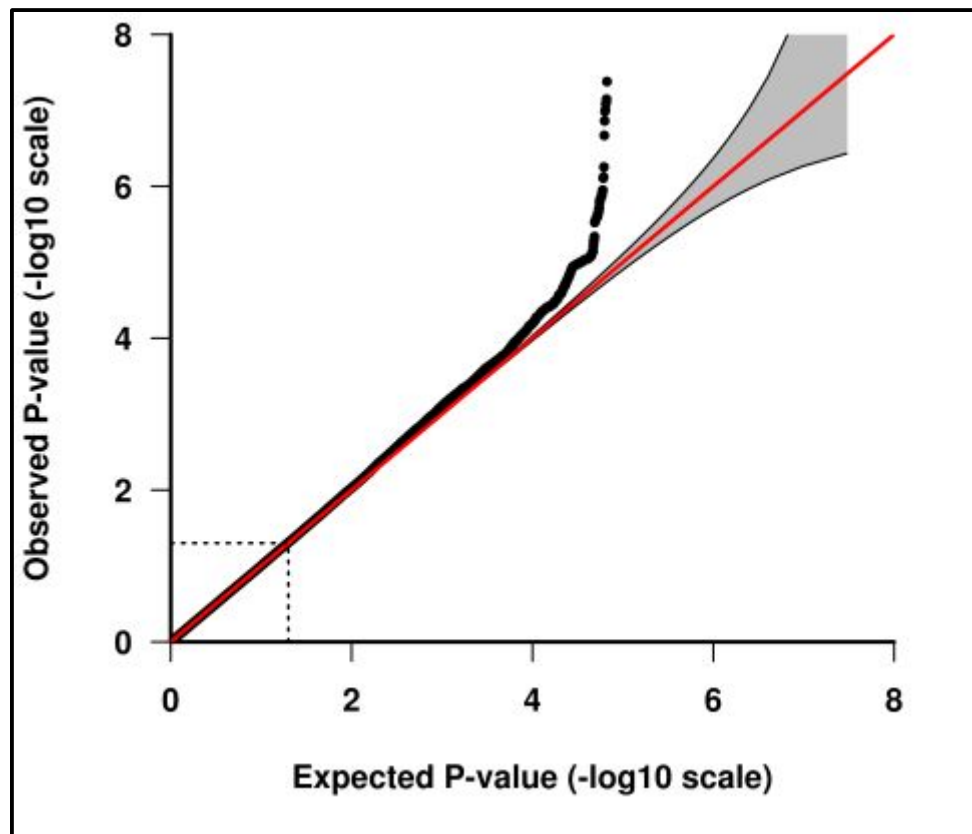
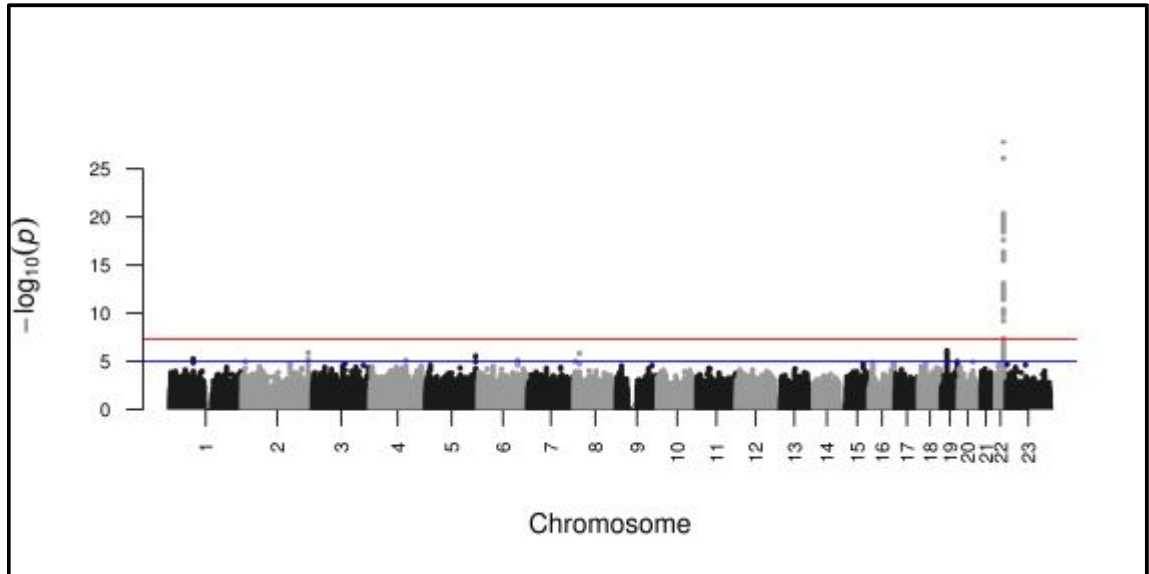
There was no evidence of population stratification in the combined meta-analysis ( $\lambda_{GC} = 1.01$ ) (Figure 2-3). A number of variants in *PNPLA3* were associated at the level of genome-wide significance and of these rs738409 was the most significant. In addition several independent regions of interest were identified which were associated at suggestive levels of significance (Table 2-6). One of these variants, rs739846 in *SUGP1*, occurs in a genomic region of strong linkage disequilibrium. This region of strong linkage disequilibrium contains the functional<sup>269</sup> variant rs58542926 (Glu167Lys) in *TM6SF2* and has previously been associated with NAFLD<sup>229,261</sup>. As this variant appears to be the most plausible causal variant at this locus, it was also genotyped in the validation cohort and underwent meta-analysis.

Table 2-6 Fixed effects meta-analysis results from the UK and German discovery cohort

Chr	SNP	Locus	Allele	Allele Frequency	$P_{META}$	$I^2$	Odds Ratio [95% CI]
22	rs738409	<i>PNPLA3</i>	G	0.27	$1.17 \times 10^{-28}$	60	2.39 [2.05-2.78]
19	rs739846	<i>SUGP1/TM6SF2</i>	A	0.08	$7.54 \times 10^{-7}$	0	1.92 [1.48-2.50]
19	rs58542926*	<i>SUGP1/TM6SF2</i>	T	0.08	$2.86 \times 10^{-6}$	0	1.87 [1.44-2.43]
2	rs62190923	<i>TM4SF20</i>	G	0.21	$1.31 \times 10^{-6}$	0	0.64 [0.54-0.77]
8	rs7812374	Intergenic	T	0.58	$1.46 \times 10^{-6}$	0	0.70 [0.60-0.81]
5	rs6556045	Intergenic	A	0.06	$2.51 \times 10^{-6}$	0	2.11 [1.55-2.87]
1	rs6605237	Intergenic	T	0.27	$5.43 \times 10^{-6}$	0	1.46 [1.24-1.71]
4	rs17886348	<i>IL21</i>	T	0.08	$7.64 \times 10^{-6}$	60	1.79 [1.39-2.31]
6	rs7769670	<i>PDE7B</i>	A	0.14	$7.84 \times 10^{-6}$	0	1.58 [1.29-1.93]
8	rs7845021	Intergenic	C	0.61	$1.02 \times 10^{-5}$	0	0.73 [0.64-0.84]
19	rs626283	<i>TMC4/MBOAT7</i>	C	0.44	$1.07 \times 10^{-5}$	0	1.36 [1.19-1.57]

\* rs58542926 included in analysis as this variant is thought to be functionally implicated in liver disease pathogenesis

Abbreviations: CHR – chromosome; SNP – single nucleotide polymorphism; CI – confidence interval;  $I^2$  – heterogeneity index



**Figure 2-3 A Manhattan and Quantile-Quantile plot of the combined UK-German GWAS meta-analysis**

Top image: Manhattan plot of German GWAS results. The blue horizontal line indicates suggestive genetic association ( $P < 5 \times 10^{-5}$ ) and the red horizontal line indicates genome-wide significant genetic association ( $P < 5 \times 10^{-8}$ ). Bottom image: A QQ plot of the observed versus the expected  $P$ -values for the German GWAS results where the red line indicates the expected distribution of results under the null hypothesis. The grey outline shows the 95% confidence interval of the expected distribution of  $P$ -values. The dotted line represents a marginal significance threshold ( $P \leq 0.05$ )

## 2.5.2 - VALIDATION ANALYSIS

All selected variants were genotyped successfully and genotype distributions were in Hardy-Weinberg equilibrium in both cases and controls ( $P > 0.05$ ). There was no heterogeneity between any of the variants genotyped in both the discovery and validation cohorts ( $I^2 < 0.25$ ). The variants in three of the top-hit loci replicated in the validation cohort, these were rs738409 in *PNPLA3*, rs10401969 and rs58542926 in the *TM6SF2/SUGP1* locus and rs626283 in the *MBOAT7/TMC4* locus (Table 2-7).

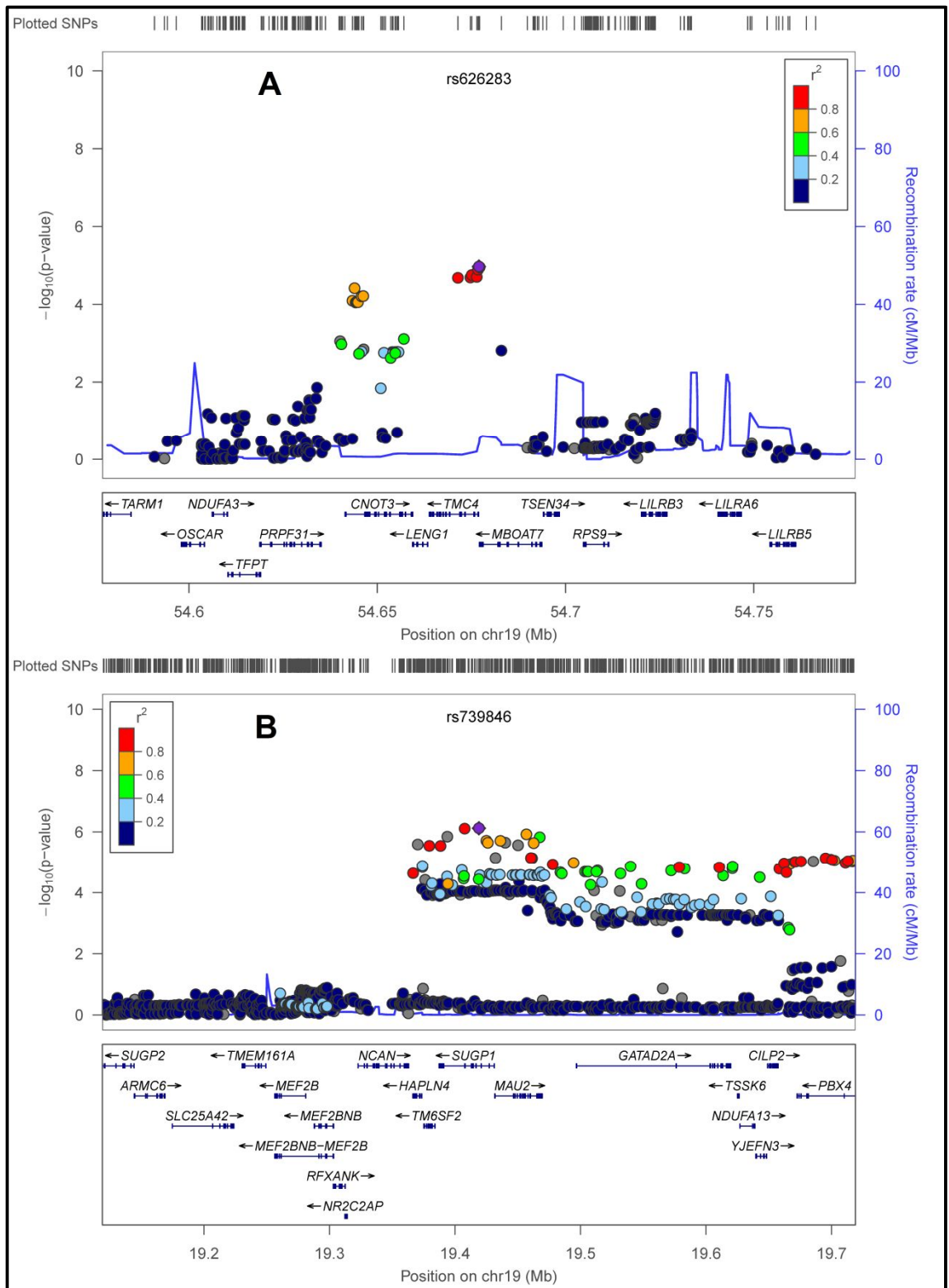
Table 2-7 Validation of the top SNPs identified in the validation cohort

Chr	SNP	Validation Cohort		Discovery and Validation Cohorts	
		$P_{\text{VALIDATION}}$	Odds Ratio [95% CI]	$P_{\text{META}}$	Odds Ratio [95% CI]
22	rs738409	$4.59 \times 10^{-22}$	2.03 [1.76-2.35]	$1.54 \times 10^{-48}$	2.19 [1.97-2.43]
19	rs10401969*	$1.24 \times 10^{-4}$	1.57 [1.25-1.97]	$7.89 \times 10^{-10}$	1.71 [1.44-2.04]
19	rs58542926	$7.34 \times 10^{-5}$	1.59 [1.26-1.99]	$1.33 \times 10^{-9}$	1.70 [1.43-2.02]
2	rs62190923	0.73	1.03 [0.87-1.22]	-	-
8	rs7812374	0.69	0.97 [0.85-1.11]	-	-
5	rs6556045	0.16	1.20 [0.93-1.56]	-	-
1	rs6605237	0.80	0.98 [0.84-1.14]	-	-
4	rs17886348	0.99	1.00 [0.77-1.30]	-	-
6	rs7769670	0.37	1.09 [0.90-1.33]	-	-
8	rs7845021	0.19	1.09 [0.95-1.26]	-	-
19	rs626283	$2.29 \times 10^{-5}$	1.33 [1.17-1.53]	$1.03 \times 10^{-9}$	1.35 [1.23-1.49]

\*The lead SNP rs739846 failed genotyping for technical reasons and therefore rs10401969 was used as a replacement ( $r^2 = 1.0$ ). Abbreviations: CHR – chromosome; SNP – single nucleotide polymorphism; CI – confidence interval. Table published in Buch et al., 2015<sup>44</sup>

The variant rs626283 sits in an extended region of linkage disequilibrium which covers several genes including *MBOAT7*, *TMC4*, *CNOT3* and *LENG1* (Figure 2-4: Image A). The most significantly associated SNP, rs626283, occurs in an intergenic region between the 3' promoter of *TMC4* and the 5' untranslated region of the *MBOAT7*. The *SUGP1/TM6SF2* variants rs10401969, which was genotyped as a proxy for rs739846, and rs58542926 sit in another extended 500 kB region of high linkage disequilibrium on chromosome 19 which includes several other genes including *NCAN*, *MAU2*, *GATAD2A* and *CILP2* (Figure 2-4: Image B). The functional variant rs58542926 in *TM6SF2* was less strongly associated with alcohol-related cirrhosis than the intronic variant rs10401969 in *SUGP1*<sup>145</sup>.





**Figure 2-4 Fine-mapping analysis of the genome-wide significant loci**

For the variants which replicate in the validation analysis locus plots are given for: A – *TMC4/MBOAT7*; and, B – *TM6SF2/SUGP1*. The  $-\log_{10}(P\text{ values})$  are plotted against SNP genomic position based on the human genome build 1937. SNPs are coloured to reflect correlation with the most significant SNP, with red denoting the highest linkage disequilibrium, ( $r^2 > 0.8$ ) with the lead SNP. Estimated recombination rates from the 1000 Genomes Project (Europeans populations) are plotted in blue to reflect the local linkage disequilibrium structure. Gene annotations were obtained from the UCSC Genome Browser.

## 2.5.3 - POST-HOC ANALYSIS

### ADJUSTED ANALYSIS

In the discovery and validation cohorts the logistic regression based allelic association was adjusted for age, sex, BMI and type II diabetes status. This analysis was performed as per the main GWAS with a discovery and replication phase but with 809 fewer samples due to missing variable information. There was no evidence for population stratification. The only variant to contain genome-wide significant associations was rs7390409 in *PNPLA3*. All of the other top-hit variants remain associated, although none reached genome-wide significance.

Table 2-8 Post-hoc association analysis adjusted for gender, age, BMI and type II diabetes status

Locus	Chromosome	SNP ID	Unadjusted		Adjusted	
			P-value	Odds Ratio [95% CI]	P-value	Odds Ratio [95% CI]
<i>PNPLA3</i>	22	rs738409	$4.59 \times 10^{-22}$	2.03 [1.76–2.35]	$4.05 \times 10^{-13}$	2.12 [1.73–2.59]
<i>TM6SF2</i>	19	rs10401969	$1.24 \times 10^{-4}$	1.57 [1.25–1.97]	0.020	1.43 [1.06–1.94]
		rs58542926	$7.34 \times 10^{-5}$	1.59 [1.26–1.99]	0.022	1.43 [1.05–1.94]
<i>MBOAT7</i>	19	rs626283	$2.29 \times 10^{-5}$	1.33 [1.17–1.53]	$3.81 \times 10^{-4}$	1.41 [1.17–1.70]
		rs641738	$1.30 \times 10^{-5}$	1.35 [1.18–1.54]	$2.11 \times 10^{-4}$	1.43 [1.18–1.72]

Table from Buch et al., 2015<sup>44</sup>

Abbreviations: CI – confidence interval

### PATHWAYS ANALYSIS

Three pathways from the REACTOME database<sup>76</sup> were selected to be representative for *TM6SF2*, *PNPLA3* and *MBOAT7* (Table 2-9). The variants present in the genes which constitute each pathway were annotated to three gene sets in the UK dataset. Both the arachidonic acid metabolism pathway and the glycerophospholipid biosynthesis pathway contain a greater number of genetic associations than would be expected by chance ( $P < 0.05$ ) (Table 2-10).

### CANDIDATE GENE ANALYSIS

Three independent loci contain variants that associate below marginal significance ( $P_{META} \leq 0.05$ ): *ADH1C*, *CYP2E1* and *GSTP1*. In both *ADH1C* and *CYP2E1*, the two associated variants (rs698; rs1693482 and rs3813867; rs2031920 respectively) are known to be in strong linkage disequilibrium<sup>1</sup>. There is significant heterogeneity for the genetic association for these variants, between the UK and German results data in this fixed effects meta-analysis.

Table 2-9 The genes in each candidate pathway

Gene	Pathway	Other Genes in Pathway
<i>TM6SF2</i>	Lipid digestion, mobilization, and transport (REACT_602.8)	<i>A2M, ABCA1, ABCG1, ABCG5, ABCG8, ABHD5, ALB, AMN, APOA1, APOA2, APOA4, APOA5, APOB, APOC2, APOC3, APOE, APOF, BMP1, CAV1, CEL, CETP, CLPS, CUBN, FABP1, FABP12, FABP2, FABP3, FABP4, FABP5, FABP6, FABP7, FABP9, HSPG2, LCAT, LDLR, LDLRAP1, LIPC, LIPE, LPA, LPL, MGLL, MTP, NPC1L1, P4HB, PLIN1, PLTP, PLTP, PNLIP, PNLIPRP1, PNLIPRP2, PPP1CA, PPP1CB, PPP1CC, PRKACA, PRKACB, PRKACG, SAR1B, SCARB1, SDC1</i>
<i>MBOAT7</i>	Arachidonic acid metabolism (10.3180/REACT_147851.3)	<i>ABCC1, AKR1C3, ALOX12, ALOX12B, ALOX15, ALOX15B, ALOX5, ALOX5AP, CBR1, CYP1A1, CYP1A2, CYP1B1, CYP2C19, CYP2C8, CYP2C9, CYP2J2, CYP2U1, CYP4A11, CYP4A22, CYP4B1, CYP4F11, CYP4F2, CYP4F22, CYP4F3, CYP4F8, CYP8B1, DPEP1, DPEP2, DPEP3, EPHX2, FAM213B, GGT1, GGT5, GPX1, GPX2, GPX4, HPGD, HPGDS, LTA4H, LTC4S, MAPKAPK2, PLA2G4A, PTGDS, PTGES, PTGES3, PTGIS, PTGR1, PTGS1, PTGS2, TBXAS1</i>
<i>PNPLA3</i>	Glycerophospholipid biosynthesis (10.3180/REACT_121401.2)	<i>ACHE, AGPAT1, AGPAT2, AGPAT3, AGPAT4, AGPAT5, AGPAT6, AGPAT9, BCHE, CDIPT, CDS1, CDS2, CEPT1, CHAT, CHKA, CHKB, CHPT1, CPNE1, CPNE3, CPNE6, CPNE7, CRLS1, DGAT1, DGAT2, EPT1, ETNK1, ETNK2, GNPAT, GPAM, GPAT2, GPCPD1, GPD1, GPD1L, HADHA, HADHB, LCLAT1, LPCAT1, LPCAT2, LPCAT3, LPCAT4, LPGAT1, LPIN1, LPIN2, LPIN3, MBOAT1, MBOAT2, MBOAT7, MGLL, PCYT1A, PCYT1B, PCYT2, PEMT, PGS1, PHOSPHO1, PISD, PITPNB, PLA2G10, PLA2G12A, PLA2G16, PLA2G1B, PLA2G2A, PLA2G2D, PLA2G2E, PLA2G2F, PLA2G3, PLA2G4A, PLA2G4B, PLA2G4C, PLA2G4D, PLA2G4E, PLA2G4F, PLA2G5, PLA2G6, PLB1, PLBD1, PLD1, PLD2, PLD3, PLD4, PLD6, PNPLA2, PNPLA3, PNPLA8, PTDSS1, PTDSS2, PTPMT1, SLC44A1, SLC44A2, SLC44A3, SLC44A4, SLC44A5, TAZ</i>

Table 2-10 Post-hoc gene set-based association test results

Pathway	SNP Number	Number Significant ( $P < 0.05$ )	Empirical Significance ( $P$ )	Top independent Associations
Lipid digestion, mobilization, and transport	4884	222	0.85	-
Arachidonic acid metabolism	4450	291	0.0091	rs5751902, rs45535831, rs11078528, rs17205398, rs3801150, rs73734172
Glycerophospholipid biosynthesis	9958	469	0.0022	rs738408, rs16972418, rs11414481, rs6439081, rs1883350

Table 2-11 Post-hoc analysis of candidate gene variants

Chr	Gene	Variant	Position (hg19)	Risk allele	$P_{META}$	Odds ratio	95% CI	$P_{HET}$	$I^2$
1	<i>IL-10</i>	rs1800872 (-627)	206946407	T	0.30	0.91	[0.93-1.07]	0.20	37.9
1	<i>IL-10</i>	rs1800896 (-1117)	206946897	C	0.27	0.92	[0.93-1.04]	0.013	83.9
2	<i>IL-1B</i>	rs1143634 (+3953)	113590390	A	0.31	0.92	[0.93-1.04]	0.013	83.9
2	<i>IL-1B</i>	rs16944 (-511)	113594867	A	0.46	1.06	[1.05-0.97]	0.027	79.7
2	<i>CTLA-4</i>	rs231775 (Thr17Ala)	204732714	G	0.39	1.06	[1.05-0.95]	0.48	0.0
3	<i>PPARG</i>	rs1801282 (Pro12Ala)	12393125	G	0.45	0.93	[0.94-1.04]	0.082	66.9
4	<i>ADH1C</i>	rs698 (Ile350Val)	100260789	C	0.038	0.86	[0.88-1.07]	0.052	73.5
4	<i>ADH1C</i>	rs1693482 (Arg272Gln)	100263965	T	0.038	0.86	[0.88-1.07]	0.055	72.9
4	<i>MTTP</i>	rs3816873 (Ile128Thr)	100504664	C	0.61	1.04	[1.04-0.97]	0.33	0.0
5	<i>CD14</i>	rs2569190 (-159)	140012916	A	0.070	1.13	[1.12-0.94]	0.15	52.7
6	<i>HFE</i>	rs1799945 (His63Asp)	26091179	G	0.94	0.99	[0.99-1]	0.71	0.0
6	<i>HFE</i>	rs1800562 (Cys282Tyr)	26093141	A	0.87	1.03	[1.02-0.99]	0.19	42.0
6	<i>TNF-<math>\alpha</math></i>	rs1800629 (-308)	31543031	A	0.26	1.12	[1.09-0.92]	0.0079	85.8
6	<i>TNF-<math>\alpha</math></i>	rs361525 (-238)	31543101	A	0.38	0.87	[0.91-1.02]	0.053	73.4
6	<i>SOD2</i>	rs4880 (Ala16Val)	160113872	A	0.22	0.92	[0.93-1.07]	0.32	0.0
10	<i>CYP2E1</i>	rs3813867 (Pst-I)	135339605	C	0.011	0.58	[0.73-0.84]	0.052	73.5
10	<i>CYP2E1</i>	rs2031920 (Rsa-I)	135339845	T	0.014	0.60	[0.74-0.85]	0.046	75.0
11	<i>GSTP1</i>	rs1695 (Ile105Val)	67352689	G	0.024	1.19	[1.16-0.85]	0.75	0.0
14	<i>NFKBIA</i>	rs696	35871093	T	0.97	1.00	[1.0-1.0]	0.15	50.5
17	<i>PEMT</i>	rs7946 (Val175Met)	17409560	C	0.73	1.03	[1.02-0.98]	0.36	0.0
19	<i>TGF-<math>\beta</math>1</i>	rs1800471 (+915)	41858876	G	0.61	0.93	[0.95-1.02]	0.74	0.0
19	<i>TGF-<math>\beta</math>1</i>	rs1800470 (+869)	41858921	G	0.75	0.98	[0.98-1.02]	0.37	0.0
19	<i>TGF-<math>\beta</math>1</i>	rs1800469 (-509)	41860296	A	0.95	1.00	[1.0-1.0]	0.28	14.7
19	<i>XRCC1</i>	rs25487 (Arg399Gln)	44055726	T	0.82	1.02	[1.01-0.99]	0.84	0.0
19	<i>XRCC1</i>	rs1799782 (Arg194Trp)	44057574	A	0.25	0.85	[0.89-1.03]	0.29	9.8
19	<i>APOE</i>	rs429358 (Cys158Arg)	45411941	C	0.52	0.93	[0.95-1.03]	0.63	0.0
19	<i>APOE</i>	rs7412 (Arg112Cys)	45412079	T	0.87	0.98	[0.98-1.01]	0.087	65.9

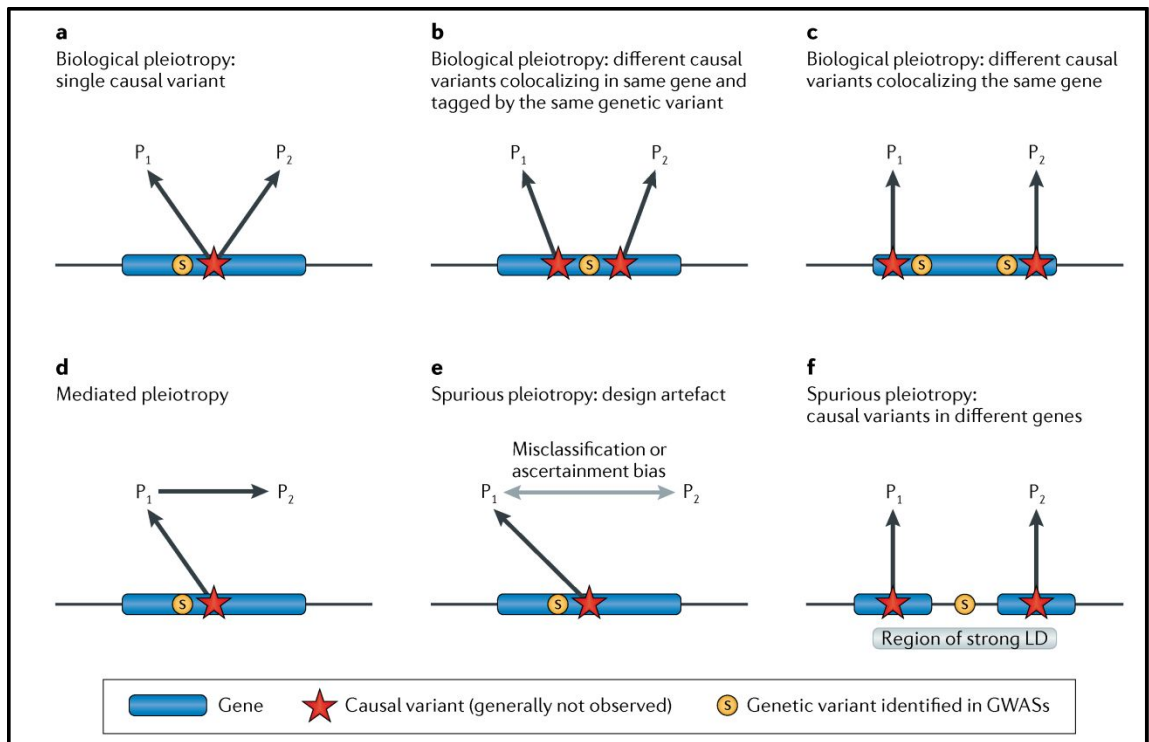
Abbreviations: Chr – chromosome; hg19 – human genome 19; CI – confidence interval;  $P_{META}$  – Fixed effects meta-analysis genetic association significance;  $P_{HET}$  – Heterogeneity test significance value

## 2.6 - DISCUSSION

Three independent genome-wide significant associations were identified in the genes/loci: *PNPLA3*, *TM6SF2/SUGP1* and *MBOAT7/TMC4*. *PNPLA3* contained the most significantly associated variant and had the largest odds ratio of the variants with genome-wide significant associations. This variant has already been identified as a risk variant for alcohol-related cirrhosis in several association studies<sup>413,428,431</sup>. This GWAS confirms the robustness of this association and demonstrates its prominent effect in comparison to other variants in these cohorts. Two other independent loci also contain variants that associate with alcohol-related cirrhosis at genome-wide significance, localising to the genes *TM6SF2/SUGP1* and *MBOAT7/TMC4*.

The genetic association at the *TM6SF2/SUGP1* locus was confirmed through the genotyping of the two variants rs58542926 and rs10401969. The variant rs58542926 occurs in *TM6SF2*, which is located on chromosome 19 at the coordinates 19,375,174-19,384,074 (human genome build 19) and encodes the 351 amino acid length protein Transmembrane 6 superfamily member 2 (TM6SF2). This variant is non-synonymous and is predicted to result in a glutamate to lysine amino-acid substitution at position 167 (Glu167Lys) in the encoded protein. This variant has already been associated with NAFLD through exome-wide association analysis<sup>229</sup> and replication genotyping<sup>261</sup>. *TM6SF2* is predominantly expressed in the liver and small intestine and at a sub-cellular level, localises to the endoplasmic reticulum and the intermediate compartment of human liver cells. Functional experimentation has demonstrated that TM6SF2 is a regulator of liver fat metabolism. The rs58542926 is also functional where carriage of either the opposing effects on the secretion of triglyceride-rich lipoproteins from the liver influencing either NAFLD or cardiovascular disease risk<sup>269</sup>. It is thought that TM6SF2 is required for normal very-dense lipoprotein secretion and that one of the rs58542926 variants alleles alter TM6SF2 function resulting in the accumulation of lipids in the liver<sup>229</sup>.

There is strong functional evidence implicating rs58542926 as a pathogenic variant in NAFLD and it therefore seems likely that this is the causal variant, which also contributes to alcohol-related cirrhosis risk. However, the variant rs10401969 in *SUGP1* associates with greater significance. A potential cause of this discrepancy could be pleiotropy (Figure 2-5). In support of this hypothesis, there are several genome-wide significant hits in this locus with a diversity of non-liver related phenotypes such as bipolar disorder<sup>299</sup> and schizophrenia<sup>366</sup> suggesting that this region may harbour several sources of genetic variation which could independently contribute to distinct phenotypic effects. Techniques such as Bayesian fine-mapping<sup>410</sup> could be useful to delineate the presence of multiple independent genetic associations.

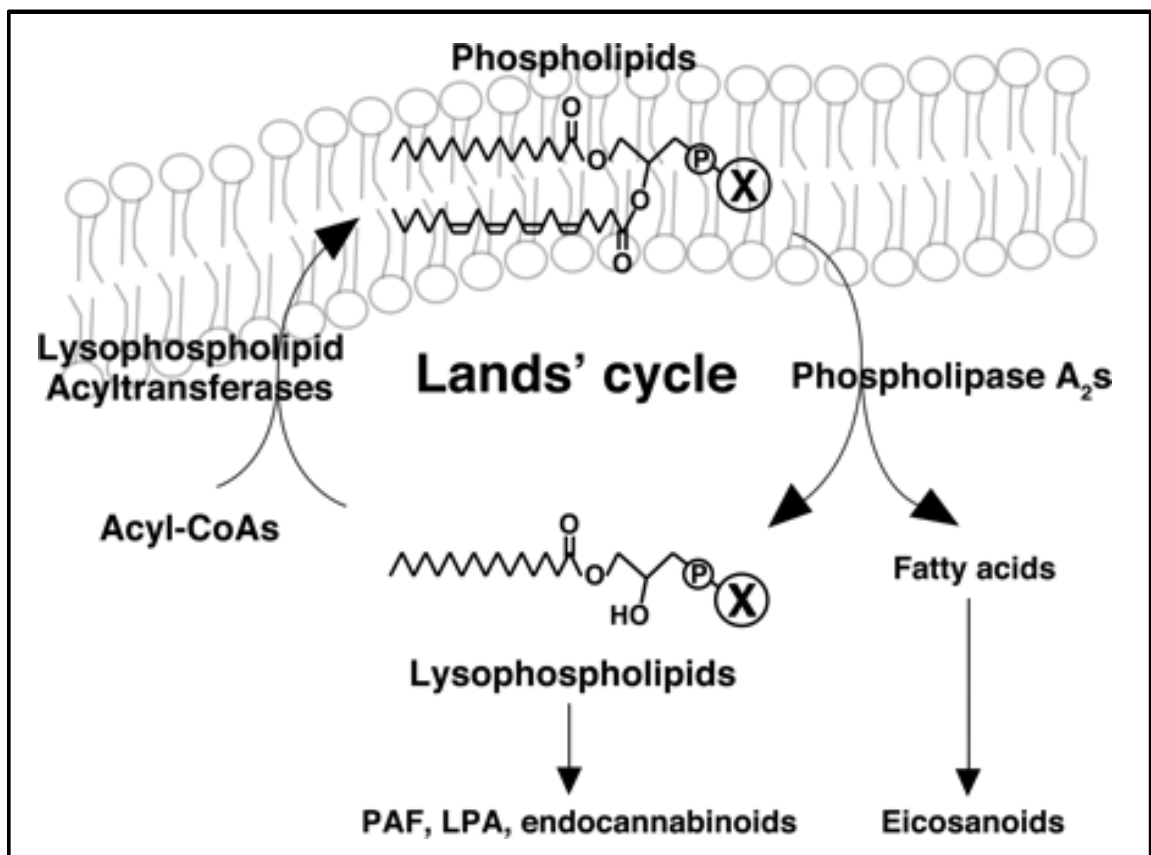


**Figure 2-5 Forms of pleiotropy between genetic variants and phenotypes at a locus**

A - Biological pleiotropy at the allelic level: the causal variant affects both phenotypes; B – Co-localizing association (biological pleiotropy): the observed genetic variant is in strong LD with two causal variants in the same gene that affect different phenotypes; C - Biological pleiotropy at the gene level: two independent causal variants in the same gene affect different phenotypes; D - Mediated pleiotropy: the causal variant affects  $P_1$ , which lies on the causal path to  $P_2$ , and thus an association occurs between the observed variant and both phenotypes; E - Spurious pleiotropy: the causal variant affects only  $P_1$ , but  $P_2$  is enriched for  $P_1$  owing to misclassification or ascertainment bias, and a spurious association occurs between the observed variant and the  $P_2$ ; F - Spurious pleiotropy: the observed variant is in LD with two causal variants in different genes that affect different phenotypes. Abbreviations:  $P_1$  – Phenotype 1;  $P_2$  – Phenotype 2; LD – Linkage Disequilibrium; GWAS - Genome-wide Association Study. Figure from Solovieff et al., 2013<sup>402</sup>

The genetic associations at the *MBOAT7/TMC4* locus were confirmed through the genotyping of the variant rs626283. This variant occurs in an intergenic region of high linkage disequilibrium between the genes *MBOAT7* and *TMC4* and is a known expression quantitative trait loci for *MBOAT7*<sup>91</sup>. This provides a hypothetical functional mechanism through which it may influence cirrhosis risk. *MBOAT7* is located on chromosome 19 at the coordinates 54,677,106-54,693,733 and encodes the 472 amino acid length protein, membrane bound o-acyltransferase domain containing 7. This protein has several aliases: *BB1*, *LRC4*, *LENG4*, *LPIAT*, *MBOA7*, *OACT7* and *hMBOA-7*. It is predicted to contain several transmembrane regions and is thought to be an integral membrane protein. It also has in vitro lysophospholipid acyltransferases activity with a substrate preference for arachidonic acid and because of this may be important for regulating free arachidonic acid levels and leukotriene synthesis in neutrophils<sup>141</sup>. Other members of the *MBOAT7* enzyme family play critically important roles in the remodelling phospholipids in cell membranes as part of the Lands'

cycle<sup>180,237</sup> (Figure 2-6). The Lands cycle also generates eicosanoids and endocannabinoids that may play important roles in the inflammatory processes which drive liver disease progression towards cirrhosis<sup>420</sup>. Genome-wide significant associations in *MBOAT7* have also been reported by other GWAS where the variants rs2576452 and rs8736 are significantly associated with the ratio of arachidonate and its metabolite 1-arachidonoylglycerophosphoinositol in the blood<sup>393</sup>. Notably, these quantitative trait loci occur in the same promoter region between *TMC4* and *MBOAT7* as the variant that is most significantly associated with alcohol-related cirrhosis providing a functional mechanism through which this variant may influence liver disease pathogenesis.



**Figure 2-6 Phospholipid metabolism in the Lands cycle**

Fatty acids of phospholipid are liberated by phospholipase A<sub>2</sub> and converted to eicosanoids. Lysophospholipids are also precursors of a different class of lipid mediators including PAF, LPA, and endocannabinoids. Lysophospholipids are converted to phospholipids in the presence of acyl-CoA by lysophospholipid acyltransferases. X indicates several polar head groups of phospholipids. Abbreviations: PAF – Poly Unsaturated; LPA - Lysophosphatidic Acid. Image from Hishikawa et al., 2008<sup>180</sup>

In contrast to many other GWAS, this study is of modest sample size. For example, GWAS cohorts for the study of schizophrenia are comprised of tens of thousands of DNA samples from which, over a hundred independent unique loci and risk variants have been identified<sup>381</sup> (many have odds ratio < 1.5). If such variants of a similar effect size contribute to alcohol-related cirrhosis risk, as would be predicted by a polygenic risk model, then the current work is largely underpowered to detect these genetic

associations. It is likely that there are still many alcohol-related cirrhosis risk loci that may be identified through genome-wide association analysis with sufficiently large sample sizes<sup>452</sup>.

This GWAS accounts for the vast majority of the common genetic variation in the genomes of two Northern-European ancestry populations and thus risk variants identified in this study may not be relevant in other ancestral populations. However, there are several reasons to suppose that this is not the case: (i) it is already known that rs738409 in *PNPLA3* is associated with both NAFLD and alcohol-related liver disease risk in diverse ancestral populations<sup>55,370</sup>; and (ii) both rs626283 in *MBOAT7* and rs58542926 in *TM6SF2* are common variants, in all major ancestral populations<sup>1</sup>.

Although this GWAS covers the majority of common genetic variation, it lacks coverage of copy number variants and rare variants (minor allele frequencies < 1%). Studies of lower frequency variation require increasing levels of statistical power and hence either large sample sizes or an analysis which is limited to a sub-set of variants. Copy number variation analysis typically requires the recalling of the raw fluorescence intensity data using a specialised calling algorithm<sup>471</sup>. Reanalysing these data, using such different methodologies, or with enlarged cohorts and other datasets, could account for these other sources of genetic variation and identify new cirrhosis risk loci.

There is a potential for GWAS to be confounded by population sub-structure whereby factors, unrelated to the phenotype of interest lead to spurious genetic associations. For example, the German no-liver disease controls used in the discovery analysis were primarily recruited for the purposes of a GWAS of alcohol-dependence<sup>433</sup> and thus may be of less phenotypic surety than those recruited from the UK, which were recruited at a hepatology clinic. To minimize error several precautions were undertaken: (i) patients of ambiguous ancestry were excluded via multi-dimensional scaling analysis with reference to the diverse HapMap<sup>192</sup> populations; (ii) the genomic inflation factor ( $\lambda_{GC}$ ) and the top-principal components were calculated in the data to detect and account for population stratification; (iii) an adjusted analysis was performed accounting for several variables which have potential to confound the genetic association with alcohol-related cirrhosis; and, (iv) allele frequencies were separately compared for variants of interest between cirrhosis cases and both no-liver disease controls and population controls. All of these measures indicated a homogeneous dataset, with phenotypic surety, in which the confounding of the association signal was limited.

Three gene sets, representative of three lipid metabolic pathways were tested for an overabundance of genetic associations in the UK dataset. The arachidonic acid metabolic pathway and the glycerophospholipid pathway gene sets both contained a



greater number of genetic associations with cirrhosis than would be expected by chance. It has long been known that disturbances in lipid processing<sup>374</sup> and metabolism<sup>51</sup> occur in chronic alcohol misuse, which contribute to the development of hepatic steatosis<sup>252</sup> and liver injury<sup>264</sup>. Serum arachidonic acid levels are known to be reduced in patients with cirrhosis<sup>204</sup> suggesting a link between this pathways analysis and experimental observation. However, this pathways analysis was selective rather than encompassing limiting the scope of these findings. Pathways analysis is a developing area of GWAS study<sup>447</sup> and there is clear potential for future studies to perform more detailed pathways analysis to validate these findings.

For the variants present in 17 previously studied alcohol-related liver disease candidate genes, there was limited evidence to support genetic associations with alcohol-related cirrhosis in the UK and German discovery cohorts. Variants in the *ADH1C*, *CYP2E1* and *GSTP1* loci were marginally associated with cirrhosis ( $P < 0.05$ ); however, there was significant heterogeneity in the meta-analysed data for the variants in both *ADH1C* and *CYP2E1*. As a fixed effects model was used to calculate these significance values, it is therefore likely that they overestimate the true significance. The only marginally associated variant without heterogeneity, rs1695 in *GSTP1*, encodes a non-synonymous substitution (Ile105Val) in the protein glutathione S-transferase pi 1. A recent meta-analysis identified significant genotypic associations between the Val/Val and Val/Ile genotype of rs1695 and alcohol-related cirrhosis ( $P = 0.03$ , OR = 2.04 95% CI [1.09-3.80])<sup>282</sup>. In this previous meta-analysis, other variants encoding null alleles of other glutathione-s-transferases, *GSTM1* and *GSTT1*, were also tested for association with risk. However, these null alleles are encoded by large insertion deletion variants and therefore were not covered by panel genotyping or imputation; it remains unknown whether these variants contribute to cirrhosis risk in the study population.

In summary, this first GWAS of alcohol-related cirrhosis demonstrates three genome-wide significant risk variants in the genes *PNPLA3*, *TM6SF2* and *MBOAT7*. These findings confirm previously identified genetic risk factors and identify novel risk loci. It is likely that increasing the number of patients studied via GWAS analysis has the potential to discover novel risk loci.

---

---

**CHAPTER 3 EXTENDED GENOME-WIDE  
ASSOCIATION STUDY**

---

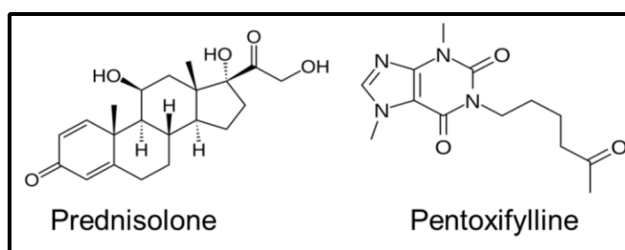
---

### 3.1 - OVERVIEW

Alcohol-related cirrhosis is a complex disease phenotype, likely with polygenic genetic risk factors. To maximise statistical power and validate new risk loci additional cohorts will be required. This chapter details an extended GWAS including an additional 850 severe-alcoholic hepatitis cases recruited through the Steroids or Pentoxifylline for Alcoholic Hepatitis Trial (STOPAH). The GWAS data for 322 STOPAH cases underwent harmonisation and merging with the UCL alcohol-related cirrhosis GWAS dataset. This expanded dataset underwent genetic association analysis followed by a meta-analysis with the genetic association results from the German alcohol-related cirrhosis GWAS dataset. Variants of interest underwent replication genotyping in an independent cohort of 528 STOPAH severe alcoholic hepatitis cases and 873 UCL alcohol dependent controls.

### 3.2 - BACKGROUND

Severe alcoholic hepatitis is a clinically evident form of alcohol related liver disease resulting from chronic alcohol misuse, which frequently co-occurs with alcohol-related cirrhosis<sup>285</sup>. It has a poor short-term prognosis and treatment options for it are limited with current guidelines<sup>59,285</sup> recommending two pharmaceuticals: the glucocorticoid based prednisolone and the phosphodiesterase inhibitor pentoxifylline (Figure 3-1). There have been several clinical trials of these treatment options, however, some studies are conflicting and hence controversy exists regarding their efficacy<sup>399</sup>. A recent systematic network review suggests that both drugs either alone or in combination reduce short-term (4 weeks) mortality although no treatment reduces medium term (3-12) mortality<sup>399</sup>. The largest study included in this systematic network review was the UK based STOPAH trial<sup>129</sup>. Germane to the subject of GWAS, the DNA samples from a significant proportion of the STOPAH cohort have been extracted and one third of these have undergone genome-wide genotyping as part of the first, and ongoing, GWAS of severe alcoholic hepatitis.



**Figure 3-1 the chemical structures of prednisolone and pentoxifylline**

The GWAS of alcohol-related cirrhosis validated three independent loci: *PNPLA3*, *TM6SF2* and *MBOAT7*. However, it is likely that there are undiscovered alcohol-

related cirrhosis risk variants, as is consistent with a polygenic model of genetic risk. The joint analysis of independent GWAS datasets has the potential to increase statistical power and identify novel loci. There are two approaches through which independent GWAS datasets may undergo joint genetic association analysis: (i) the merging of independent GWAS datasets before undergoing genetic association analysis; or, (ii) the meta-analysis of the genetic association results from independent GWAS datasets. The use of these approaches are not mutually exclusive.

The merging of GWAS datasets is a complex process involving joining datasets containing the genotype information for hundreds of thousands of genetic variants, in potentially thousands of individuals; this process has the potential to introduce substantial error unless stringent quality control procedures are adhered to<sup>385,484</sup>. Joint analysis is particularly useful when a significant proportion of the same DNA samples overlap between different datasets. Before merging, dataset compatibility must be assessed by a number of criteria, primarily: (i) is the phenotypic information sufficiently similar between datasets to allow accurate and equivalent comparison; (ii) is sample demography equivalent between datasets; and, (iii) is between GWAS panel genotype concordance sufficient to cover the majority of common variation. The process of dataset merging requires that all concordant variants present in both datasets are aligned to the same DNA strand, have the same reference genomic coordinates and have the same allele-coding schema. Ambiguous SNPs in particular have a significant potential to confound the harmonisation process. Following merging, a dataset may be treated as if a primary dataset and undergo genome-wide imputation and genetic association analysis.

Meta-analysis is a statistical method for systematically combining the results of different studies to provide a single summary result; it is widely used in GWAS<sup>119</sup>. It is advantageous as it allows datasets to be co-analysed when merging of datasets is unsuitable (e.g. significant ancestral differences between cohorts, primary genotype data is unavailable...). Before such an analysis can occur, dataset compatibility must be assessed, primarily: (i) is phenotypic information sufficiently similar between datasets to allow accurate and equivalent comparison; (ii) is there significant statistical heterogeneity between the results of each GWAS dataset; (iii) does each genetic association file contain equivalent measures of genetic association (e.g. allelic, dominant or recessive models) and effect size (e.g. OR or beta, 95% confidence interval or standard error); and, (iv) are the variants in each genetic association results dataset concordant. The latter criterion is generally not an issue as most GWAS contain imputed genotype data and therefore have high overlapping genotype coverage and the universal alignment of allele codes to the forward strand. A more significant issue is statistical heterogeneity, or variance between results data.

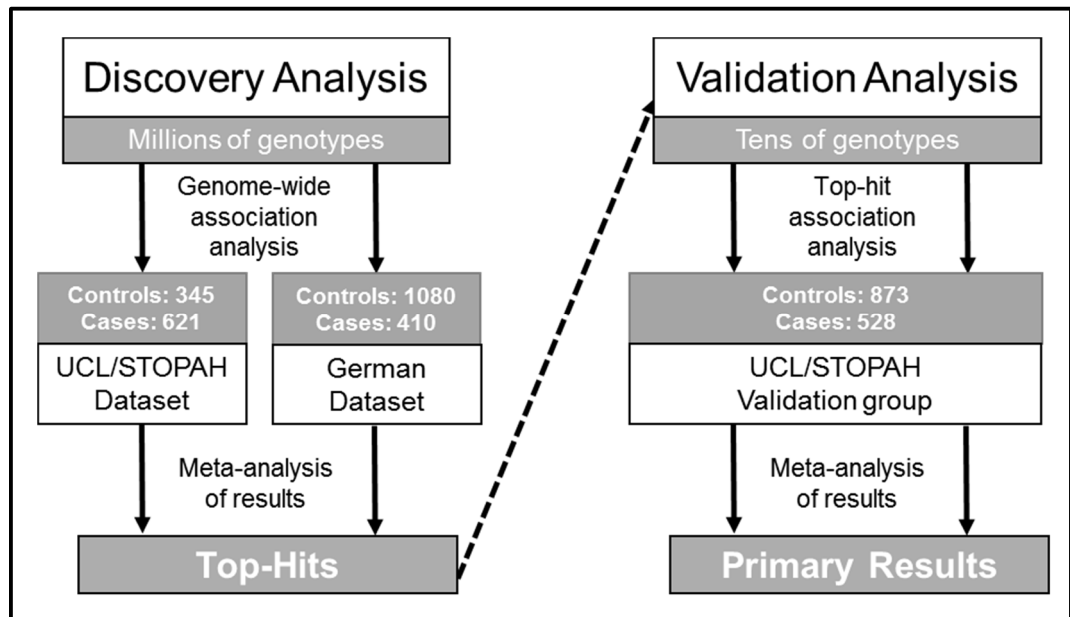
Heterogeneity is widely quantified using the Cochran's Q statistic<sup>68</sup>. From this statistic, a significance value may be calculated that indicates the probability of their being heterogeneity in the data between the samples. It is also possible to calculate a percentage measure of the total variation between studies<sup>176</sup>, known as the  $I^2$ . The presence of significant heterogeneity influences the choice of meta-analytical using either a fixed effects or random effects model. Fixed effects models are the widely used for GWAS meta-analysis. However, in the presence of between-study heterogeneity, this model may underestimate confidence intervals and overestimate significance values<sup>87</sup>. Random effects models alternately assume variance between results datasets and are appropriate in the presence of statistical heterogeneity.

### **3.3 - AIM**

The primary objective of this study was to merge the GWAS dataset for the STOPAH severe alcoholic hepatitis cases with the UCL alcohol-related cirrhosis GWAS dataset and to perform an enlarged genome-wide association analysis focussing on the identification and verification of novel loci.

### **3.4 - MATERIALS AND METHODS**

This study was by design, split into two stages: (i) a discovery analysis in which an enlarged UCL/STOPAH GWAS dataset underwent genome-wide association analysis followed by meta-analysis with the German GWAS dataset results; and (ii) a validation stage in which novel variants identified during the discovery analysis underwent replication/validation genotyping in the validation STOPAH cases and UCL controls (Figure 3-2).



**Figure 3-2 The study design plan of the extended genome-wide association study**

### 3.4.1 - COHORTS

#### UCL COHORT

The UCL cohort used in this analysis, comprised of 1,521 alcohol dependent samples that were used in either the discovery or the validation analysis: (i) there were 302 alcohol-related cirrhosis cases and 346 no-significant liver injury controls that underwent genome-wide genotyping in the discovery analysis; and (ii) there were 873 alcohol-dependent samples without clinically evident alcohol-related liver disease that were used as controls in the validation analysis (Table 3-1).

#### STOPAH COHORT

The STOPAH cohort comprised solely of cases with severe alcoholic hepatitis and were recruited as per a published protocol<sup>427</sup>. All of these cases had a long history of alcohol misuse and compatible clinical, and/or biopsy features of severe alcoholic hepatitis. Subjects that had a prolonged history of jaundice (>3 months) or whose liver biochemistry was not suggestive of alcohol-related hepatitis (ALT >300 international units per litre, AST >500 international units per litre) were excluded. All of these cases were British, although their ancestral information was limited to a broad classification of self-reported ethnic group (e.g. white, South Asian, black etc.). All cases that were of none white ancestry were excluded so in total, 850 severe alcoholic hepatitis cases were included from the STOPAH cohort.

DNA samples were extracted from whole blood using a commercially available Blood DNA extraction mini kit (QIAGEN, Manchester, UK). These samples were split into two

groups that were used in the discovery and validation analyses: (i) the STOPAH exploratory group comprising of 322 severe alcoholic hepatitis cases which were used in the in the discovery analysis (Table 3-1); and, (ii) STOPAH validation group comprising of 528 severe alcoholic hepatitis cases which were used in the validation analysis (Table 3-2).

Table 3-1 Demographic features of the UCL and STOPAH discovery group

<b>Analysis</b>	<b>UCL</b>		<b>STOPAH</b>
<b>Group</b>	<b>Case</b>	<b>Controls</b>	<b>Case</b>
Number	302	346	322
Genome-wide genotyping	Yes	Yes	Yes
Median age years (IQR)	53 (46-60)	48 (41-55)	48 (41-55)
Gender (% male)	67.6%	76.6%	67.4%
Proven alcohol-related cirrhosis (%)	100%	0	37.8%

Table 3-2 Demographic features of the UCL and STOPAH validation group

<b>Analysis</b>	<b>UCL</b>	<b>STOPAH</b>
<b>Group</b>	<b>Controls</b>	<b>Case</b>
Number	873	528
Genome-wide genotyping	No	No
Median age years (IQR)	45 (39–52)	49 (42 - 57)
Gender (% male)	65.4%	60.8%
Proven alcohol-related cirrhosis (%)	0	48.4%

## 3.4.2 - DISCOVERY ANALYSIS

### GWAS DATASETS

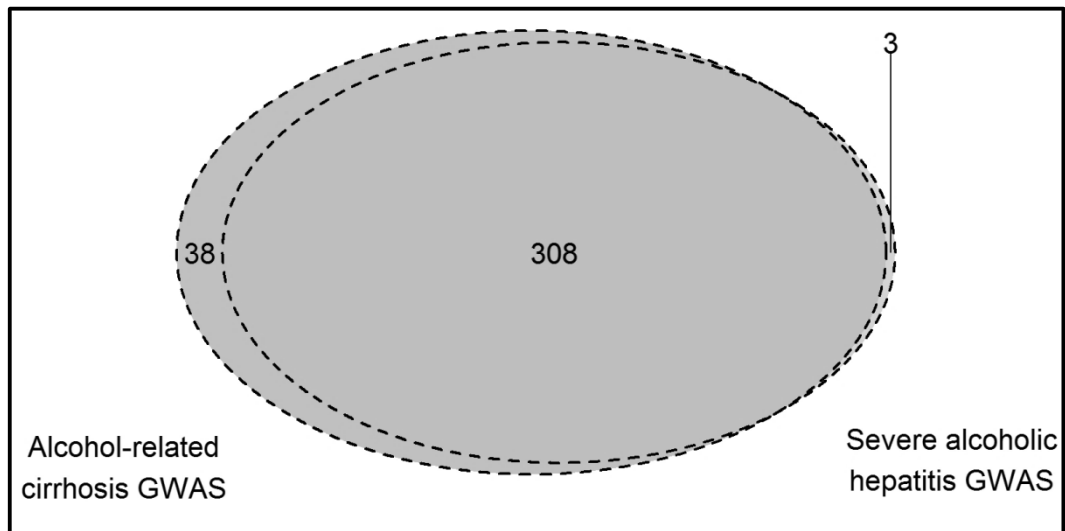
#### UCL Alcohol-related Cirrhosis GWAS Dataset

The UCL exploratory cohort comprised of alcohol-related cirrhosis cases and no-significant controls that underwent genome-wide genotyping as part of the first alcohol-related cirrhosis GWAS (chapter 2). This dataset contains the genotype information for 609,879 variants in 346 alcohol-related cirrhosis cases and 302 no-significant liver injury controls.

#### STOPAH-UCL Severe Alcoholic Hepatitis GWAS Dataset

The STOPAH severe alcoholic hepatitis GWAS dataset forms part of an ongoing GWAS of severe-alcoholic hepatitis. The samples in this dataset include 322 severe alcoholic hepatitis cases from the STOPAH cohort and 311 no-significant liver injury controls from the UCL cohort; these controls were the same as those used in the initial alcohol-related cirrhosis GWAS (Figure 3-3). All samples were genotyped on an Illumina Core Exome array (Illumina, San Diego, USA) at the Wellcome Trust Sanger Institute. At the Wellcome Trust Sanger Institute, raw fluorescence intensities underwent genotype calling using two algorithms: the zCall algorithm for rare variants (allele frequency < 1%)<sup>146</sup> and the proprietary GenCall algorithm (Illumina, San Diego, USA) for all other variants. This raw genotype data underwent quality control including per sample quality control (gender checks, duplicate sample checks, heterozygosity ( $\pm 3$  standard deviations from mean), relatedness ( $\hat{\pi} \geq 0.185$ ) and removal of tri-allelic variants) and per variant quality control (genotyping success rate, Hardy-Weinberg equilibrium ( $P_{\text{HWE}} \leq 1 \times 10^{-6}$ )), case/control differences in genotyping rates and removal of redundant genotypes for those genotyped more than once). This data was stored in binary PLINK format.





**Figure 3-3 The overlap in no-significant liver injury controls**

This Venn diagram shows the proportion of shared controls between the UCL alcohol-related cirrhosis GWAS dataset and the STOPAH severe alcoholic hepatitis GWAS dataset

## GWAS DATASET PROCESSING

### Data Harmonisation

The STOPAH GWAS dataset was harmonised with the UCL alcohol-related cirrhosis GWAS dataset using the software Genotype Harmonizer (version 1.4.15)<sup>89</sup>. The naming of variants in all datasets were standardized to those the UCL dataset schema. Variants that were none concordant between datasets and those that did not meet selected criteria (minor allele frequency < 0.01, minimum linkage disequilibrium with at least 3 neighbouring variants  $r^2 < 0.3$ ) were removed from the merged dataset.

### Duplicate Quality Control

Duplicate samples were identified in both the UCL and STOPAH harmonised datasets; these samples were extracted and merged in PLINK (version 1.9)<sup>58</sup>. Variants with conflicting genotypes between duplicate samples or any sample with a conflicting genotyping rate (< 98%) or differential missing genotype rates ( $P_{\text{missing}} < 5 \times 10^{-4}$ ) were excluded.

### Data Merging

The harmonized datasets were merged using default settings in PLINK (version 1.9). All STOPAH samples and UCL alcohol-related cirrhosis samples were coded as cases and the UCL no-liver disease samples were coded as controls. The harmonised, merged dataset underwent sample level quality control. This included identical by descent estimation to infer relationship and calculation of autosomal heterozygosity to identify abnormal samples. Related samples that were no closer than second cousins ( $\hat{\pi} \geq 0.185$ ) and samples with abnormal autosomal heterozygosity ( $\pm 3$  standard

deviations from the group autosomal heterozygosity mean) were excluded from the dataset. A principal components analysis was performed on a linkage-disequilibrium<sup>346</sup> pruned dataset to visually confirm genetic homogeneity within the merged dataset.

### Imputation

The merged and quality controlled UCL/STOPAH dataset underwent haplotype phasing using SHAPEIT (version v2.r790)<sup>312</sup>. The phased data underwent imputation using Impute2 (version 2.3.2)<sup>277</sup> using the 1000 genomes phase 3 haplotype data as a reference<sup>1</sup>. Standard parameters were used for imputation (buffer = 250 kb, effective size = 20,000 and input threshold = 0.9) sequentially on 500 megabase interval chunks of each autosome. The output GENS files were concatenated and converted into the Oxford BGEN format using QCTOOL (version 1.4) excluding variants of poor imputation quality (info score < 0.3) and low minor-allele frequency (< 1%).

## **GENETIC ASSOCIATION ANALYSIS**

### Discovery Association Analysis

Genetic association analysis was performed on the merged UCL-STOPAH dataset using SNPTEST (version 2.5.1) under a frequentist additive model on genotype dosages (i.e. method expected). Subsequent analyses of genetic association results were performed using R (version 3.2.2) using the GENABEL package for genomic inflation factor estimation and the ggplot2 package for data visualization.

### Meta-Analysis

The UCL/STOPAH GWAS genetic association results dataset underwent a fixed-effect meta-analysis with the German alcohol-related cirrhosis GWAS dataset using the software META (version 1.3.2)<sup>278</sup>. The German GWAS genetic association data contained the results for 6,866,425 imputed and directly genotyped variants that had undergone allelic logistic regression based association test conditioning on gender and the top-principal component (chapter 2) (Table 3-3).

Table 3-3 Demographics of the German discovery group

German Cohort		
Analysis	Meta	Meta
Group	Alcohol-related cirrhosis Cases	No-significant liver injury controls
Number	410	1080
Genome-wide genotyping	Yes	Yes
Median age years (IQR)	53 (47-61)	42 (36-48)
Gender (% Male)	71	100
Proven alcohol-related cirrhosis (%)	100%	0

Abbreviations: IQR – inter-quartile range

### 3.4.3 - VALIDATION ANALYSIS

#### SNP SELECTION AND GENOTYPING

Variants were selected for validation genotyping based on three criteria:

- (i) The genetic association  $P$ -value is below the marginal significance threshold ( $P_{META} \leq 1 \times 10^{-5}$ )
- (ii) The variant is in an independent locus (i.e. is not in linkage disequilibrium with any other variant undergoing validation genotyping)
- (iii) The locus has not been identified in the previous alcohol-related cirrhosis GWAS

Those identified variants were genotyped in-house using the K-Biosciences allele specific PCR (KASPAR) (LGC Genomics, Hoddesdon, UK) genotyping platform and custom designed primers. If the variant was unsuitable for PCR based genotyping (e.g. due to the location in a highly repetitive region of nucleotide sequence) then a nearby variant in strong linkage disequilibrium was genotyped as a proxy. The PCR amplification and detection of fluorescence data was performed on a LightCycler® 480 Real-Time PCR machine (Roche Molecular Diagnostics, Burgess Hill, UK). Genotype calling was performed automatically using proprietary software (Roche Molecular Diagnostics, Burgess Hill, UK)<sup>367</sup> with some minor manual editing of genotype calls.

#### ANALYSIS

The raw genotype data underwent quality control (e.g. removal of conflicting samples, removal of duplicates) and was converted into PED and MAP using a custom designed

R package, KASParclean. Further quality controls and genetic association tests including tests of Hardy-Weinberg equilibrium, differential genotype missing rate tests were performed using PLINK<sup>58</sup> (version 1.9).

### Concordance analysis

A proportion of samples from both the STOPAH and UCL cohort, with imputed genotype data, underwent direct genotyping to validate the accuracy of imputed genotype calls. The concordance between imputed and direct genotype data was directly assessed using a custom package<sup>79</sup> in R (version 3.2.2) from which a concordance value was determined. The correlation between this concordance value and the imputation accuracy (Impute2 info score) was assessed using linear regression.

### Genetic association analysis

Genetic association tests were performed by comparing the allele frequencies between the validation STOPAH severe alcoholic hepatitis cases and the validation UCL alcohol-dependent controls. The discovery genetic association was performed in PLINK<sup>58</sup> (version 1.9) using an allelic logistic regression model conditioning on gender as a covariate. Validation genetic association data were converted into an appropriate format and underwent both a fixed effects and random effects meta-analysis using the software META<sup>278</sup> (version 1.3.2).

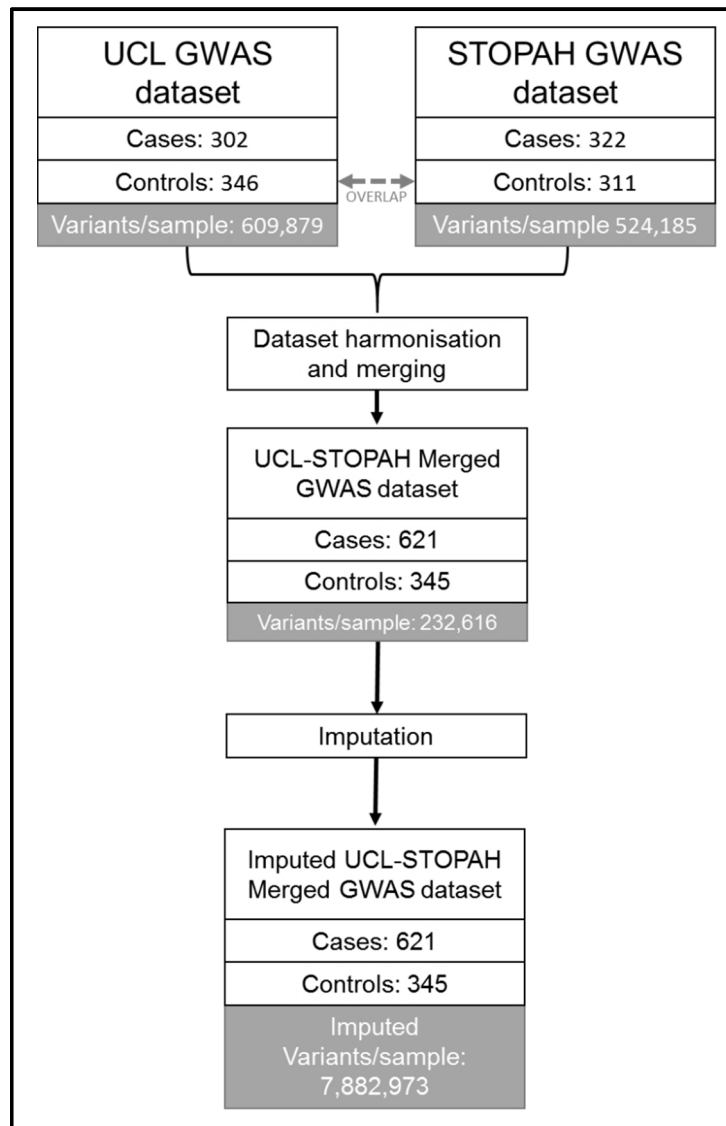
## **3.5 - RESULTS**

### **3.5.1 - GWAS DATASET PROCESSING**

Before imputation, there were genotypes for 233,076 variants in the UCL/STOPAH GWAS dataset, (Figure 3-4); these represent those variants that are concordant between the original UCL and STOPAH GWAS datasets. In the 308 overlapping controls with no significant liver injury, the mean conflict rate between the genotype calls was 16 conflicts per variant (0.0069%) all of which occurred in 446 variants. In total, 15 variants ( $P_{\text{threshold}} < 5 \times 10^{-4}$ ) had significantly different missing genotype rates. An autosomal heterozygosity check identified five samples with abnormal autosomal heterozygosity; these samples were removed from the merged dataset. Identity-by-descent analysis identified a single case that was independently recruited into both the STOPAH and UCL studies. Following these stages of quality control, there were genotypes for 232,616 variants in 932 samples (Figure 3-4).

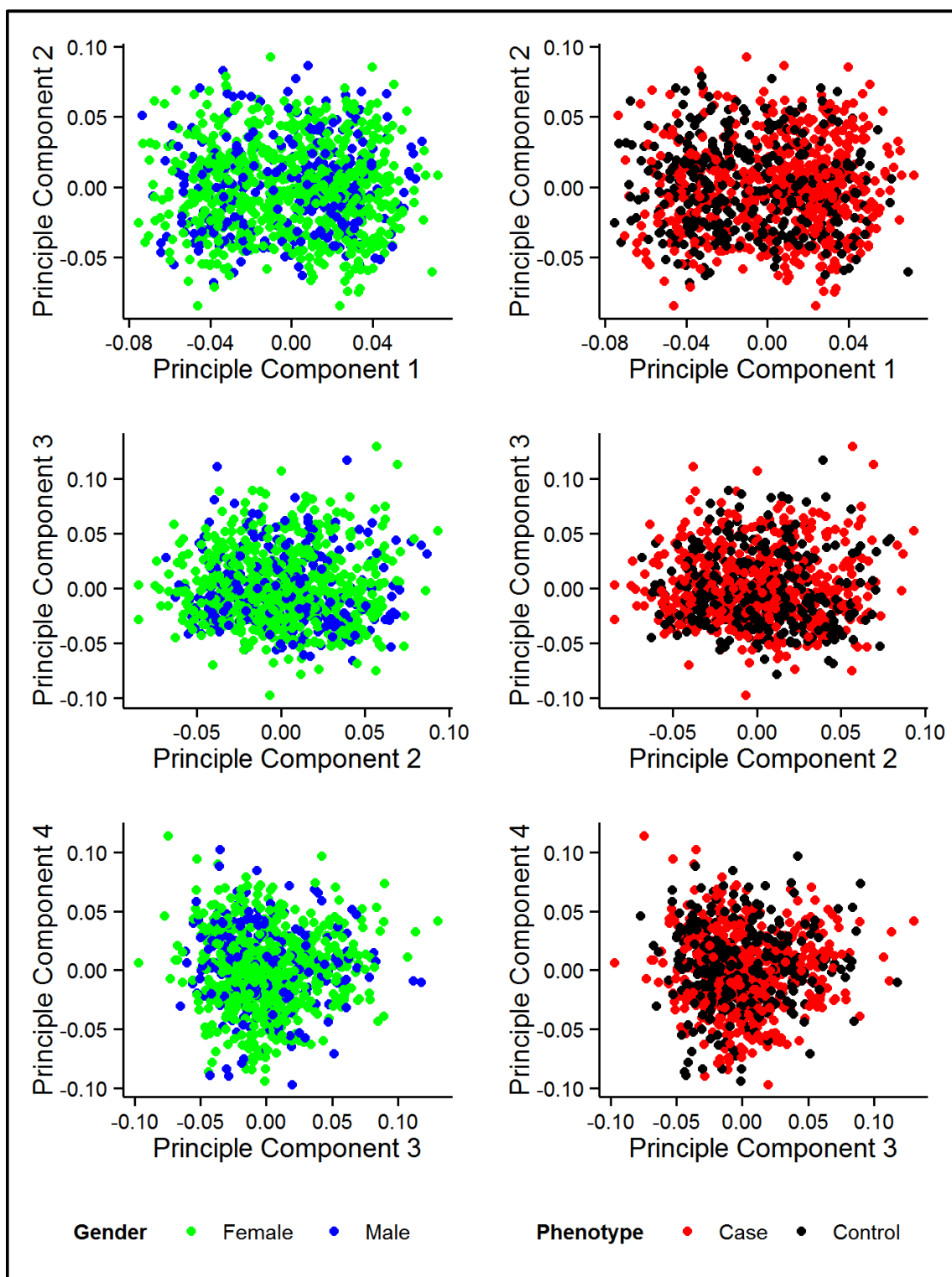
A post-quality control principal components analysis did not reveal significant stratification issues in the dataset (Figure 3-5). The merged UCL/STOPAH dataset was split into separate files by each autosome and underwent imputation producing

genotype information for 81,454,666 variants. The majority of this information was either redundant or of poor quality as when crudely filtering on imputation quality and minor allele frequency (info score < 0.3; minor allele frequency < 1%) only 9,613,508 variants were retained in the dataset (Figure 3-4). Of these, 7,882,973 (81.9%) variants were of sufficient imputation quality for genetic association analysis (impute2 info score  $\geq$  0.8).



**Figure 3-4 A schematic of the stages of the discovery analysis**

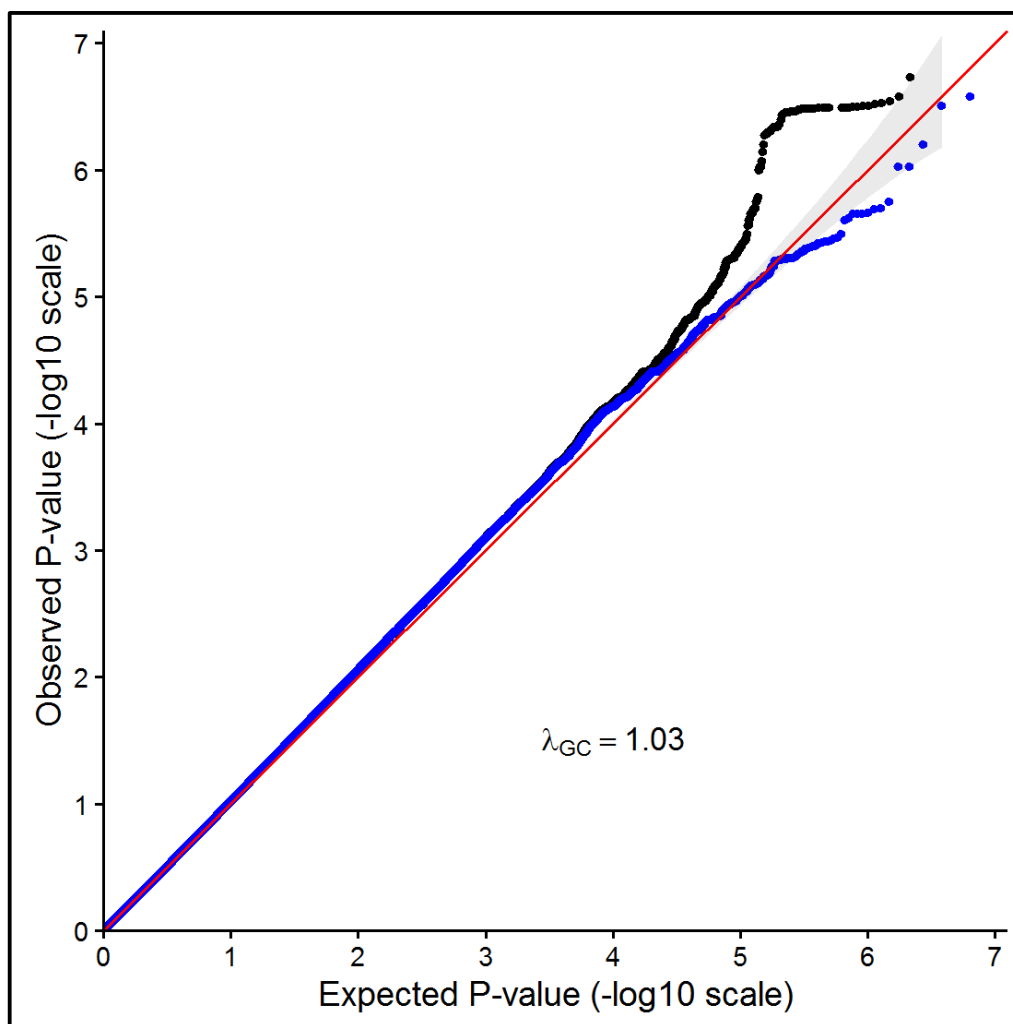
The UCL and STOPAH GWAS datasets were underwent harmonisation and merging. A number of controls overlapped between these datasets. The merged dataset underwent quality control followed by imputation including autosomal variants only.



**Figure 3-5 Principal component plot in the UCL-STOPAH merged dataset**  
 The top three principal components are compared for each sample by plotting on X-Y scatter plot and colouring by phenotype status (female = green, male = blue) or phenotype status. The clustering of points into a single group, and is suggestive of a homogeneous dataset.

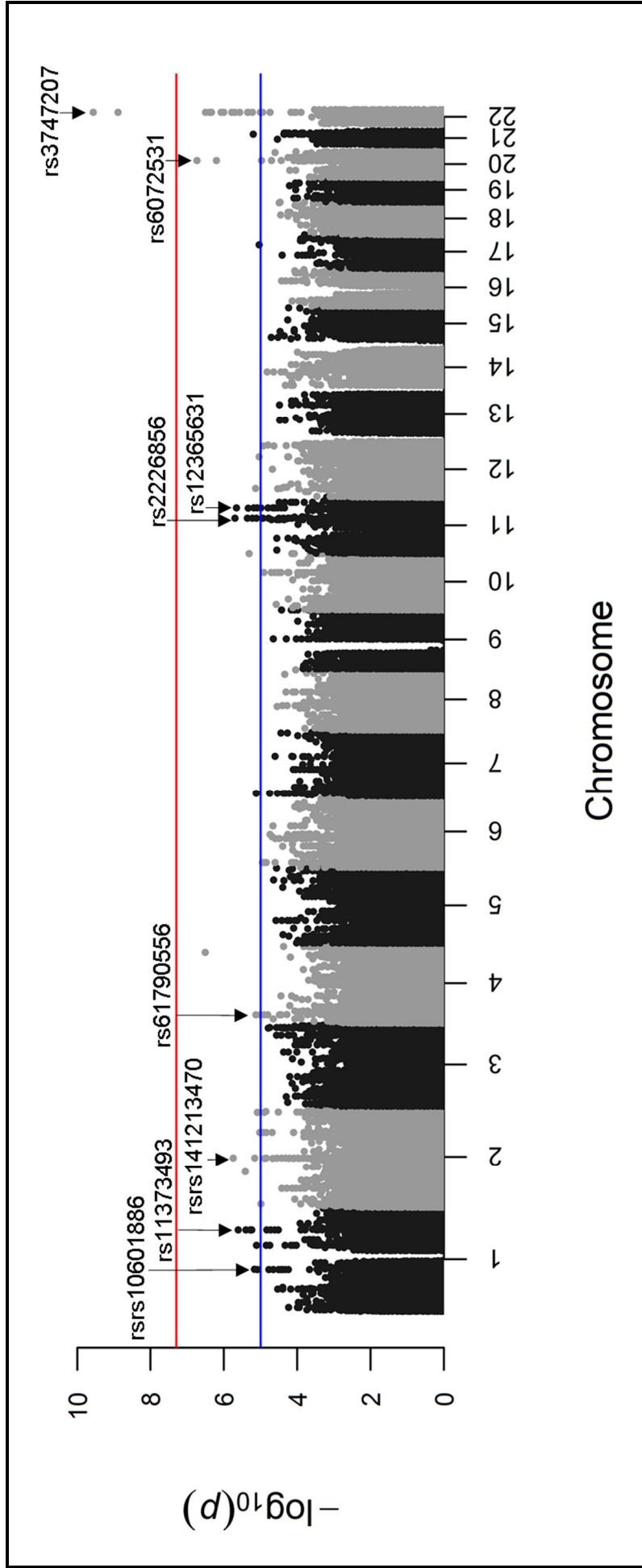
### 3.5.2 - DISCOVERY ASSOCIATION ANALYSIS

In the discovery association analysis, there were 621 alcohol-related liver disease cases and 345 no-significant liver injury controls. There was not any indication for population stratification (Figure 3-6) as measured by the genomic inflation factor ( $\lambda_{GC} = 1.04$ , standard error =  $3.05 \times 10^{-6}$ ). The *PNPLA3* locus contains the greatest number of significantly associated variants including several genome-wide significant hits ( $P \leq 5 \times 10^{-8}$ ) (Figure 3-7). In total 146 variants had a  $P$ -value below the marginal significance threshold ( $P \leq 1 \times 10^{-5}$ ). These variants occurred in 19 distinct loci, although only eight contained multiple supporting variants below the marginal significance threshold (Table 3-4, Figure 3-7). Two of these occur in genes: rs10601886 in *RTCA* and rs2226856 in *DLG2*.



**Figure 3-6 Quantile-quantile plot of the genetic association results in the extended UCL/STOPAH dataset**

This plot shows the distribution of  $P$ -values in the entire dataset (black dots) and excluding  $P$ -values from the *PNPLA3* locus (blue dots). The red line represents the expected distribution of  $P$ -values under the null hypothesis and the grey area represents a 95% confidence interval for the expected distribution. The genomic inflation factor,  $\lambda_{GC}$  is estimated from all of the  $P$ -values.



**Figure 3-7 A Manhattan plot of the extended UCL/STOPAH analysis** The genetic association results for each variant are given by position along the chromosome (x-axis) versus the association  $P$ -value on a  $-\log_{10}$  scale. The blue line denotes the marginal significance threshold ( $P \leq 1 \times 10^{-5}$ ) and the red line a genome wide significant threshold ( $P \leq 5 \times 10^{-8}$ ). The peaks of the most significantly associate variants are highlighted by their name.



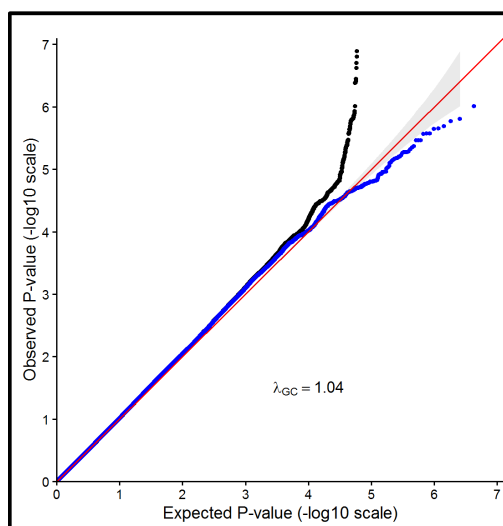
Table 3-4 The most-significant genetic associations\* in the UCL/STOPAH dataset

Chromosome	Variant	Position	Major allele	Minor allele	Impute2 info score	Imputed Genotypes	Minor allele frequency	P	Odds Ratio	95% Confidence Interval																																																																																										
1	rs10601886	100740606	GAACTT	G	0.96	Cases	595.7/25.3/0	6.77X10 <sup>-6</sup>	0.30	[0.18-0.51]																																																																																										
						Controls	305.2/39.8/0				5.8%	1	rs11373493	194818918	A	AT	0.95	Cases	567.2/50.8/3	2.46X10 <sup>-6</sup>	5.46	[2.27-13.1]	Controls	338.5/6.5/0	0.9%	2	rs141213470	118788583	A	T	0.93	Cases	608.8/12.2/0	1.78X10 <sup>-6</sup>	0.19	[0.09-0.4]	Controls	315.6/29.4/0	4.3%	4	rs61790556	19820482	C	T	0.99	Cases	346.8/244.5/29.7	7.29X10 <sup>-6</sup>	0.62	[0.5-0.76]	Controls	151.5/151.6/41.9	34.1%	11	rs2226856	84548416	C	G	0.96	Cases	126/323.7/171.3	2.00X10 <sup>-6</sup>	1.61	[1.32-1.96]	Controls	111.3/169.7/64	43.1%	11	rs12365631	108971949	C	G	0.85	Cases	555/64.9/1	2.21X10 <sup>-6</sup>	4.54	[2.21-9.34]	Controls	333.6/11.4/0	1.7%	20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]	Controls	153.9/152/39.1	33.4%	22	rs3747207	44324855	G	A
1	rs11373493	194818918	A	AT	0.95	Cases	567.2/50.8/3	2.46X10 <sup>-6</sup>	5.46	[2.27-13.1]																																																																																										
						Controls	338.5/6.5/0				0.9%	2	rs141213470	118788583	A	T	0.93	Cases	608.8/12.2/0	1.78X10 <sup>-6</sup>	0.19	[0.09-0.4]	Controls	315.6/29.4/0	4.3%	4	rs61790556	19820482	C	T	0.99	Cases	346.8/244.5/29.7	7.29X10 <sup>-6</sup>	0.62	[0.5-0.76]	Controls	151.5/151.6/41.9	34.1%	11	rs2226856	84548416	C	G	0.96	Cases	126/323.7/171.3	2.00X10 <sup>-6</sup>	1.61	[1.32-1.96]	Controls	111.3/169.7/64	43.1%	11	rs12365631	108971949	C	G	0.85	Cases	555/64.9/1	2.21X10 <sup>-6</sup>	4.54	[2.21-9.34]	Controls	333.6/11.4/0	1.7%	20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]	Controls	153.9/152/39.1	33.4%	22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]	Controls	226.4/108.8/9.7	18.6%					
2	rs141213470	118788583	A	T	0.93	Cases	608.8/12.2/0	1.78X10 <sup>-6</sup>	0.19	[0.09-0.4]																																																																																										
						Controls	315.6/29.4/0				4.3%	4	rs61790556	19820482	C	T	0.99	Cases	346.8/244.5/29.7	7.29X10 <sup>-6</sup>	0.62	[0.5-0.76]	Controls	151.5/151.6/41.9	34.1%	11	rs2226856	84548416	C	G	0.96	Cases	126/323.7/171.3	2.00X10 <sup>-6</sup>	1.61	[1.32-1.96]	Controls	111.3/169.7/64	43.1%	11	rs12365631	108971949	C	G	0.85	Cases	555/64.9/1	2.21X10 <sup>-6</sup>	4.54	[2.21-9.34]	Controls	333.6/11.4/0	1.7%	20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]	Controls	153.9/152/39.1	33.4%	22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]	Controls	226.4/108.8/9.7	18.6%																			
4	rs61790556	19820482	C	T	0.99	Cases	346.8/244.5/29.7	7.29X10 <sup>-6</sup>	0.62	[0.5-0.76]																																																																																										
						Controls	151.5/151.6/41.9				34.1%	11	rs2226856	84548416	C	G	0.96	Cases	126/323.7/171.3	2.00X10 <sup>-6</sup>	1.61	[1.32-1.96]	Controls	111.3/169.7/64	43.1%	11	rs12365631	108971949	C	G	0.85	Cases	555/64.9/1	2.21X10 <sup>-6</sup>	4.54	[2.21-9.34]	Controls	333.6/11.4/0	1.7%	20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]	Controls	153.9/152/39.1	33.4%	22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]	Controls	226.4/108.8/9.7	18.6%																																	
11	rs2226856	84548416	C	G	0.96	Cases	126/323.7/171.3	2.00X10 <sup>-6</sup>	1.61	[1.32-1.96]																																																																																										
						Controls	111.3/169.7/64				43.1%	11	rs12365631	108971949	C	G	0.85	Cases	555/64.9/1	2.21X10 <sup>-6</sup>	4.54	[2.21-9.34]	Controls	333.6/11.4/0	1.7%	20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]	Controls	153.9/152/39.1	33.4%	22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]	Controls	226.4/108.8/9.7	18.6%																																															
11	rs12365631	108971949	C	G	0.85	Cases	555/64.9/1	2.21X10 <sup>-6</sup>	4.54	[2.21-9.34]																																																																																										
						Controls	333.6/11.4/0				1.7%	20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]	Controls	153.9/152/39.1	33.4%	22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]	Controls	226.4/108.8/9.7	18.6%																																																													
20	rs6072531	40511892	A	T	0.89	Cases	363.5/227.2/30.3	1.86X10 <sup>-7</sup>	0.55	[0.44-0.69]																																																																																										
						Controls	153.9/152/39.1				33.4%	22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]	Controls	226.4/108.8/9.7	18.6%																																																																											
22	rs3747207	44324855	G	A	0.95	Cases	302.1/255.8/63.1	2.28X10 <sup>-10</sup>	2.09	[1.65-2.65]																																																																																										
						Controls	226.4/108.8/9.7				18.6%																																																																																									

\*The results are only given in loci with multiple supporting variants below the marginal significance threshold

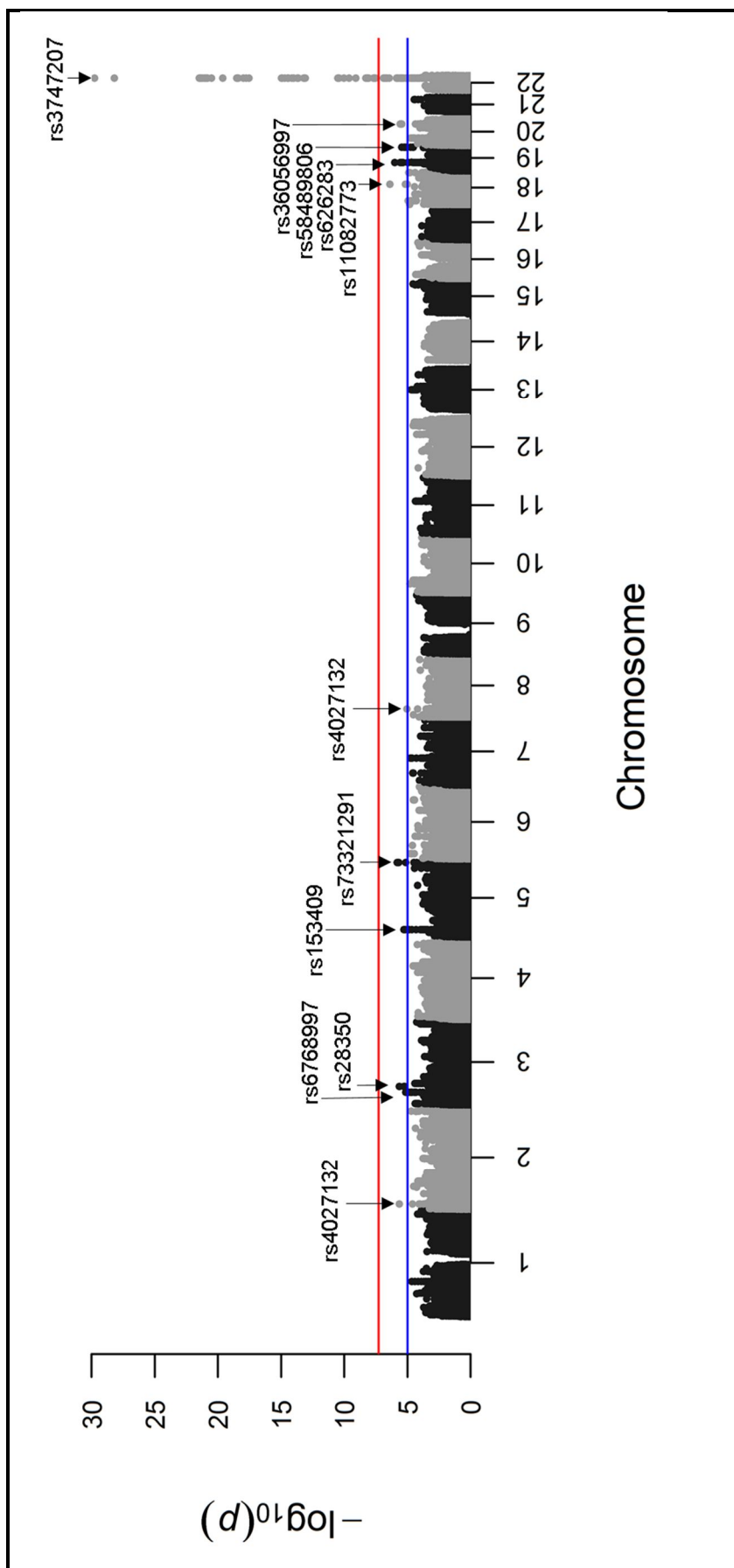
## META-ANALYSIS

In the discovery association meta-analysis, the UCL/STOPAH discovery genetic association results from the 621 alcohol-related liver disease cases and the 345 no-significant liver injury controls were meta-analysed with the German discovery genetic association results from 410 alcohol-related cirrhosis cases and 1080 no-significant liver injury controls. In total, 6,440,581 variants were concordant, and suitable for meta-analysis, between the UCL/STOPAH and the German datasets (Figure 3-9). The distribution of  $P$ -values in the meta-analysis was not indicative of systematic deviation from the expected distribution of  $P$ -values (Figure 3-8) ( $\lambda_{GC} = 1.04$ , standard error =  $5.73 \times 10^{-5}$ ). In total, 191 variants from 11 distinct loci were associated below the marginal significance threshold ( $P_{META} \leq 1 \times 10^{-5}$ ). The *PNPLA3* locus contained 155 of the top hit variants including a number below the genome-wide significance level ( $P \leq 1 \times 10^{-8}$ ). None of the variants in the remaining 10 loci surpassed the genome-wide significance threshold. Several of the most significant variants were analysed previously, including rs3747207 (*PNPLA3*), rs58489806 (*TM6SF2*), rs626283 (*MBOAT7*) and rs73321291 (intergenic between *DUSP1* and *ERGIC1*). In total six variants met the criteria for validation analysis, that is, they occurred in independent loci, were below the marginal significance threshold and were unique to this analysis. These variants occur in genomic regions that have been associated with diverse traits through GWAS including lipid and cholesterol metabolism (Table 3-6).



**Figure 3-8 Quantile-Quantile plot of the meta-analysis  $P$ -values**

This plot shows the observed versus the expected distribution of  $P$ -values on a  $-\log_{10}$  scale. It shows the distribution of all variants tested in the meta-analysis (black points) and with the variants from the *PNPLA3* locus removed (blue points). The red line shows the expected distribution of  $P$ -values when there is no systematic deviation from the expected distribution of  $P$ -values with a 95% confidence interval (grey box). The estimated genomic inflation factor,  $\lambda_{GC}$  was derived from all of the  $P$ -values in the data



**Figure 3-9 A Manhattan plot of meta-analysis P-values**

The UCL/STOPAH and German GWAS genetic association datasets showing each the results for each variant by position along the chromosome (x-axis) versus the association P-value on a  $-\log_{10}$  scale. The blue line denotes the marginal significance threshold ( $P \leq 1 \times 10^{-5}$ ) and the red line a genome wide significant threshold ( $P \leq 5 \times 10^{-8}$ )

Table 3-5 The most significant associations from a fixed effects meta-analysis in the extended GWAS analysis

Chromosome	Variant	Position	Major allele	Minor allele	$P_{META}$	Odds Ratio	95% Confidence Interval	Heterogeneity $\frac{P}{I^2}$	$P_{UCL/STOPAH}$	$P_{GERMANY}$
2	rs4027132	12037492	A	G	$2.23 \times 10^{-6}$	1.38	[1.21-1.58]	0.94 0.00	$1.05 \times 10^{-3}$	$5.90 \times 10^{-4}$
3	rs6768997	29449644	T	C	$7.00 \times 10^{-6}$	1.40	[1.21-1.63]	0.52 0.00	$4.61 \times 10^{-4}$	$3.14 \times 10^{-3}$
3	rs28350	42418446	G	A	$2.20 \times 10^{-6}$	1.53	[1.28-1.83]	0.31 3.93	$8.81 \times 10^{-5}$	$2.70 \times 10^{-3}$
5	rs153409	16790483	C	T	$5.25 \times 10^{-6}$	0.73	[0.64-0.84]	0.80 0.00	$9.17 \times 10^{-4}$	$1.63 \times 10^{-3}$
*5	rs73321291	172225349	A	G	$1.55 \times 10^{-6}$	2.11	[1.56-2.87]	0.90 0.00	$2.61 \times 10^{-3}$	$1.09 \times 10^{-4}$
8	rs7812374	18979088	C	T	$8.77 \times 10^{-6}$	1.37	[1.19-1.57]	0.11 60.18	$7.20 \times 10^{-2}$	$1.27 \times 10^{-5}$
18	rs11082773	47221491	A	G	$4.11 \times 10^{-7}$	0.64	[0.54-0.76]	0.50 0.00	$3.51 \times 10^{-5}$	$2.78 \times 10^{-3}$
*19	rs626283	54677001	A	G	$9.71 \times 10^{-7}$	0.56	[0.44-0.7]	0.54 0.00	$3.12 \times 10^{-3}$	$5.82 \times 10^{-5}$
*19	rs58489806	19456917	C	T	$3.43 \times 10^{-6}$	1.36	[1.19-1.54]	0.56 0.00	$5.27 \times 10^{-3}$	$1.65 \times 10^{-4}$
20	rs36056997	49664218	G	GA	$2.65 \times 10^{-6}$	1.40	[1.22-1.61]	0.29 9.15	$9.82 \times 10^{-5}$	$3.95 \times 10^{-3}$
*22	rs3747207	44324855	G	A	$1.46 \times 10^{-30}$	2.41	[2.07-2.80]	0.14 55.20	$2.28 \times 10^{-10}$	$5.61 \times 10^{-23}$

\*These loci were identified in previous analyses

Table 3-6 Positions and features of the novel loci identified during the discovery meta-analysis

Chromosome	Top Variant	Locus Start	Locus End	GWAS Associated Traits ( $\pm 100\text{kB}$ of Locus)	Nearby Genes ( $\pm 100\text{kB}$ )
2	rs4027132	12031341	12048788	Hair morphology, Bipolar disorder	<i>LPIN1</i> , <i>MIR4262</i> ,
3	rs6768997	29433727	29474961	Major depressive disorder, Obesity-related traits	<b><i>RBMS3</i></b>
3	rs28350	42417887	42498693	Response to antipsychotic	<i>LYZL4</i> , <i>VIPR1</i> , <i>SULT4A1</i>
5	rs153409	16788851	16819642	NA	<b><i>MYO10</i></b>
8	rs7812374	18891534	19006584	Urinary albumin excretion, Smooth-surface caries, Telomere length	<i>PSD3</i> , <i>SH2D4A</i>
18	rs11082773	47182737	47405530	Testicular tumour, Blood metabolite levels, HDL cholesterol (8 GWAS), Heschl's gyrus morphology, Lipid metabolism phenotypes (2 GWAS), Obesity-related traits	<i>LIPG</i> , <i>ACAA2</i> , <i>SCARNA17</i> , <i>MYO5B</i> ,
20	rs36056997	49663480	49663480	Molar-incisor hypomineralization, heart cycle re-polarisation	<i>DPM1</i> , <i>MOCS3</i> , <i>KCNG1</i>

Abbreviations: Chr – chromosome; kB – kilobase; HDL – high density lipoprotein

\*Genes in bold typeface represent those where the primary variant is located within the highlighted gene  
Genome-wide association study and gene position data was obtained from the UCSC genome browser<sup>216</sup>

### 3.5.3 - VALIDATION GENOTYPING

A total of 873 UCL controls and 528 STOPAH cases from the UCL/STOPAH validation group underwent direct genotyping, concordance analysis and genetic association analysis for those variants identified in the discovery meta-analysis.

#### Genotyping

All of the selected variants except for rs6768997 were suitable for validation genotyping. Rs6768997 occurs on a repetitive short interspersed nuclear element and is therefore unsuitable for PCR based genotyping so instead, the nearby variant rs9861686 was selected as a proxy. These variants underwent KASPAR genotyping using allele specific primer PCR amplification (Table 3-7). Visual inspection of genotyping cluster plots demonstrates distinct clustering of the three possible genotypes for each variant and the negative controls (Figure 3-10). All of the variants were in Hardy-Weinberg equilibrium in the entire validation group and in both case control groups ( $P_{HWE} > 0.01$ ).

#### Concordance analysis

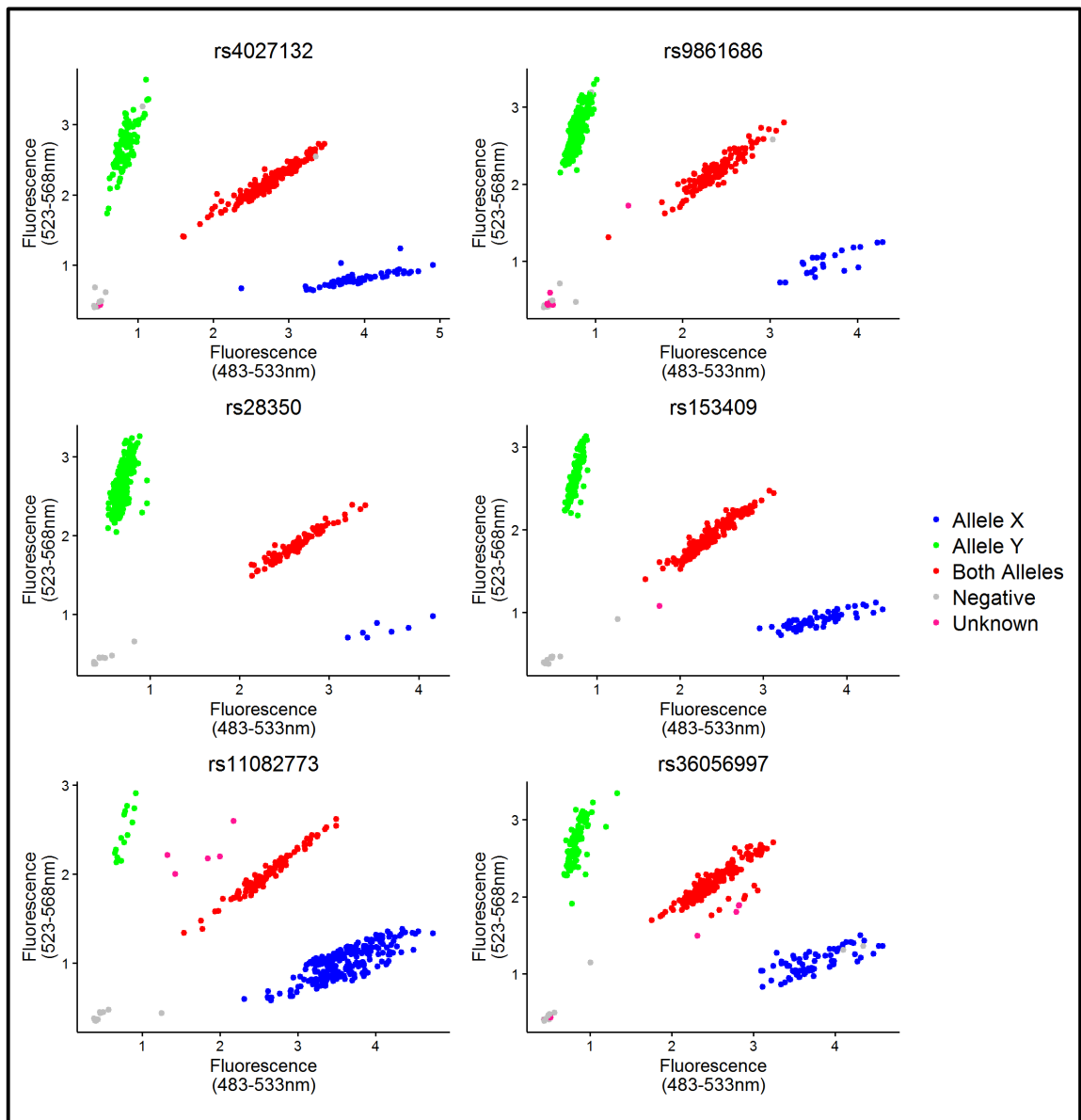
The samples with both imputed and direct genotype data underwent analysis to determine their concordance (Table 3-8; Figure 3-12). The genotypes for the majority of variants were highly concordant between the imputed and direct genotype data (concordance > 97%), however, the variants rs28350 and rs36056997 had a lower concordance. There was a clear linear relationship ( $R^2 = 0.99$ ) between the Impute2 information score of the imputed genotype data and the measured concordance rate between the imputed and genotyped data (Figure 3-11).

#### Genetic association analysis

In the validation group, for the several variants genotyped none of the allele frequencies differ between cases and controls (Table 3-9). When this validation data is meta-analysed with the discovery UCL/STOPAH and German genetic association data, it introduces significant heterogeneity ( $P_{HET} \leq 0.05$ ). Despite none of these genetic associations being significant, all variants share the same directionality of effect as in both the extended UCL/STOPAH and German genetic association datasets.

Table 3-7 Nucleotide sequences of the primers used for validation genotyping of the six most significantly associated variants

SNP	Name	Sequence
rs4027132	Allele specific (G)	GAAGGTCGGAGTCAACGGATTCTCTGGCTATGACAAATTCAAAGTTCAAA
	Allele specific (GA)	GAAGGTGACCAAGTTCATGCTGGCTATGACAAATTCAAAGTTCAAG
	Reverse 1	CTTCTCAAGTCCTCCTAAATAAATGTTTAA
	Reverse 2	CACTCTTTCTTCTCAAGTCCTCCTAAATA
rs9861686	Allele specific (A)	GAAGGTGACCAAGTTCATGCTAAAATGTCATTTAATATCTAAGCCTCTGAATA
	Allele specific (G)	GAAGGTCGGAGTCAACGGATTAATGTCATTTAATATCTAAGCCTCTGAATG
	Reverse 1	GCATAATCACACATTGCTCTGTGCTATTT
	Reverse 2	CATTGCTCTGTGCTATTTACAAATAGGTTT
rs28350	Allele specific (A)	GAAGGTGACCAAGTTCATGCTGTTGCCTAGACAATGGCATATTAAATCA
	Allele specific (G)	GAAGGTCGGAGTCAACGGATTGCCTAGACAATGGCATATTAAATCG
	Reverse 1	CTGGAAGATTATTACATTCCTTGTCTTA
	Reverse 2	CATTCCCTTGTCTATAGATCAAGGACAA
rs153409	Allele specific (C)	GAAGGTCGGAGTCAACGGATTCATATCATGTGGTCTTTGCAGTACG
	Allele specific (T)	GAAGGTGACCAAGTTCATGCTCCATATCATGTGGTCTTTGCAGTACA
	Reverse 1	AATTAGCCACAGGCCTTGTCTGTATTATA
	Reverse 2	GGCCTTGTCTGTATTATACACCTTAAGAA
rs11082773	Allele specific (A)	GAAGGTGACCAAGTTCATGCTAAGATTTCAAACCTCCTAACAGCCAGAT
	Allele specific (G)	GAAGGTCGGAGTCAACGGATTTCAAACCTCCTAACAGCCAGAC
	Reverse 1	CCAGGCTTGGTGAAAAAGAAATGAATGAA
	Reverse 2	GGTAGTTAGTCCAGGCTTGGTGAAA
rs36056997	Allele specific (C)	GAAGGTCGGAGTCAACGGATTGTAAGTACCTGCTCTTATACAAGGAC
	Allele specific (A)	GAAGGTGACCAAGTTCATGCTGTAAGTACCTGCTCTTATACAAGGAA
	Reverse 1	TCCAGTTCTGAGAAGTCCTCTAAGATTT
	Reverse 2	GAGAAGTCCTCTAAGATTTCCAGTTGATT



**Figure 3-10 Fluorescence cluster plots of validation genotyping experiments**

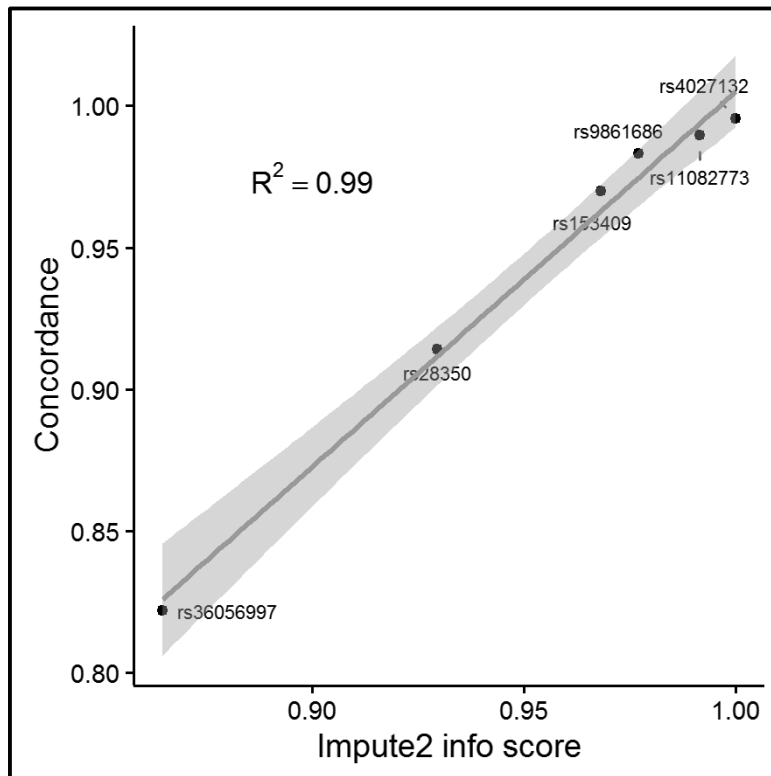
Each scatter plot represents the recorded fluorescence at two different wavelengths for a single KASPAR genotyping experiment containing 384 DNA samples. Each point represents a single genotype coloured by clustering (red, green and blue) from which genotyping is inferred. Negative controls (grey) should have limited fluorescence and cluster together towards zero. Samples with poor clustering were classified as unknown (pink).



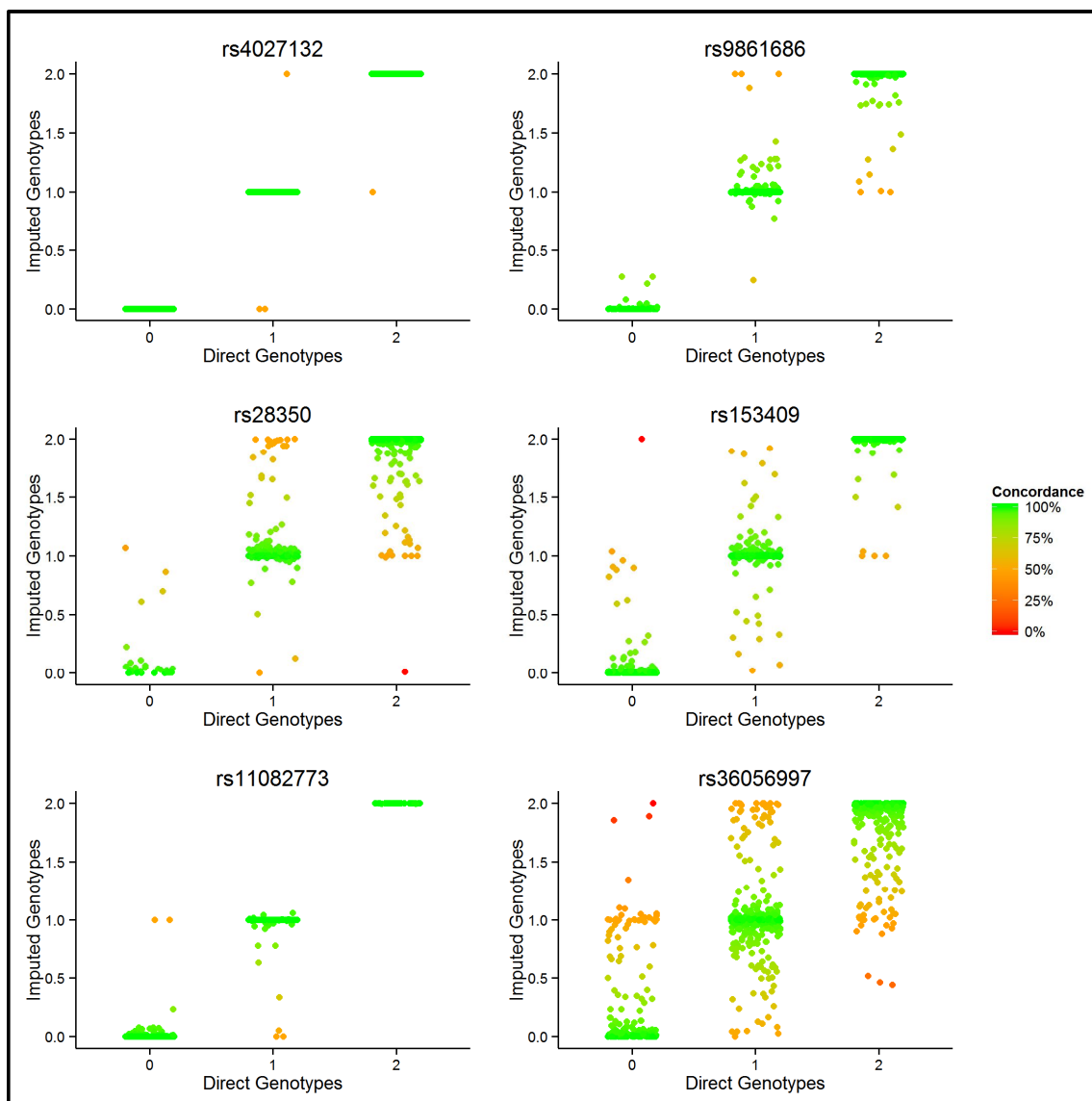
Table 3-8 A concordance comparison of direct and imputed genotype data

Chromosome	Variant	Number of genotypes compared	Concordance*
2	rs4027132	917	1.00
3	rs9861686	922	0.98
3	rs28350	919	0.91
5	rs153409	890	0.97
18	rs11082773	910	0.99
20	rs36056997	918	0.82

\*Pearson's correlation coefficient



**Figure 3-11 Linear regression plot of genotyping concordance versus Impute2 info score**  
 From a linear regression of Impute2 info scorer versus genotype concordance a line of best fit (blue line) with its 95% confidence interval (grey polygon)



**Figure 3-12 Comparing the concordance between imputed and direct genotype data**

Each plot shows the discrete direct genotype values converted into genotype dosage format (0=homozygote allele 1, 1=heterozygote, 2= homozygote allele 2) (x-axis) plotted versus the continuous imputed genotype dosage values (y-axis) for the six variants that underwent validation genotyping. Each genotype data point is coloured on a continuous scale (from red to green) by its percentage concordance between the imputed and direct genotype data

Table 3-9 The results of a genetic association analysis in the validation group

Chromosome	Variant	Minor Allele	Major Allele	Genotype Counts		Minor Allele Frequency		P*	Odds Ratio	95% Confidence Interval
				Cases	Controls	Cases	Controls			
2	rs4027132	G	A	105/259/147	169/383/282	45.9%	43.2%	0.18	1.11	[0.95-1.30]
3	rs9861686	A	G	40/196/277	67/317/453	26.9%	26.9%	0.98	1.00	[0.84-1.19]
3	rs28350	A	G	15/136/363	17/230/599	16.2%	15.6%	0.71	1.04	[0.84-1.29]
5	rs153409	T	C	78/272/173	151/402/295	40.9%	41.5%	0.76	0.98	[0.83-1.14]
18	rs11082773	G	A	13/134/363	29/220/580	15.7%	16.8%	0.46	0.92	[0.75-1.14]
20	rs36056997	GA	G	103/253/162	164/404/279	44.3%	43.2%	0.58	1.05	[0.89-1.22]

\*Allelic association test performed using logistic regression when conditioning on gender as a covariate

Table 3-10 The results of a meta-analysis including the validation genetic association data

Chromosome	Variant	Effects Model	$P_{META}$	Heterogeneity		$P_{GERMANY}$	$P_{UCL/STOPAH}$	$P_{VALIDATION}$
				$P$	$I^2$			
2	rs4027132	Fixed	$1.40 \times 10^{-5}$	0.09	58.38	$5.90 \times 10^{-4}$	$1.05 \times 10^{-3}$	0.26
		Random	$3.45 \times 10^{-3}$					
3	rs9861686	Fixed	$2.18 \times 10^{-3}$	0.00	81.53	$5.09 \times 10^{-3}$	$3.93 \times 10^{-4}$	0.63
		Random	0.13					
3	rs28350	Fixed	$3.84 \times 10^{-5}$	0.03	71.30	$2.70 \times 10^{-3}$	$8.81 \times 10^{-5}$	0.49
		Random	0.019					
5	rs153409	Fixed	$2.48 \times 10^{-4}$	0.02	73.15	$1.63 \times 10^{-3}$	$9.17 \times 10^{-4}$	0.79
		Random	0.040					
18	rs11082773	Fixed	$1.07 \times 10^{-5}$	0.03	72.32	$2.78 \times 10^{-3}$	$3.51 \times 10^{-5}$	0.48
		Random	0.016					
20	rs36056997	Fixed	$7.57 \times 10^{-5}$	0.02	74.86	$3.95 \times 10^{-3}$	$9.82 \times 10^{-5}$	0.50
		Random	0.027					

### 3.6 - DISCUSSION

The merging of datasets genotyped on different panels has the potential to confound GWAS<sup>484</sup>. These analyses utilised strict quality control procedures and state of the art software<sup>89</sup> to remove errors resulting from dataset harmonisation and merging. The quality control also included an analysis using the overlapping control samples genotyped on two arrays for the identification of discordant genotypes. In the final UCL/STOPAH dataset, genotype concordance was high, and there was variant coverage over all of the autosomes. Principal components analysis demonstrates that there are no significant differences between samples in the merged dataset.

The *PNPLA3* locus contained both the greatest number of significant genetic associations and the most significant overall genetic associations. The genetic association at the variant rs738409 was 100-fold more significant in the merged dataset in comparison to the UCL dataset alone ( $P_{\text{UCL}} = 2.51 \times 10^{-8}$  versus  $P_{\text{UCL/STOPAH}} = 1.68 \times 10^{-30}$ ), a finding consistent with the increased number of cases. Other variants identified from the discovery analysis were nearing genome-wide levels of significance such as the candidate gene *DLG*. Other GWAS have also associated the *DLG2* locus with energy intake in childhood obesity<sup>69</sup> and the sera levels of certain glycerophospholipid species<sup>94</sup>.

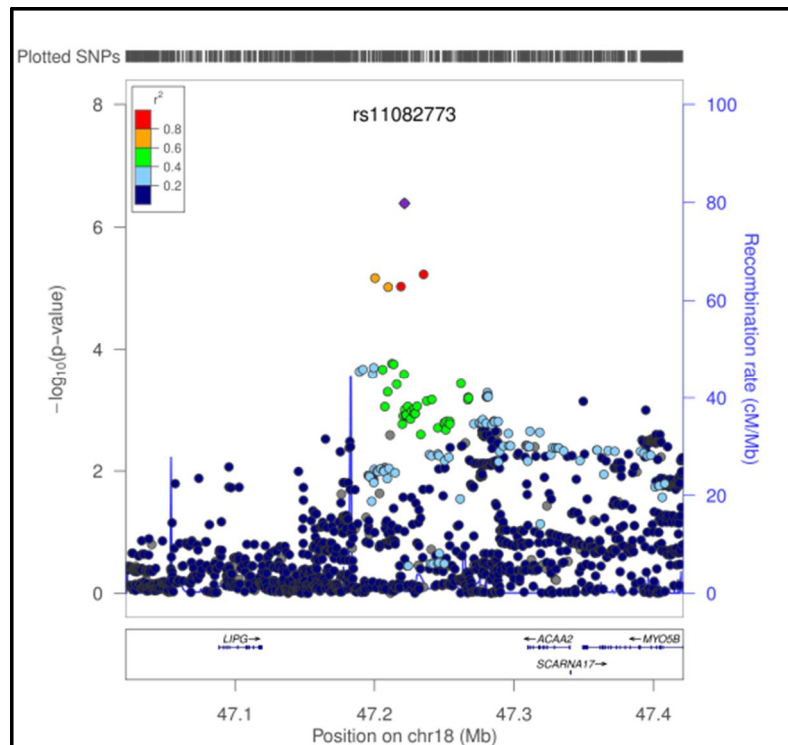
This extended GWAS confirms the primacy of *PNPLA3* as a risk locus as it was the only locus to contain genome-wide significant associations in both the expanded UCL/STOPAH GWAS analysis and the German association analysis. Notably, rs3747207 was the most significantly associated variant in *PNPLA3*. This variant, which occurs in an intergenic region between the second and third exons, had a marginally greater association signal than rs738409 which has been identified as the primary variant driving associations at this locus<sup>370,413,428</sup> ( $P_{\text{rs3747207}} = 1.46 \times 10^{-30}$ ,  $P_{\text{rs738409}} = 1.68 \times 10^{-30}$ ). These two variants are close to each other (130 bases) and in perfect linkage disequilibrium ( $r^2 = 0.99$ ,  $D' = 1.0^1$ ). This slight discrepancy may have resulted from the lower imputation information score of rs374207 in both the German dataset (Impute2 info score: rs738409 = 0.943 vs rs3747207 = 0.934) and the UCL/STOPAH dataset (Impute2 info score: rs738409 = 0.951 vs rs3747207 = 0.946).

The meta-analysis identified several novel loci containing variants nearing genome-wide significance. Of these, there was an intriguing novel loci containing several highly associated variants (rs11082773,  $P_{\text{META}} = 4.11 \times 10^{-7}$ ) between the endothelial lipase gene, *LIPG* and the acetyl-Coenzyme A acyltransferase 2 gene, *ACAA2* (Figure 3-13). This region is a quantitative trait loci for the expression high-density lipoprotein (HDL) cholesterol<sup>16,60,450,457</sup> and the nearby endothelial lipase gene may be functionally involved in this mechanism<sup>196</sup>. The *ACAA2* gene encodes the enzyme acetyl-

Coenzyme A acyltransferase 2 which catalyses the last stage of the mitochondrial fatty acid beta oxidation spiral<sup>177</sup>. Other validated alcohol-related cirrhosis risk loci such as *PNPLA3*, *MBOAT7* and *TM6SF2* are also involved in lipid metabolic pathways making this locus an intriguing candidate for further analysis.

For the most significantly associated variants, the imputed and direct genotypes were compared in a large number of samples. The concordance between the imputed and direct genotypes was high for most variants (concordance > 97%) highlighting the accuracy of imputation. There was a near perfect correlation between the imputation information measure and the concordance between direct and imputed genotype data. The imputation information measure is used by the frequentist association test performed in SNPTEST, to account for missing information in imputed data in genetic association calculations<sup>277</sup>. Hence, even when the imputation information score was lower, as was the case for rs36056997, this was unlikely to introduce bias to the genetic association test.

Validation genotyping did not provide supporting evidence for the novel genetic associations identified in the extended GWAS meta-analysis. Independent replication is the gold standard<sup>59</sup> for validating a genetic association, although non-replication alternately does not refute a genetic association per se<sup>150</sup>. There are several potential causes of non-replication such as complex genetic architecture<sup>153</sup>, insufficient statistical power<sup>150</sup>, or phenotypic differences between datasets<sup>230</sup>. There is evidence to suppose that phenotype definitions in the replication cohort differed from those used in the primary GWAS datasets as the genotyping significant heterogeneity into the meta-analysis summary results<sup>119</sup>. There are phenotypic differences between the STOPAH cohort cases and the UCL alcohol-related cirrhosis cases. First, a large number of the severe alcoholic cases were of unidentified alcohol-related cirrhosis status whereas all of the UCL cases had confirmed alcohol-related cirrhosis. Second, there was less ancestral information for samples in the STOPAH cohort, bar self-reported white British status, giving a greater potential for population stratification.



**Figure 3-13 A locus plot of the genetic association signal near *LIPG* and *ACAA2***

The  $-\log_{10}(P \text{ values})$  are plotted against SNP genomic position based on the human genome build 19. Variants are coloured to reflect correlation with the most significant SNP, with red denoting the highest linkage disequilibrium, ( $r^2 > 0.8$ ) with the lead SNP. Estimated recombination rates (blue line) from the 1000 Genomes Project (hg19/genomes March 2012 release, EUR population) reflect the local linkage disequilibrium structure. Gene annotations were from the UCSC Genome Browser. The plot was generated using LocusZoom<sup>347</sup>.

Many of the same criticisms levied at the original GWAS pertain here also (e.g. merging of different datasets, lack of coverage of CNVs and rare variants). Another limitation of this analysis is the non-autosomal coverage of the sex chromosomes or the mitochondrial DNA. These genomic regions were not included in this analysis due to the increased challenges of harmonizing, merging and imputing haploid variants. There are well-known differences in alcohol-related liver disease risk by gender<sup>295</sup>, and therefore the sex chromosomes may contain important loci that contribute to disease risk. The mitochondria are also linked to liver disease because they are intimately involved in the oxidative stress mechanisms which contribute to the pathogenesis of alcohol-related liver disease<sup>182</sup>. Therefore, mitochondrial DNA may also contain important alcohol-related liver disease loci. Despite these limitations, there is high confidence in the data quality of this enlarged meta-analysis because it still identifies the confirmed loci: *PNPLA3*, *TM6SF2* and *MBOAT7* as top-hits. The functionality of the novel top-hit loci, especially in the region near *LIPG/ACAA2*, provide intriguing evidence that genes involved in lipid metabolic pathways are important in the pathogenesis of alcohol-related cirrhosis and the broader alcohol-related liver disease phenotype.

In summary, this extended GWAS revealed several novel variants nearing genome-wide significance levels while variants in *PNPLA3*, *TM6SF2* and *MBOAT7* also remained as the most significant associations. Despite showing the same directionality in all instances, none of these variants replicated. One of the loci identified via this analysis was between the genes *LIPG* and *ACAA2*, is also a validated cholesterol level quantitative trait locus<sup>11,52,410,416</sup> and hence shares a similar pathway to the validated risk loci *TM6SF2*. Genetic variation in *PNPLA3* and in particular, the variant rs738409, remained as the most significant association in this dataset.



---

---

**CHAPTER 4 GENETIC VARIATION IN  
*PNPLA3* AND ALCOHOL-RELATED  
CIRRHOSIS RISK**

---

---

## 4.1 - OVERVIEW

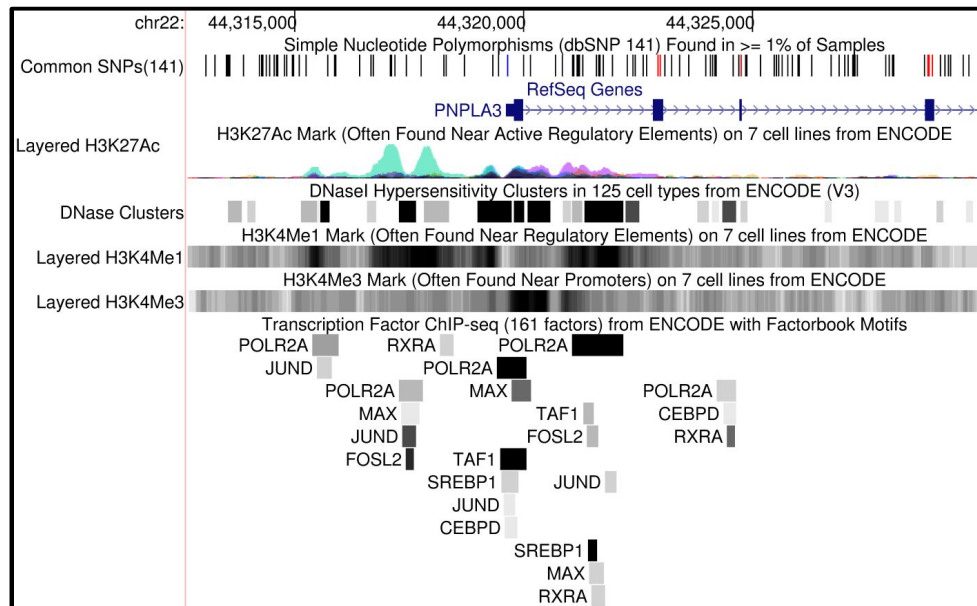
The variant rs738409 in *PNPLA3* has the largest effect on alcohol-related cirrhosis risk of any variant so far identified. While the robust genetic associations at this locus have been addressed, much still remains to be clarified. In this chapter the consequences of the carriage of rs738409 are further explored by determining: (i) if it is a risk factor for the development of alcohol dependence per se; (ii) if it is associated with other important phenotypic features of alcohol-related cirrhosis such as age at presentation, survival and the development of HCC; and (iii) its contribution to the population attributable risk of cirrhosis in the British and Irish ancestry population of England.

## 4.2 - BACKGROUND

### 4.2.1 - *PNPLA3*

*PNPLA3* occurs on the long arm of chromosome 22 towards the end of the 22q13.2 region. The reference sequence for *PNPLA3* contains nine exons and spans nearly 24 kilo-bases. It has five known transcripts and three of these do not code for proteins; the transcript ENST00000216180 encodes the reference protein sequence<sup>126</sup> (Supplementary Sequence 1). The transcription of *PNPLA3* is detected primarily in the liver but also in kidney, adipose, cerebral, duodenal and skin tissue<sup>121</sup>.

A review of transcription factor binding motifs in the *PNPLA3* promoter region from the ENCODE dataset (Figure 4-1)<sup>116</sup> demonstrates several nutritionally regulated transcription factor binding motifs and others which recruit the core machinery for transcription such as SREBP and retinoid X receptor binding motifs. As with any transcribed gene, it also contains all of the sequence necessary to recruit the core transcription machinery such as RNA polymerase II (POLR2A) and myc-associated factor X (MAX). Several of these transcription factors are known to regulate *PNPLA3* expression via a feed-forward loop<sup>188</sup>.

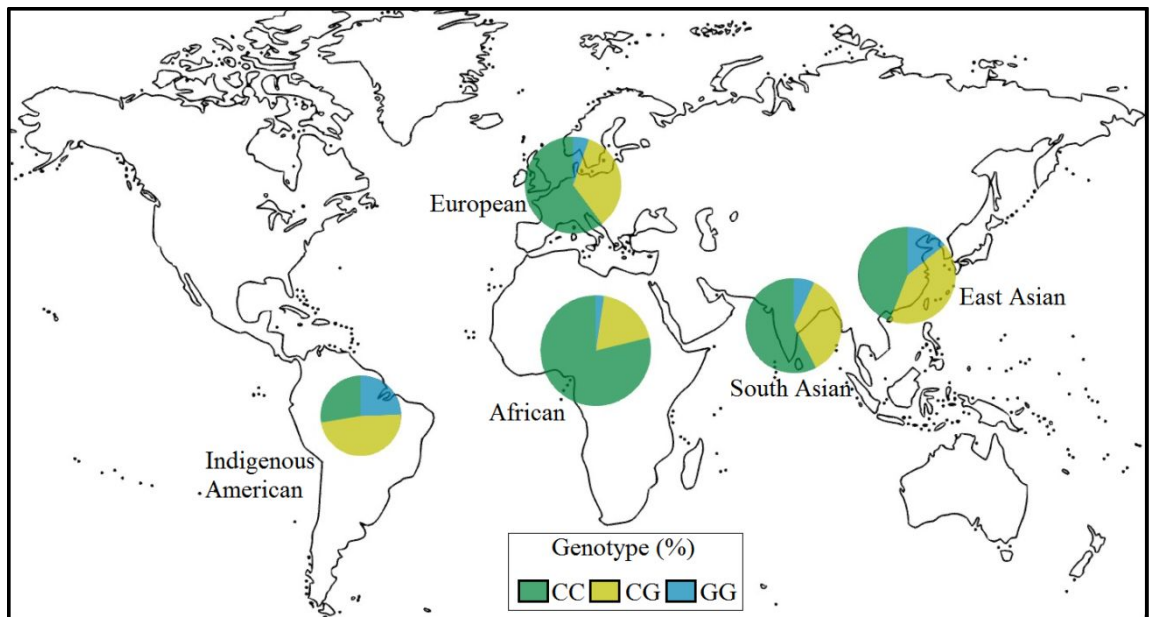


**Figure 4-1 Transcription factor binding sites and regulatory elements in *PNPLA3***

Several transcription factor binding sites that are involved in nutritional regulation are found in the promoter region of *PNPLA3*. Abbreviations: chr – chromosome; SNP – single nucleotide polymorphism; ENCODE - The Encyclopaedia of DNA Elements; POLR2A – RNA polymerase II; JUND - jun D proto-oncogene; MAX - myc-associated factor X; RXRA - Retinoic acid receptor RXR-alpha; FOSL2 - FOS-Like Antigen 2; CEBPD - CCAAT/Enhancer-Binding Protein Delta; SREBP1 - Sterol regulatory element-binding transcription factor 1. Image obtained from the UCSC genome browser<sup>216</sup>

#### 4.2.2 - GENETIC VARIATION IN *PNPLA3* AND LIVER DISEASE

*PNPLA3* was first identified as a nutritionally regulated mRNA transcript in adipose tissue<sup>22</sup> and following from this the earliest genetic association studies of *PNPLA3* were undertaken in obesity related phenotypes. In the first candidate gene study of *PNPLA3*, several SNPs were found to marginally associate with obesity risk<sup>201</sup>. Interest in the role of genetic variation in *PNPLA3* escalated when a GWAS in patients with NAFLD identified highly significant association between the variant rs738409 in *PNPLA3* and liver fat content and measures of liver inflammation<sup>370</sup>. This variant, rs738409, encodes a cytosine to guanine substitution on the forward strand of chromosome 22 at the position 43,928,847 on human genome build 19 (hg19)<sup>1</sup> and is common with the minor allele (rs738409[G]) having frequencies varying from 10-50% in global populations (Figure 4-2).



**Figure 4-2 Map of rs738409 genotype frequencies**

The variant rs738409 is common in all global populations with minor allele frequencies ranging from as low as ~10% to as high as ~50%. Data source: 1000 genomes, 2012<sup>1</sup>.

Following from the initial GWAS, several independent studies have confirmed significant associations between rs738409 and NAFLD as well as several related phenotypes and traits (Table 4-1). In the largest GWAS of NAFLD to date, rs738409 conferred the largest effect on the chance of developing advanced stage NAFLD ( $P=3.60 \times 10^{-43}$ , OR 3.26 95% CI [2.11–7.21])<sup>405</sup>. There is an increasing awareness of the effect of rs738409 as an independent risk factor in the development of HCC in NAFLD<sup>169,398</sup> and its potential use in prognostic models<sup>161</sup>.

Table 4-1 Genome-wide association studies where *PNPLA3* has been identified as a significant locus

Reference	Disease/Trait	Discovery Sample (n)	Replication Sample (n)	Reported Gene(s)	Most significant variant [Risk Allele]	P-Value
DiStefano, 2015 <sup>100</sup>	Hepatic lipid content in extreme obesity	Caucasian (2,300)	-	<i>PNPLA3</i>	rs4823173[G]	2.8x10 <sup>-7</sup>
Kitamoto, 2013 <sup>225</sup>	NAFLD	Japanese Cases (392) Controls (934)	Japanese cases (172), controls (1012)	<i>PNPLA3</i> , <i>SAMM50</i> , <i>PARVB</i>	rs2896019[G]	2x10 <sup>-20</sup>
Kawaguchi, 2012 <sup>217</sup>	NAFLD	Japanese cases (529), Controls (932)	-	<i>PNPLA3</i>	rs738409[G]	1x10 <sup>-10</sup>
Chambers, 2011 <sup>53</sup>	Alanine transaminase levels	European (52,350) Asian Indian (8,739)	-	<i>PNPLA3</i> , <i>SAMM50</i>	rs738409[G]	1x10 <sup>-45</sup>
Kim, 2011 <sup>224</sup>	Liver enzyme levels	Korean (12,545)	East Asian (30,395)	<i>PNPLA3</i>	rs12483959[A]	2x10 <sup>-18</sup>
Speliotes, 2011 <sup>405</sup>	NAFLD	Amish (880), European (6,296)	European cases (592) controls (1,405)	<i>PNPLA3</i>	rs738409[G]	4x10 <sup>-34</sup>
Paré, 2011 <sup>323</sup>	Soluble ICAM-1	Multi-ethnicity American women (22,435)	Predominantly European (9,813)	<i>PNPLA3</i>	rs738409[G]	5.8x10 <sup>-8</sup>
Kamatani, 2010 <sup>211</sup>	Alanine transaminase levels	Japanese (14,402)	-	<i>PNPLA3</i> , <i>SAMM50</i> , <i>PARVB</i>	rs2896019[G]	2x10 <sup>-12</sup>
Yuan, 2008 <sup>475</sup>	Alanine transaminase levels	European (7,751)	European (1,005) Asian Indian (3,669)	<i>PNPLA3</i> , <i>SAMM50</i>	rs2281135[T]	8x10 <sup>-16</sup>
Romeo, 2008 <sup>370</sup>	NAFLD, fat content & inflammation	Hispanic, African & European American individuals (9,229)	-	<i>PNPLA3</i>	rs738409[G]	5.9x10 <sup>-10</sup>

\*Data contains imputed genotypes.

Abbreviations: NAFLD – Non-alcohol-related fatty liver disease

### 4.2.3 - *PNPLA3* IN ALCOHOL-RELATED LIVER DISEASE

NAFLD and alcohol-related liver disease share clinical, histological and pathogenic features<sup>266,473</sup>. Following from this, several groups have genotyped rs738409 and other variants in *PNPLA3* to investigate genetic associations with alcohol-related liver disease phenotypes (Table 4-3).

The seminal genetic association between rs738409 and alcohol-related cirrhosis was reported in a mixed European/Native American Mestizo population from Mexico in which 15 SNPs around *PNPLA3* were genotyped<sup>428</sup>. On conditional analysis, the most significant association was observed between the rs738409 and the presence of clinically diagnosed alcohol-related cirrhosis. This association has replicated in a modest number of independent cohorts. In the first of these, a highly significant allelic association was reported between rs738409 and alcohol-related cirrhosis in a population from the UK<sup>388</sup>. A further two studies have also demonstrated significant associations between this variant and alcohol-related liver disease risk. The first of these studies reported a strong association rs738409 and biopsy proven alcohol-related cirrhosis in a predominantly male German population<sup>413</sup> while the second study demonstrated association with the broader phenotype of alcohol-related liver injury in a Belgian population<sup>431</sup>. In all studies the minor allele (rs738409[G]) encoding the 148Met allele was found to increase risk of alcohol-related cirrhosis.

It seems likely that the genetic association between rs738409 and alcohol-related cirrhosis are independent of the comorbid phenotype, alcohol dependence<sup>413</sup>. However, the case/control comparisons required to validate this hypothesis have not been tested in the majority of genetic association studies<sup>55</sup>. The case/control comparisons have been either between cases and population controls<sup>431</sup> or between cases and alcohol-misusers without significant liver injury<sup>388,428</sup>. As a case in point, such a confounding genetic effect occurs for the genetic association between rs671 (Glu504Lys) *ALDH2* and alcohol-related cirrhosis in East Asian ancestry populations, which is primarily driven due to its association with alcohol-dependence/misuse. It remains unclear whether the alcohol dependence phenotype is confounding genetic association studies of rs738409 and alcohol-related cirrhosis.

Beyond influencing cirrhosis predisposition per se, rs738409 is also associated with HCC following presentation with alcohol-related cirrhosis<sup>432</sup> (Table 4-2). In particular, carriage of the rs738409[G] allele increases HCC risk following the alcohol-related development of cirrhosis (OR = 1.77)<sup>432</sup>.

Table 4-2 The association between rs738409 and HCC risk

Study	Odds Ratio [95% CI]	P
Falletti et al., 2011 <sup>122</sup>	1.64 [0.98-2.73]	0.057
Guyot et al., 2013 <sup>161</sup>	2.23 [1.44-3.45]	3.05x10 <sup>-4</sup>
Hamza et al., 2012 <sup>164</sup>	1.67 [0.96-2.89]	0.067
Nischalke et al., 2011 <sup>309</sup>	2.68 [1.48-4.85]	0.0011
Trépo et al., 2012 <sup>430</sup>	2.51 [1.84-3.41]	4.30 x 10 <sup>-9</sup>
Meta-analysis	2.20 [1.80-2.67]	4.71 x10 <sup>-15</sup>

Abbreviation: CI – confidence interval.  
Data from Trépo et al., 2011<sup>432</sup>

There is emerging evidence that rs738409 is associated with the clinical presentation with alcohol-related cirrhosis and its prognosis. This has been explored in a limited way using the rs738409 genotype to stratify groups and following features of alcohol-related liver disease such as presentation with cirrhosis from alcohol-misuse onset<sup>46</sup>, time on a transplantation waiting list until death or liver transplant<sup>133</sup> and presentation with HCC and survival in alcohol-related cirrhosis<sup>443</sup>. It has been shown that carriers of rs738409[G] cirrhosis risk allele: (i) present with cirrhosis after a shorter drinking history; (ii) have a significantly increased risk of developing hepatocellular carcinoma once they have cirrhosis; and, (iii) either die or develop decompensated cirrhosis after a shorter time-period. There are several potentially confounding issues in these time-to-event analyses, including lack of control of ancestral background<sup>443</sup>, potential violations of the proportional hazards assumption<sup>133,151</sup> and lack of inclusion of key variables in multivariate time to event models<sup>133</sup>.

Table 4-3 Published candidate genetic association studies of *PNPLA3* in alcohol-related liver disease

Reference	Disease/Trait	Sample Size*	Number of SNPs Genotyped	Strongest SNP [Risk Allele]	P-Value	Odds Ratio	95% Confidence Interval
Friedrich et al., 2014 <sup>133</sup>	Alcohol-related liver disease vs. population controls	European (2055)	1	rs738409 [G]	<0.005	-	-
Dutta 2013 <sup>106</sup>	Alcohol-related cirrhosis vs. healthy controls	Indian (220)	10 (1 in <i>PNPLA3</i> )	rs738409 [G]	0.037	2.12	[1.29-3.47]
Burza, 2014 <sup>46</sup>	Alcohol-related cirrhosis (time to clinical presentation)	Italian (384)	1	rs738409 [G]	0.021	**1.53	**[1.07-2.19]
Rausch et al., 2013 <sup>356</sup>	Histological damage in alcohol-related liver disease	German (217)	1	rs738409 [GG]	<0.005	-	-
Stickel, 2011 <sup>413</sup>	Alcohol-related cirrhosis vs. no-liver disease controls	German (838)	1	rs738409 [G]	7.84x10 <sup>-8</sup>	2.62	[1.73-3.97]
Trépo, 2011 <sup>431</sup>	Alcohol-Related liver disease vs. population controls	Belgian and French (658)	1	rs738409 [G]	8x10 <sup>-3</sup>	1.54	[1.12-2.11]
Nischalke et al, 2011 <sup>309</sup>	Alcohol-related cirrhosis vs. population controls	German (271)	1	rs738409 [G]	<0.002	1.91	[1.28-2.86]
Nguyen-Khac et al., 2011	Severe alcoholic hepatitis vs population controls			rs738409 [G]	<0.001	2.79	[1.39-5.64]
Seth, 2010 <sup>388</sup>	Alcohol-related liver disease vs. no-liver disease controls	British (454)	1	rs738409 [G]	2x10 <sup>-5</sup>	2.2	[1.53-3.18]
Tian, 2010 <sup>428</sup>	Clinical alcohol-related cirrhosis vs. population controls	Mexican Mestizo (1221)	291 (15 in <i>PNPLA3</i> )	rs738409 [G]	4.7x10 <sup>-5</sup>	1.81	[1.36-2.41]

\*Only includes alcohol-related liver disease population and control population from which statistical comparisons were made  
 \*\*Hazard ratio



## **4.3 - AIMS**

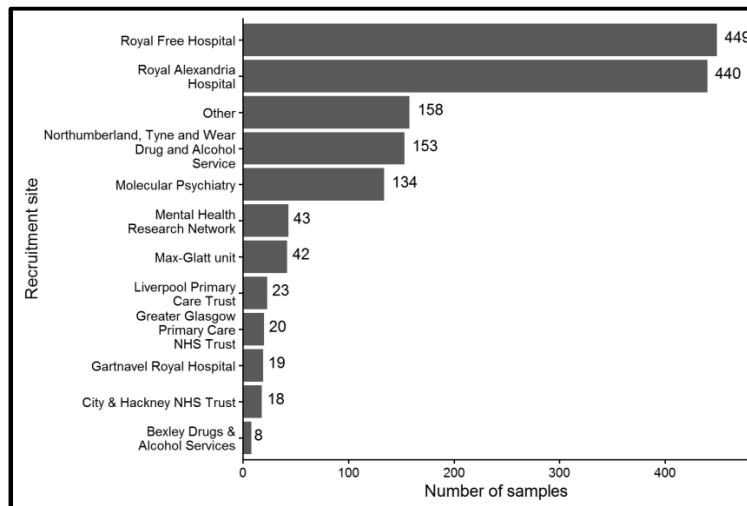
There were three aims:

1. To investigate the association between the rs738409 and alcohol-related cirrhosis when comparing both alcohol-dependent and non-dependent populations controls via direct genotyping
2. To analyse the association between rs738409 and time to presentation with cirrhosis and time to death using time to event modelling
3. To determine the population attributable risk for alcohol-related cirrhosis resulting from carriage of rs738409 in the British population

## **4.4 - MATERIALS AND METHODS**

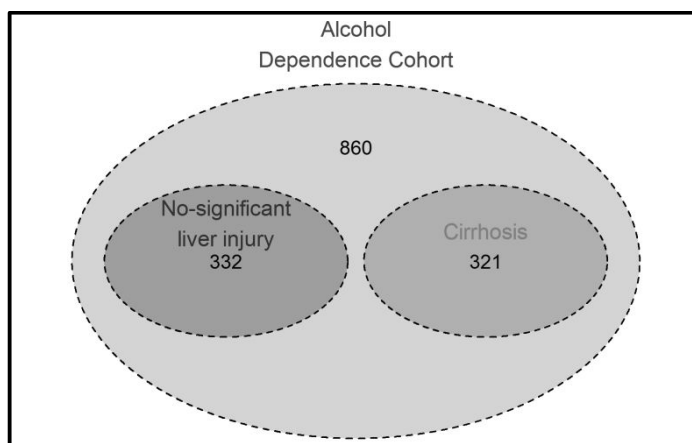
### **4.4.1 - ALCOHOL-DEPENDENCE COHORT**

The UCL alcohol-dependence cohort comprised of the DNA samples from 1513 individuals that were recruited when attending a variety of UK community and hospital-based services providing support and treatment for alcohol use disorders, between 1997 and 2013 (Figure 4-3). A diagnosis of alcohol-dependence was made, by experienced staff/trained research workers, using the DSM-IV<sup>10</sup> or the ICD-10<sup>463</sup> criteria. All participants were unrelated and of English, Scottish, Welsh or Irish descent with a maximum of one grandparent of Northern-European but non-Jewish origin. Details of alcohol history including the ages of onset of alcohol misuse, abuse and dependence; the duration of hazardous drinking; average daily alcohol consumed in grams and family history of alcohol abuse/dependence were recorded in the majority of samples in this cohort.



**Figure 4-3 The sites at which alcohol-dependent subjects were recruited**

A subset of 652 alcohol dependent individuals, recruited primarily from the Hepatology service at the Royal Free Hospital, London were more extensively evaluated to determine the presence and degree of any associated alcohol-related liver injury. The clinical history, including further details of alcohol consumption, were recorded with particular attention paid to previous encounters with medical practitioners, laboratory and radiological investigations and hospitalizations and clinical episodes suggestive of liver injury or decompensation including: jaundice, variceal haemorrhage, fluid retention and neuropsychiatric disturbance. All patients were examined by two experienced, senior clinicians for signs of alcohol misuse and liver injury and a detailed nutritional assessment was undertaken, in the majority, by a senior dietitian. Blood was taken at the time of presentation for standard liver function tests including plasma albumin, serum bilirubin, ALP, AST, ALT, and GGT, full blood count, prothrombin time and the INR; urea, creatinine and electrolytes; serological testing for antibodies to hepatitis A, B, C, D and E, cytomegalovirus, Epstein–Barr virus, herpes simplex and varicella; mitochondrial, nuclear, smooth muscle and liver kidney autoantibodies; iron, total iron binding capacity, ferritin; copper, caeruloplasmin; alpha-one antitrypsin and tissue transglutaminase. All patients underwent abdominal ultrasound and/or abdominal computed tomography/magnetic resonance imaging scans, as indicated; all underwent routine upper gastrointestinal endoscopy; histological examination was undertaken, whenever possible, of liver biopsy material obtained by percutaneous, ultrasound guided or transjugular routes; or else of explant or post-mortem liver tissue. Patients were excluded if they had any other potential cause of liver injury e.g. chronic viral hepatitis, autoimmune liver disease; genetic haemochromatosis; Wilson’s disease; alpha-one antitrypsin deficiency or coeliac disease or if they had a body mass index (BMI) > 30 and were also excluded from this group if they had a diagnosis of diabetes. Based on this assessment two groups were defined from the larger cohort (Figure 4-4):



**Figure 4-4 Venn diagram of the UCL alcohol-dependence cohort samples stratified by liver disease status**

### Alcohol-related cirrhosis

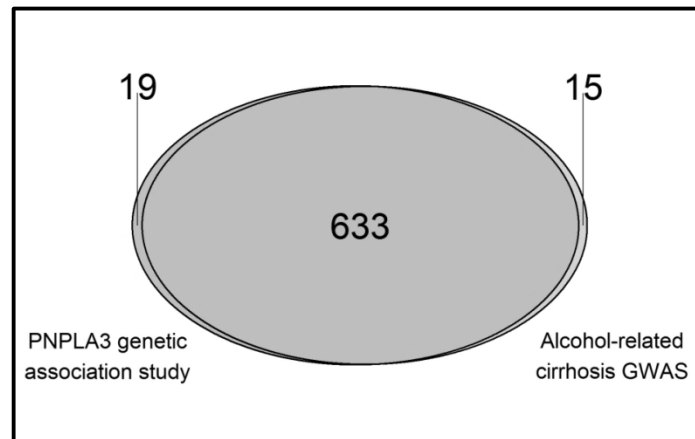
The alcohol-related cirrhosis group ( $n=321$ ) comprised of individuals with either clinically evident or biopsy proven alcohol-related cirrhosis. For these participants, data regarding length of drinking history and self-reported daily alcohol consumption were available from the date of recruitment. In 224 (69.8%) of this group, the diagnosis of cirrhosis was made on the basis of a sustained history of prolonged, harmful drinking; the presence of alcohol-dependence and histological examination of liver tissue; in the remaining 97 (30.2%), the diagnosis was made on the basis of a sustained history of prolonged, hazardous drinking; the presence of alcohol-dependence and compatible historical, clinical, laboratory, radiological and endoscopic features.

### No significant liver injury

This population comprised of 332 individuals - all subjects were actively drinking at the time of enrolment of which data were available regarding length of drinking history and self-reported daily alcohol consumption were available from the date of recruitment. The majority of these individuals had been hazardous drinking for over fifteen years. In 118 (36%) the absence of significant alcohol-related liver injury was confirmed on liver biopsy. The remainder had no historical or clinical features suggestive of significant liver injury either at presentation or during prolonged follow-up; 21 (6.3%) had isolated hyper-bilirubinaemia, most likely reflecting the presence of *Gilbert's syndrome*, but the remainder had normal serum bilirubin levels; all had normal plasma albumin concentrations and clotting profiles; serum GGT activity was raised in the majority; serum ALT activity was within the laboratory reference range in the majority but when elevated rarely exceeded twice the upper laboratory reference range; none had

evidence of parenchymal liver injury or portal hypertension on imaging; upper gastrointestinal endoscopy was normal in those in whom it was performed.

The majority of the individuals in the combined alcohol-related cirrhosis and no-significant liver injury groups ( $n=633$ ) also underwent genome-wide genotyping as part of the first GWAS of alcohol-related cirrhosis (chapter 1)(Figure 4-5).



**Figure 4-5 The overlap in samples between the genetic association study of PNPLA3 and the GWAS of alcohol-related cirrhosis**

## POPULATION CONTROLS

The ancestrally matched population controls ( $n=1,249$ ) were recruited from two sources. The majority ( $n=780$ ) were recruited from London branches of the National Health Service blood transfusion service, from family doctor surgeries and from amongst university students. Individuals were excluded if screening with the *Schedule for Affective Disorders and Schizophrenia* identified a life-time history of depression, bipolar disorder, schizophrenia or alcohol/drug use disorders. Thus, none of the control subjects had a family history of bipolar disorder, schizophrenia or alcohol-dependence; none currently drank above the current UK recommended weekly maximum of 21 (168 g) units for men and 14 (112 g) units for women nor had done so previously. All of the individuals providing blood donations had normal standard liver function tests and were negative for hepatitis surface antigen and core antibody and antibodies to HCV and HIV. The remaining population control samples ( $n=468$ ) were obtained from a separate cohort of controls of British ancestry, that were not screened for psychiatric or alcohol use disorders; these were purchased from the European Collection of Cell Cultures (Health Protection Agency Culture Collections, Salisbury, UK).

## 4.4.2 - SNP GENOTYPING

The variant rs738409 was genotyped in-house using the KASPAR (LGC Genomics, Hoddesdon, UK) genotyping platform and custom designed primers (Table 4-4). Amplification and detection was undertaken using a LightCycler® 480 Real-Time PCR machine (Roche Molecular Diagnostics, Burgess Hill, UK). Genotype calling was performed automatically using proprietary software<sup>367</sup> with some minor manual editing of genotype calls. Approximately 12% of samples, randomly selected a priori, were genotyped in duplicate to ensure consistent genotype calling.

Table 4-4 Primers used for genotyping rs738409 in *PNPLA3*

SNP	Name	Sequence
rs738409	Allele specific (C)	GAAGGTGAACCAAGTTCATGCTCCTTGGTATGTTCTGCTTCATC
	Allele specific (G)	GAAGGTCGGAGTCAACGGATTCTTGGTATGTTCTGCTTCATG
	Reverse 1	CGCCTCTGAAGGAAGGAGGGAT
	Reverse 2	AAGGAGGGATAAGGCCACTGTAGAA

## DATA PROCESSING AND ANALYSIS

Tests for primary allelic association, conditional logistic regression association analysis and quality control measures such as tests of Hardy-Weinberg equilibrium and missingness were performed using the software PLINK (version 1.9)<sup>58,349</sup>. A proportion of samples from the UCL cohort, with imputed genotype data underwent direct genotyping for rs738409. The concordance between imputed and direct genotype data was directly assessed using a custom package<sup>79</sup> in R (version 3.2.2) from which a concordance value was determined.

## 4.4.3 - TIME-TO-EVENT ANALYSIS

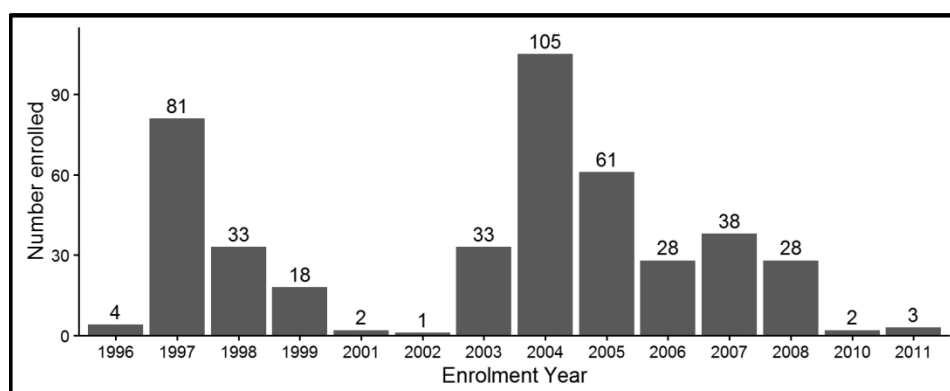
### DATASET

A proportion of the alcohol-dependence cohort recruited from the Royal Free Hospital, London were retrospectively followed until 26<sup>st</sup> October 2015. The study enrolment of these patients occurred from 23<sup>rd</sup> August 1996 until the 1<sup>st</sup> December 2011 (Figure 4-6). In all of these individual patients, several variables were collected including gender, age, length of alcohol misuse (and the age of misuse onset) and self-reported alcohol intake

The time-point at which this information was collected varied by liver disease status:

1. Those patients classified with alcohol-related cirrhosis were phenotypically categorised at the date of clinical presentation with cirrhosis.
2. Those patients classified with no-significant liver injury were phenotypically categorised at the date of enrolment.

In those classified as alcohol-related cirrhosis, the data for several variables relevant to liver function were recorded including LFTs and complications of liver dysfunction such as variceal haemorrhage, ascites and hepatic encephalopathy. These clinical data were used to derive the Pugh score<sup>348</sup> and the MELD score<sup>213</sup>; these scores were dichotomised as compensated or decompensated liver function in each patient at the date of presentation with alcohol-related cirrhosis (MELD score >15: decompensated; Pugh's score >7: decompensated).



**Figure 4-6 The total enrolment of patients by year from the Royal Free Hospital**

The enrolment of alcohol dependent patients from the Royal-Free hospital occurred over a period from 1996 until 2011

## ANALYSIS

### Entire dataset

Differences between the no-significant liver injury and the cirrhosis groups were statistically compared at the date of enrolment/presentation. Dichotomous variables (e.g. gender, decompensation status) were tested using a Pearson-Chi-squared test and continuous variables (e.g. alcohol consumption, age) were tested using a two sample t-test. The difference in mortality between these groups was also compared using Kaplan-Meier time-to-event estimates<sup>215</sup> from the date of birth until an event (death or orthotopic liver transplant). The differences in morbidity between these groups (i.e. presentation with cirrhosis) was compared by rs738409 genotype from the date of birth. Kaplan-Meier time to event estimates were statistically compared using the log-rank test.

### Cirrhosis patients

To detect potential confounding all variables were tested for association with rs738409 genotype at the date of clinical presentation with cirrhosis. To test association Pearson Chi-Squared tests were used for dichotomous variables (e.g. gender, decompensation) and Wald tests were used for continuous variables (e.g. alcohol consumption, age). Univariate and multivariate time-to-event analysis were performed using a Cox proportional hazards regression model<sup>73</sup>. Each variable was tested for association with the time from presentation with cirrhosis until the date of death or orthotopic liver transplant. Variables that were significantly associated with the time-to-event in the univariate analysis ( $P < 0.05$ ) were entered into a multivariate Cox proportional hazards regression model as covariates. The multivariate models, was tested for deviation from Cox proportional hazards assumption using the method of Grambsch and Thernau<sup>151</sup>. Covariates that were in violation of the proportional hazards assumption were entered as time dependent strata in the Cox proportional hazards model (at the minimum strata size in which the proportional hazards assumption is not violated in the model).

Several variables were analysed post hoc in the alcohol-related cirrhosis group. These variables may have influenced the time-to-event but were not suitable for survival modelling either because the factor occurred after presentation with cirrhosis (e.g. development of HCC) or, where the information was incomplete in the majority of samples (e.g. non-abstinence or the cause of death). The development of HCC and patient abstinence following presentation with cirrhosis were analysed by rs738409 genotype status using Chi-Squared test for dichotomous variables and Wald tests for continuous variables. The cause of death was incomplete in the vast majority of samples and therefore was not statistically analysed.

### No-significant liver injury patients

At the date of enrolment all variables were tested for association with rs738409 genotype to detect potential sources of confounding. Pearson Chi-Squared test were performed for dichotomous variables (e.g. gender, decompensation) and Wald tests for continuous variables (e.g. alcohol consumption, age). Univariate and multivariate time-to-event analysis were performed using a Cox proportional hazards regression model<sup>73</sup>. Each variable was tested for association with the time from presentation with cirrhosis until the date of death or orthotopic liver transplant. Variables that were significantly associated with the time-to-event in the univariate analysis ( $P < 0.05$ ) were entered into a multivariate Cox proportional hazards regression model as covariates. The multivariate model was tested for deviation from Cox proportional hazards assumption using the method of Grambsch and Thernau<sup>151</sup>. Covariates that were in

violation of the proportional hazards assumption were entered as time dependent strata in the Cox proportional hazards model (at the minimum strata size in which the proportional hazards assumption is not violated in the model).

### Software

All time-to-event analyses were performed in R<sup>352</sup> using the survival package<sup>426</sup> and several other packages were used for data manipulation, processing and plotting including ggplot2, gridExtra, reshape and plyr. Tests for association between variables collected and the date of presentation/enrolment and rs738409 were performed in PLINK<sup>349</sup>.

#### **4.4.4 - POPULATION ATTRIBUTABLE RISK**

The population attributable risk (PAR) and its 95% confidence intervals were calculated using effect size estimates from the allelic association of this variant when compared between alcohol-related cirrhosis cases and no-significant liver injury controls in the UCL cohort. The equation used for PAR estimation is suitable without the presence of confounding in the genetic association<sup>368</sup>:

$$PAR(\%) = \frac{f_{cases}(RR - 1)}{f_{cases}(RR - 1) + 1} \times 100$$

Abbreviations: PAR – population attributable risk;  $f_{cases}$  – risk allele frequency in cases, RR – relative risk

This calculation requires the relative risk (RR), a measure of the magnitude of an association between an exposed and non-exposed group. However, case/control genetic association tests cannot empirically quantify the RR and instead the OR a related measure of is calculated as a measure of effect size in genetic association studies. The OR and its 95% confidence intervals may be used as proxies for the RR and its 95% confidence intervals<sup>72</sup>. However, this may lead to inflation of the true effect size estimate and hence overestimation of the PAR. Because of this, RR estimates were calculated from the OR and its 95% confidence intervals using the method of Zhang and Yu<sup>478</sup> were used to calculate a separate PAR based on this RR estimate:

$$RR = \frac{OR}{(1 - f_{controls}) + (f_{controls} \times OR)}$$

Abbreviations: RR – relative risk; OR – odds ratio;  $f_{controls}$  – alcohol-related cirrhosis incidence in C-allele carriers

An estimate of the number of individuals affected by alcohol-related cirrhosis of British and Irish ancestry in England was calculated using published data on the prevalence of



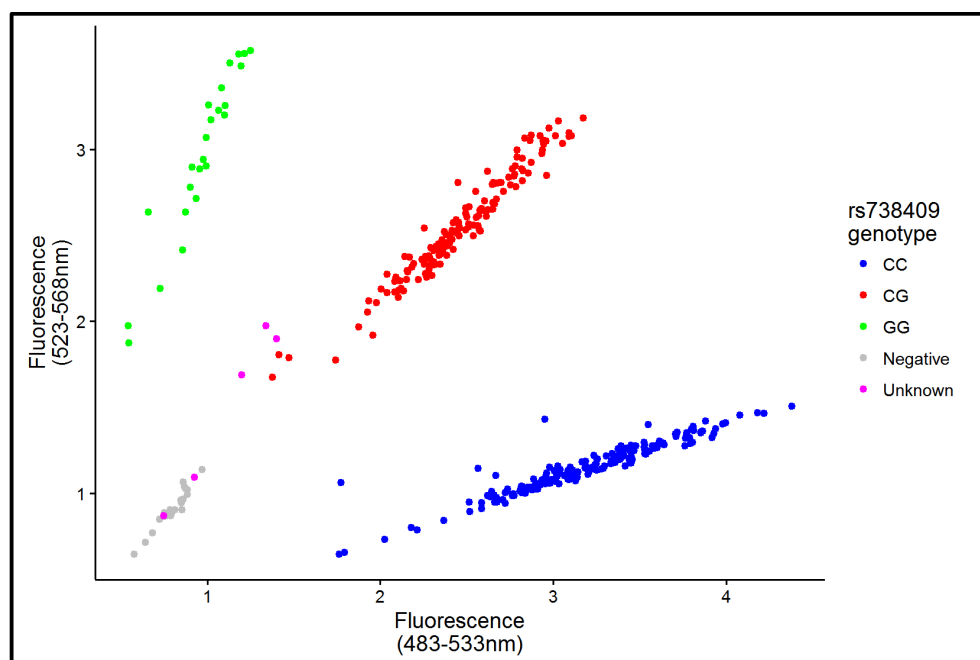
alcohol-dependence in England and Wales<sup>288</sup> and contemporaneous total population estimates of alcohol dependence<sup>318</sup>. The effect of the PAR estimates were compared at range of alcohol-related cirrhosis incidence rates in this population.

## 4.5 - RESULTS

### 4.5.1 - GENOTYPING

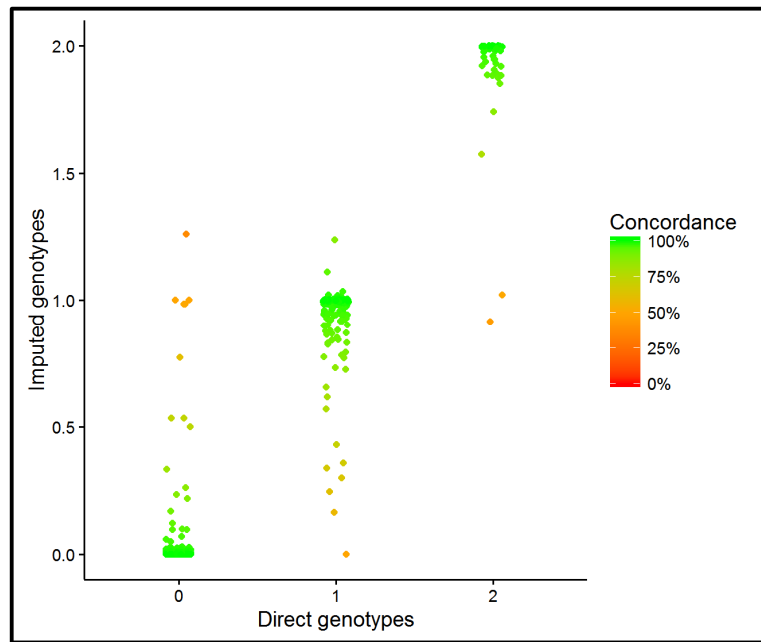
#### GENOTYPING ACCURACY

The KASPAR genotyping succeeded with a genotyping success rate > 95%. In the proportion of samples genotyped in duplicate <1% were in conflict. Visual inspection of genotyping cluster plots demonstrates distinct clustering of the three possible genotypes for each SNP and also the negative controls (Figure 4-7). Rs738409 followed Hardy-Weinberg equilibrium ( $P > 0.01$ ) in all of the comparator populations used for association analysis. In the 628 alcohol-dependent cases with both imputed and direct genotype data for rs738409, the results were highly concordant (96.9%) (Figure 4-8) demonstrating the accuracy of the imputed data, which was used in the GWAS.



**Figure 4-7 A KASPAR genotyping cluster plot for rs738409**

These images demonstrate the fluorescence data used to determine sample genotype for a SNP where each point on the scatter plot represents the recorded fluorescence at two different wavelengths. Genotypes are shown by coloured groups (red, green and blue) which are assigned by a computational algorithm and manual editing. Grey samples represent negative controls which should have limited fluorescence and cluster together. Pink samples represent samples where the genotype cannot be ascertained from these genotype data and are hence classified as unknown.



**Figure 4-8 Concordance between direct and imputed genotype data for rs738409**

The discrete direct genotype values (x-axis) plotted versus the continuous imputed genotype values (y-axis) for rs738409. Each genotype data point is coloured on a continuous scale (from red to green) by its percentage concordance between the imputed and direct genotype data

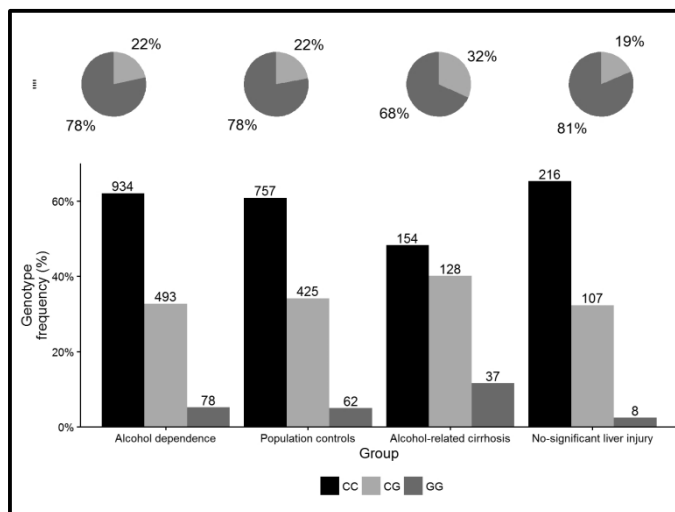
## 4.5.2 - GENETIC ASSOCIATION ANALYSIS

The variant rs738409 in *PNPLA3* was genotyped in a total of 1513 alcohol dependent cases and 1249 population controls. There were no significant differences in the allele frequencies between the alcohol dependent cases and population controls (Table 4-5). In the alcohol-dependent patients that were phenotypically classified by liver disease status a total of 331 alcohol-related cirrhosis cases and 323 no significant liver injury controls underwent genotyping. The rs738409[G] allele was strongly associated with alcohol-related cirrhosis when compared against no-significant liver injury controls ( $P=1.55 \times 10^{-7}$ ; OR=2.01, 95% CI [1.55-2.61]).

Table 4-5 Allelic association analysis of rs738409 in the UCL cohort

Comparison	Group	Number	P-values	Odds Ratio [95% confidence interval]
Alcohol dependence vs. population controls	Cases	1509	0.66	0.97 [0.86-1.10]
	Controls	1249		
Cirrhosis vs. population controls	Cases	323	$8.19 \times 10^{-7}$	1.62 [1.34-1.96]
	Controls	1249		
Cirrhosis vs. no significant liver injury	Cases	323	$1.55 \times 10^{-7}$	2.01 [1.55-2.61]
	Controls	331		

\*Calculated using an allelic Pearsons Chi-squared test



**Figure 4-9 The genotype distribution and allele frequency of rs738409 in the UCL cohort**  
The genotype frequencies for each group of patients in the UCL cohort as bar plots with the number of genotypes in each group given above each bar. The allele frequency of the major and minor allele are given in the pie-charts above each bar plot group.

### 4.5.3 - TIME-TO-EVENT ANALYSIS

Phenotypic data necessary for time-to-event data was collected in 437 patients with alcohol-dependence recruited at the Royal Free hospital. The data necessary for time-to-event analysis was complete in the majority of individuals (95.2%). There were a greater number of men (70.9%) and at the date of enrolment or presentation with cirrhosis: the mean age was 53.2 years; the mean length of drinking history was 25.2 years; the mean age of alcohol-misuse onset was 28.0 years; and, the mean self-reported alcohol consumption was 190.9 units/week.

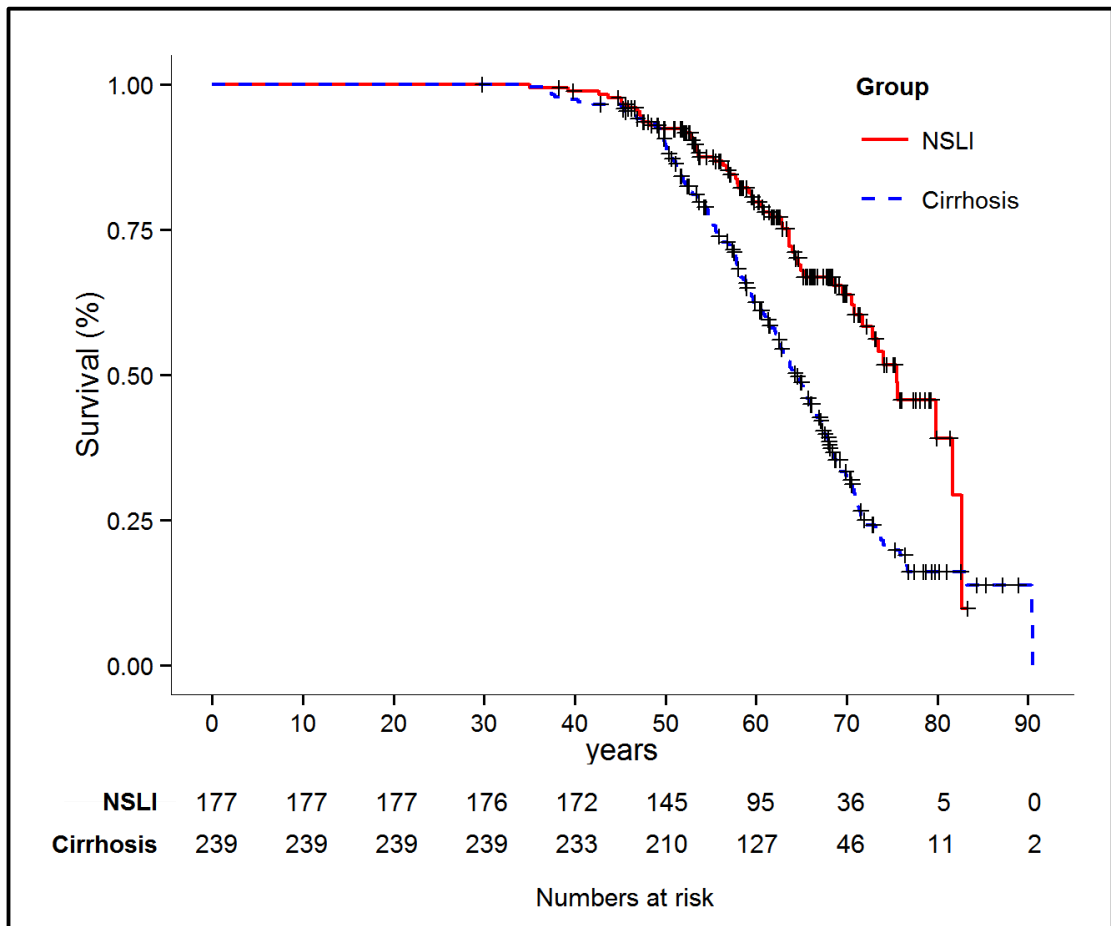
There were two phenotypically categorised groups: those with alcohol-related cirrhosis and those with no-significant liver injury (Table 4-6). At the baseline (date of enrolment or the date of clinical presentation with cirrhosis), there were significant differences between these groups: the alcohol-related cirrhosis group were more likely to be female, began misusing alcohol at an older age, had a higher frequency of the rs738409[G] allele and had a reduced lifespan (Table 4-6). Patients with cirrhosis, have on average, a ten-year shorter lifespan than those with no-significant liver injury (Figure 4-10).

In all patients, the time from the date of birth until the date of clinical presentation with cirrhosis was estimated and stratified by rs738409 genotype. Those patients that did not present with cirrhosis were censored at the date of death or the final time point in which they were last known to be alive. There were significant differences by rs738409 genotype ( $P_{\text{log-rank}} = 5.74 \times 10^{-6}$ ) (Figure 4-11, Table 4-7) and on average carriers of the rs738409[GG] genotype presented with cirrhosis approximately 10 years earlier than those with the rs738409[CC] genotype.

Table 4-6 Base lines demographics of the time-to-event cohort by liver disease status

Variable	Cirrhosis (n=254)	No-Liver Injury (n=183)	P-value
Gender (Male)	160 (66.9%)	135 (76%)	0.049*
Alcohol consumption (units/week)	186.1	197.7	0.3**
Age of Alcohol-misuse onset (years)	29.3	26.4	$3.58 \times 10^{-3}$ **
Length of drinking history (years)	25.8	24.5	0.16**
rs738409 genotype CC/CG/GG (%)	109/99/31 (45.6%/41.4%/12.9%)	114/61/2 (64.4%/34.5%/1.12%)	$2.31 \times 10^{-6}$ *
Median KM time-to-event estimate (years)	64.3	75.5	$1.95 \times 10^{-6}$ **

The average is given for continuous variables and the count and frequency for discrete variables. Abbreviations: KM – Kaplan-Meier. Statistical tests: \* Pearson's Chi-squared test; \*\* Welch Two Sample t-test; \*\*\*Log-rank test

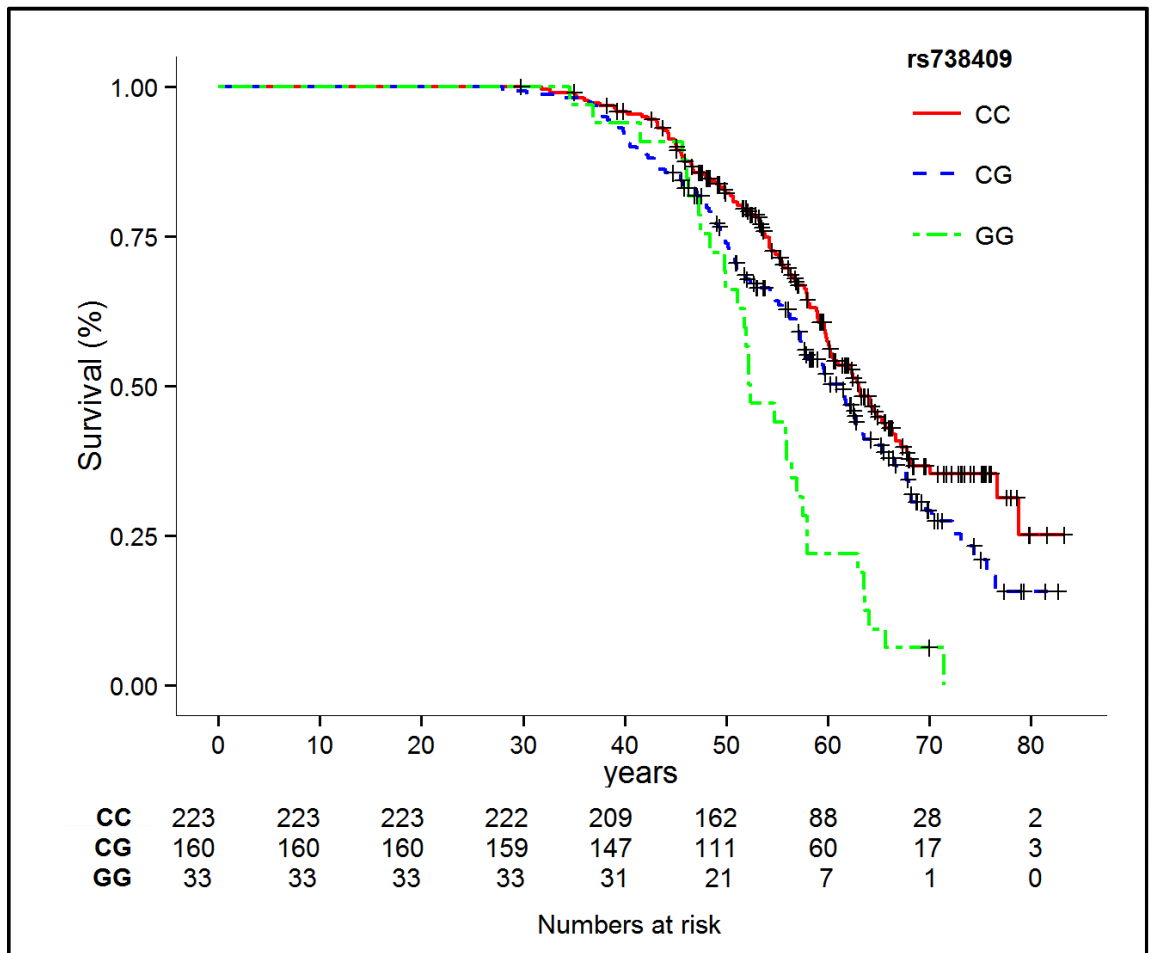


**Figure 4-10 Kaplan-Meier curve in all patients stratified by liver status**

This Kaplan-Meier curve shows the estimated survival curves stratified by liver disease status (cirrhosis or no-significant liver injury) from birth until and death or liver transplant  
 Abbreviations: NSLI – no-significant liver injury

**Table 4-7 Median Kaplan-Meier time-to-event estimates from the date of birth until the date of presentation with cirrhosis stratified by rs738409 genotype**

Variable	Strata	Median time-to-event (years) [95% CI]
rs738409 genotype	CC	63.0 [60.1–66.7]
	CG	61.4 [57.4–64.3]
	GG	52.4 [51.1-57.6]



**Figure 4-11 Kaplan-Meier curve in all patients stratified by rs738409 genotype**

This Kaplan-Meier curve shows the estimated time-to-event curves stratified by rs738409 genotype from the date of birth until the date of presentation with cirrhosis

## CIRRHOSIS PATIENTS

Over the period of data collection, an event (death or orthotopic liver transplant) occurred in 161 (68.6%) of the patients that presented with alcohol-related cirrhosis and the 78 patients alive at the date of last observation were censored. Of all the observed events, 142 (88.2%) were deaths and 19 were orthotopic liver transplants (11.8%). The baseline clinical and demographic variables collected at the date of presentation did not differ by rs738409 genotype (Table 4-8). Several variables associated with this time-to-event (Table 4-8) including the carriage of the rs738409[G] allele, which was associated with an increased the chance of death or liver transplantation by 26% at any time-point.

Covariates deemed significantly associated in the univariate test ( $P_{\text{UNIVARIATE}} < 0.05$ ) were entered as covariates into a multivariate Cox proportional hazards model. None of the variables in this model deviated from the proportional hazards assumption. Despite its significance, the MELD score was not included in the multivariate model for two reasons: (i) because it is an approximate for the Pugh score; and, (ii) because the

MELD score data was missing in a greater proportion ( $n=37$  (15.5%)). In the multivariate model, both gender and the Pugh score remained independently associated with the time-to-event ( $P_{\text{MULTIVARIATE}} < 0.05$ ). Survival curves were calculated in the alcohol-related cirrhosis group when stratified by those variables that were significantly associated with the time-to-event ( $P_{\text{UNIVARIATE}} < 0.05$ ) (Figure 4-12, Table 4-10). On average, death or liver transplant occurred 5 years earlier for individuals that are homozygote for the rs738409[GG] genotype than carriers of the rs738409[CC] genotype.

Table 4-8 Baseline clinical and demographic features in patients with cirrhosis stratified by rs738409 genotype

Variable	CC ( $n=109$ )	CG ( $n=99$ )	GG ( $n=31$ )	P-value
Sex (male)	67 (61.46%)	69 (69.69%)	24 (77.4%)	0.19*
Pugh status (decompensated)	40 (43.01%)	47 (51.08%)	14 (51.85%)	0.49*
MELD status (decompensated)	31 (34.06%)	50 (35.29%)	10 (38.46%)	0.92*
Alcohol consumption (units/week)	184.35	187.28	188.7	0.77**
Age at presentation (years)	53.76	53.12	53.26	0.70**
Age of Alcohol-misuse onset (years)	55.69	54.41	55.43	0.62**
Length of drinking history (years)	26.22	25.14	26.77	0.91**

The mean is given for continuous variables and the count and frequency for discrete variables.  
Statistical tests: \*allelic Pearson's Chi-squared test; \*\*Wald-test

Table 4-9 Cox proportional hazards analysis of the time from presentation with cirrhosis until death or orthotopic liver transplant using univariate and multivariate models

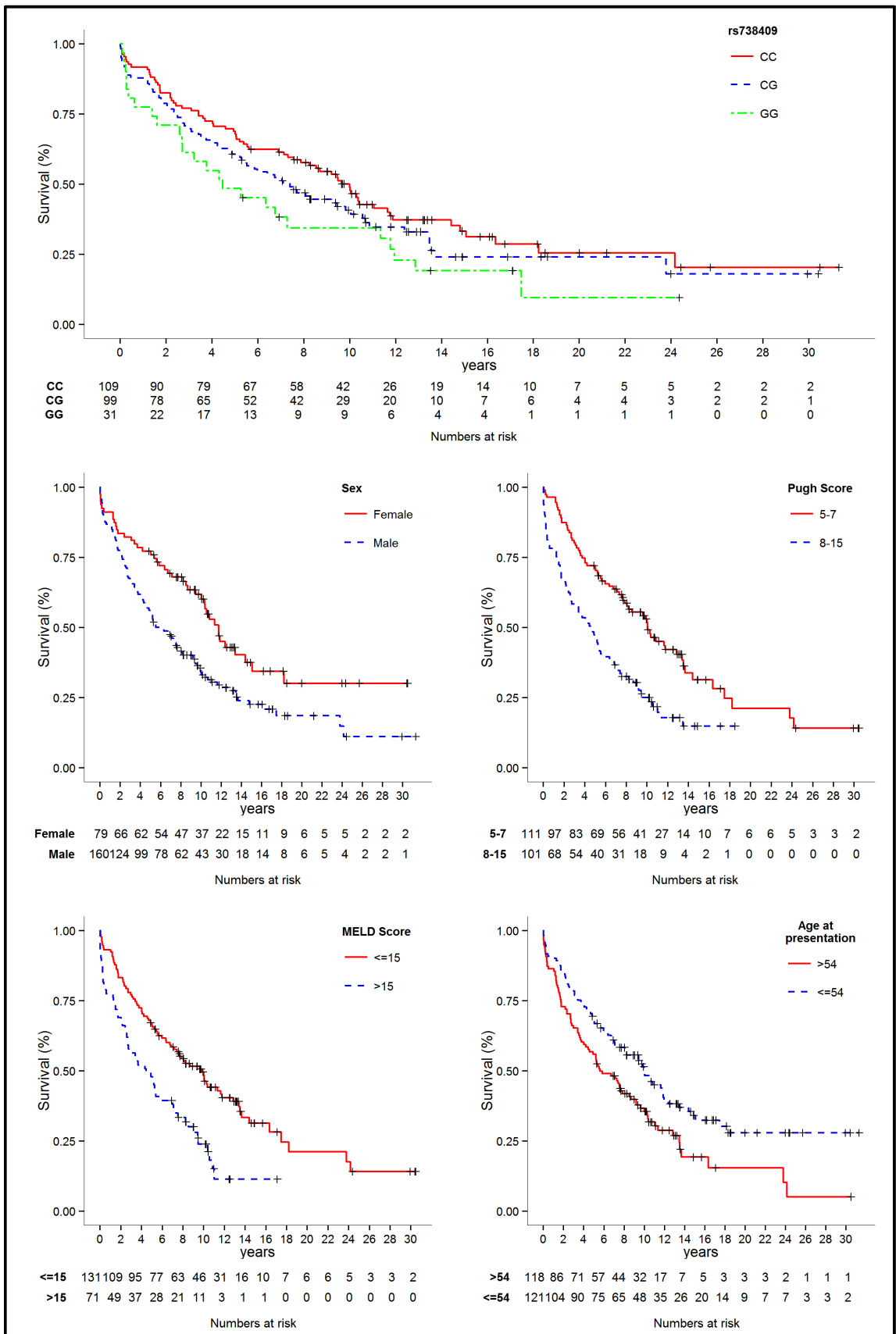
Covariate	Univariate			Multivariate		
	Hazard Ratio	P-value	Number	Hazard Ratio	P-value	Number
Sex (male)	1.76	0.0016	239	1.57	0.018	212
Age at presentation (years)	1.02	0.0051	239	1.02	0.075	212
Pugh score 5-7 (decompensated)	2.02	3.13x10 <sup>-5</sup>	212	2.01	4.3x10 <sup>-5</sup>	212
MELD score ≤15 (decompensated)	2.01	9.16x10 <sup>-5</sup>	202	-	-	-
rs738409 genotype (CC=0, CG=1, GG=2)	1.26	0.040	239	1.15	0.24	212
Abstinence	0.86	0.68	110	-	-	-
Length of alcohol misuse (years)	1.01	0.37	239	-	-	-
Alcohol consumption (units/week)	1.00	0.79	239	-	-	-
Age of misuse onset (years)	0.99	0.97	239	-	-	-

Table 4-10 Kaplan-Meier median time-to-event estimates for several variables associated with the time to death or transplant from cirrhosis presentation

Variable	Strata	Median time-to-event in years [95% CI]
rs738409 genotype	CC	10.00 [7.87-11.8]
	CG	7.40 [5.29-10.6]
	GG	4.47 [2.69-11.8]
Gender	Male	5.55 [4.71-7.87]
	Female	11.76 [10.06-18.23]
Pugh Score	Compensated (5-7)	10.06 [8.09-13.52]
	Decompensated (8-15)	4.59 [2.73-6.42]
MELD Score	Compensated (≤15)	9.82 [7.14-12.90]
	Decompensated (>15)	4.47 [2.71-7.10]
Age at Presentation (years)	≤54	10.06 [8.09-12.40]
	>54	5.61 [4.47-8.71]

Abbreviations: MELD – Model of end-stage liver disease; CI – Confidence interval



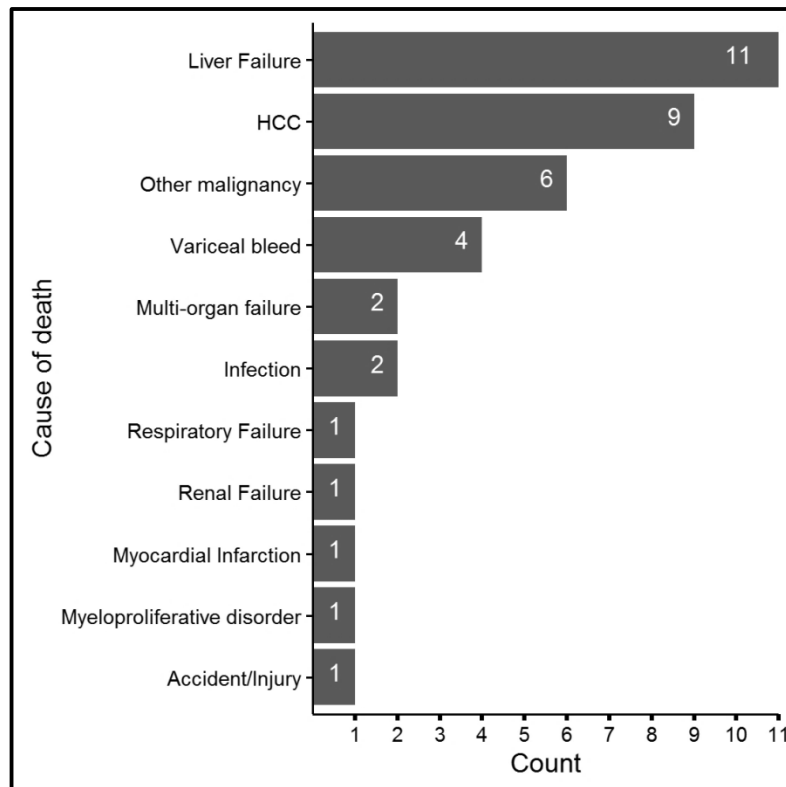


**Figure 4-12 Kaplan-Meier survival curves in patients with cirrhosis**

Kaplan-Meier survival curves from the date of presentation with cirrhosis until the date of death or orthotopic liver transplant were stratified by the variables deemed significant ( $P < 0.05$ ) in univariate Cox-regression analysis with the at risk tables for each strata given below each plot. Continuous covariates were stratified by their median value.

## Causes of Death

The cause of a death was identified in a small number of patients with cirrhosis (Figure 4-13). The predominant causes of death were liver failure and HCC. Many of the other causes of death were associated with complications arising from cirrhosis (e.g. variceal bleeds, renal failure).



**Figure 4-13 The causes of death or liver transplant in patients with cirrhosis**

## HCC

In total 16 patients with alcohol-related cirrhosis group developed HCC. Several clinical and demographic variables, including rs738409 genotype were statistically compared by HCC status (Table 4-11). Patients with HCC were more likely to be male and carriers of the rs738409[G] allele. The development of HCC was not significantly associated with the time from presentation until death or liver transplant ( $P < 0.05$ ) although the development of HCC on average reduces the time until death or transplant by 3.5 years in patients with cirrhosis.

Table 4-11 Baseline clinical demographic features of patients with cirrhosis as stratified by the subsequent development of hepatocellular carcinoma

	HCC & Cirrhosis (n=16)	Cirrhosis (n=223)	P value
Gender (Male)	15 (93.75%)	145 (65.02%)	0.037*
Pugh score 8-15 (decompensated)	8 (46.5%)	93 (66.7%)	0.29*
MELD score >15 (decompensated)	5 (34.5%)	66 (45.4)	0.68*
Alcohol consumption (units/week)	164.7	187.7	0.22**
Age at clinical presentation with cirrhosis (years)	56.6	53.2	0.06**
Length of alcohol-misuse (years)	30.5	25.5	0.051**
rs738409 genotype (CC/CG/GG)	3/7/6 (18.8%/43.8%/37.5%)	106/92/25 (47.5%/41.2%/11.2%)	0.0015***
KM median time to death or OLTx (years)	5.2	8.5	0.21****

The mean is given for continuous variables and the count and frequency for discrete variables.  
Abbreviations: KM – Kaplan-Meier, OLTx – orthotopic liver transplant. Statistical tests:  
\*Pearson's Chi-squared test; \*\* Welch Two Sample t-test; \*\*\*Allelic Pearson's Chi-squared test;  
\*\*\*\*Log-rank test

## Abstinence

Alcohol use recidivism data following presentation with cirrhosis was available in a minority of patients (n=110). None of the clinical, genetic or demographic variables differed by abstinence status (Table 4-12).

Table 4-12 Test for association between several variables and alcohol use recidivism in the alcohol-related cirrhosis time-to-event dataset

	Abstinent (n=63)	Non-abstinent (n=47)	P value
Gender (Male)	33 (52.3%)	28 (59.5%)	0.58*
Pugh (decompensated)	20 (40.8%)	21 (50.0%)	0.51*
MELD (decompensated)	17 (35.4%)	12 (29.2)%	0.70*
Alcohol consumption (units/week)	178.6	177.7	0.96**
Age at cirrhosis presentation (years)	51.9	50.7	0.59**
Length of drinking history (years)	24.5	24.6	0.92**
rs738409 genotype (CC/CG/GG)	38/16/9 (60.3%/25.4%/14.28%)	20/24/3 (42.5%/51.1%/6.38%)	0.43***
KM median time to event (years)	NA	NA	0.61****

The mean is given for continuous variables and the count and frequency for discrete variables.  
Abbreviations: KM – Kaplan-Meier. Statistical tests: \* Pearson's Chi-squared test; \*\* Welch Two  
Sample t-test; \*\*\*Allelic Pearson's Chi-squared test; \*\*\*\*Log-rank test

## NO-SIGNIFICANT LIVER INJURY PATIENTS

Out of the 177 patients with no-significant liver, 57 (32.2%) died between the date of enrolment and the last follow-up. The remaining 120 patients (67.8%) were censored at the date of last observation. At the baseline (date of enrolment) there were no differences by rs738409 genotype for the several clinical and demographic variables collected (Table 4-13). Several variables were associated with the time-to-event (Table 4-14) including carriage of the rs738409[G] allele, which was associated with a decreased chance of death by 48% at any time-point. Other significantly associated variables included the age at enrolment (4% increased chance of death for every additional year of age at enrolment) and age of alcohol-misuse onset (3% increased chance of death for every additional year of age of alcohol misuse onset).

The significantly associated variables ( $P_{UNIVARIATE} < 0.05$ ) were entered into a multivariate Cox proportional hazards model. The age of enrolment covariate violated the Cox-proportional hazards assumption; this violation was accounted for by stratifying the model into tripartite strata by the age of enrolment. In the final model both rs738409 allele status and the age at enrolment were associated with the time to event ( $P_{MULTIVARIATE} < 0.05$ ) (Table 4-14) (Figure 4-14).

Table 4-13 Baseline demographics of the no-significant liver injury patients stratified by rs738409 genotype

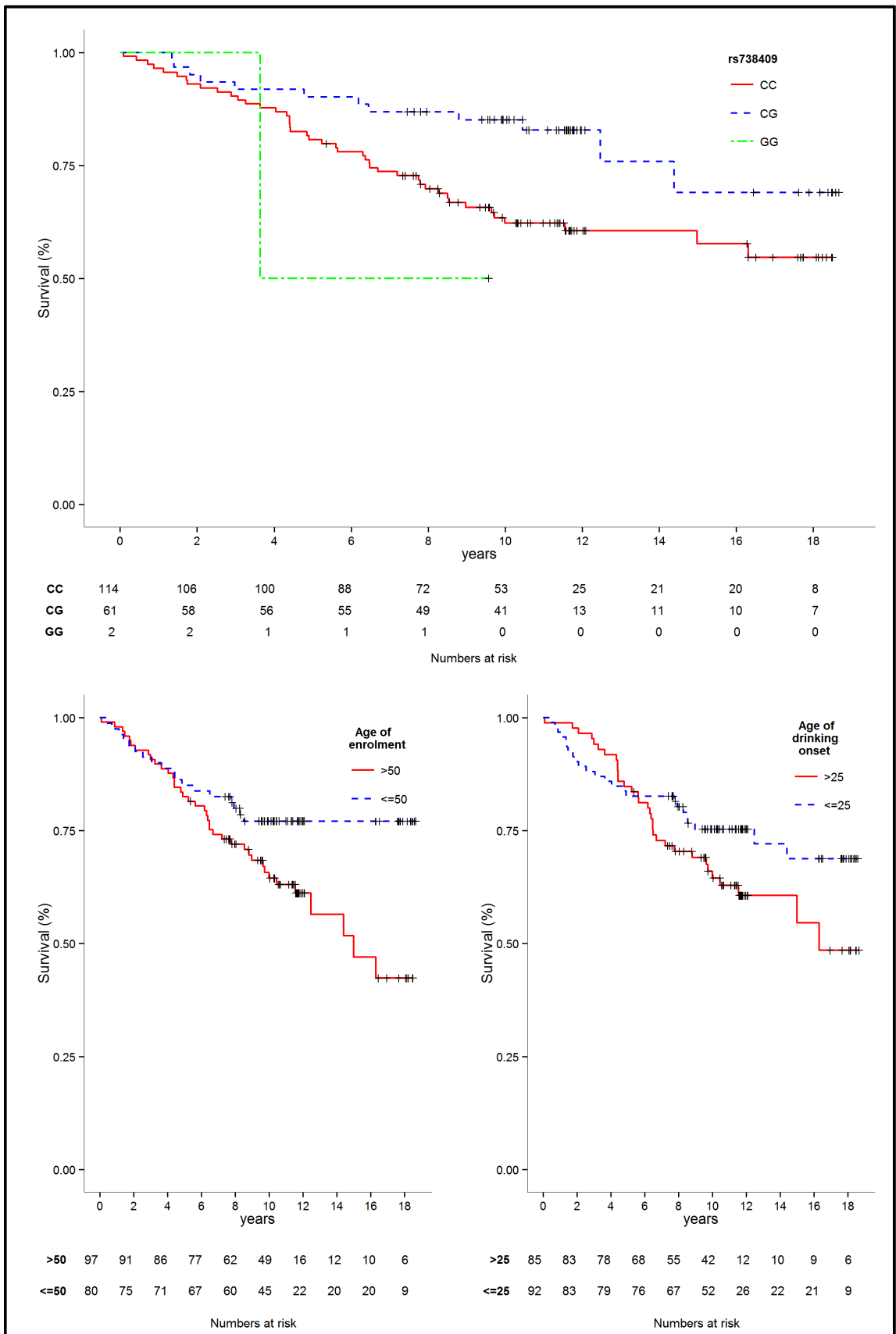
Variable	CC (n=117)	CG (n=64)	GG (n=2)	P-value
Sex (male)	88 (75.2%)	49 (76.5%)	2 (100%)	0.71*
Alcohol consumption (units/week)	197.2	198.2	270	0.74**
Age at enrolment (years)	50.71	50.69	51.8	0.97**
Age of Alcohol-misuse onset (years)	26.73	25.35	25.8	0.38**
Length of drinking history (years)	23.97	25.34	26	0.96**

The mean is given for continuous variables and the count and frequency for discrete variables. Statistical tests: \*allelic Pearson's Chi-squared test; \*\*Wald-test

Table 4-14 Cox proportional hazards time to death from enrolment analysis in the no-significant liver injury group

Covariate	Univariate			Multivariate*		
	Hazard Ratio	P-value	Number	Hazard Ratio	P-value	Number
Sex (male)	0.70	0.22	177	-	-	-
Age at enrolment (years)	1.04	0.0051	177	1.03	0.016	177
rs738409 genotype (CC=0, CG=1, GG=2)	0.52	0.033	177	0.52	0.035	177
Length of alcohol misuse (years)	1.02	0.24	177	-	-	-
Alcohol consumption (units/week)	1.00	0.71	177	-	-	-
Age of misuse onset (years)	1.03	0.051	177	0.97	0.41	177

\*The multivariate model was stratified by tripartite strata at the age at enrolment to avoid violations of the proportional hazards assumption.

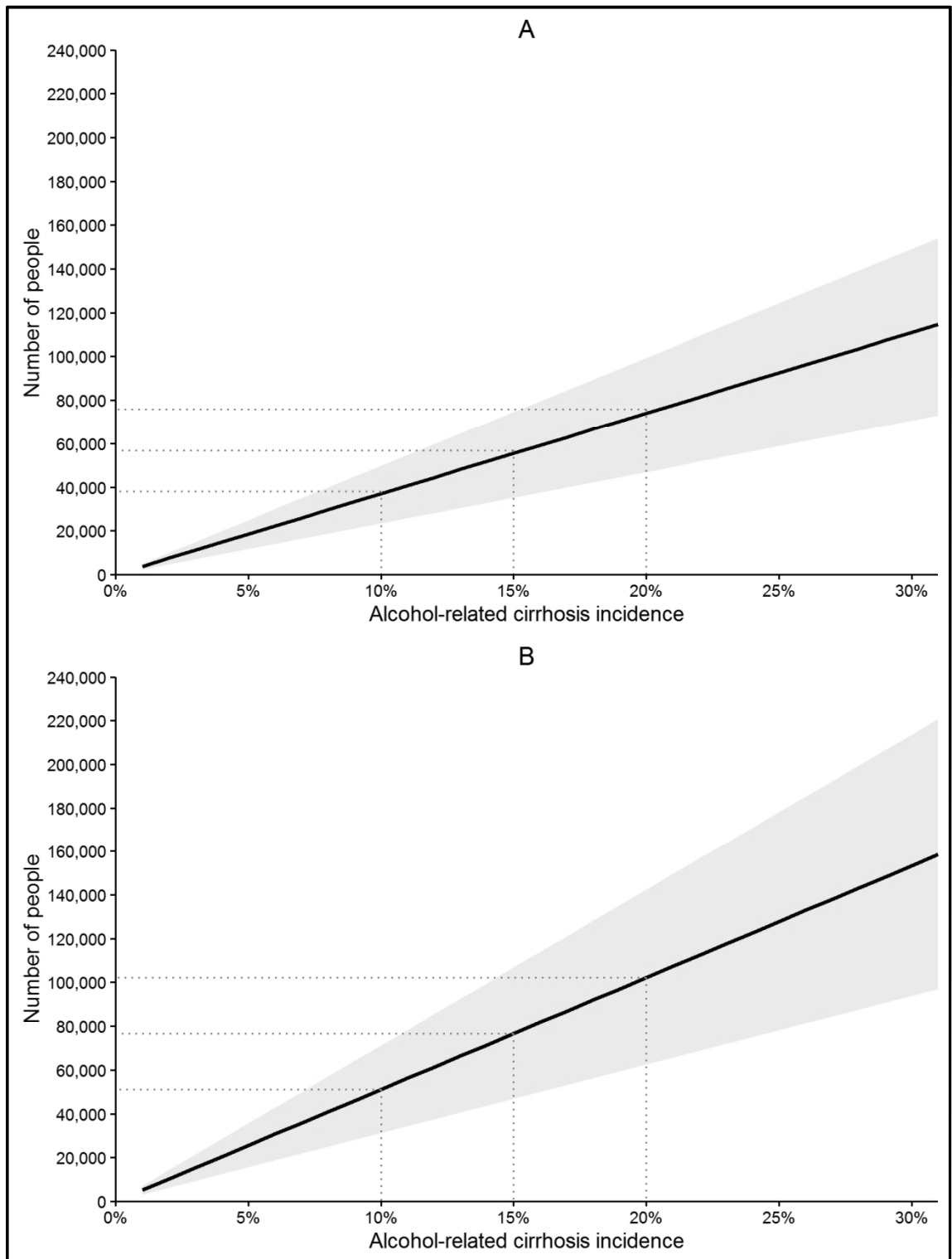


**Figure 4-14 Kaplan-Meier survival curves in patients with no significant liver injury**  
 Kaplan-Meier survival curves from the date enrolment until the date of death were stratified by the variables deemed significant ( $P < 0.05$ ) in univariate Cox-regression analysis with the at risk tables for each strata given below each plot. Continuous covariates were stratified by their median value.

#### 4.5.4 - POPULATION ATTRIBUTABLE RISK

The PAR estimates for rs738409 were derived from the OR of the allelic association test between alcohol-related cirrhosis cases and no-significantly liver injury controls (OR = 2.01, 95% CI [1.55-2.61]). This OR was used to calculate an estimate of the RR of the genetic association ( $RR_{\text{estimate}} = 1.69$ , 95% CI [1.41-2.01]). In this British and Irish cirrhosis study population the PAR for the carriage of the rs738409[G] cirrhosis risk allele was substantial accounting for around one fifth of the PAR when using the OR (PAR%=24.23, 95% CI [14.06-20.09]) and when using an RR estimate (PAR%=17.98%, 95% CI [11.39-24.21]).

In 2009, the recorded population of ethnic British and Irish adults (16-75+ years) living in England was 35,597,700. In this population, 8.7% of men and 3.3% of women have reported alcohol-dependence, totalling 2,109,229 alcohol-dependent individuals. Based on a 20% incidence rate of alcohol-related cirrhosis in patients with alcohol dependence<sup>241,423</sup>, it is estimated that 421,846 individuals in England of British or Irish ancestry, will develop alcohol-related cirrhosis over their lifetime. Based on this estimate and the calculated population attributable risk for rs738409, if the rs738409[G] allele was not present in the British and Irish population of England there would be 102,209 (95% CI [62,562-142,427]) fewer cases of alcohol-related cirrhosis in this population (Figure 4-15). Calculations based on the relative risk estimate are lower although still substantial.



**Figure 4-15 Population attributable risk estimates of alcohol-related cirrhosis by rs738409 genotype at difference incidence levels**

These images show the estimated number of alcohol-related cirrhosis cases that would not be present if the rs738409[G] allele was absent from British and Irish ancestry population of England and Wales. These estimates were derived from the odds ratio (image A) and a relative risk estimate (image B). Both images show the estimated number of people (y-axis) at different incidence levels of alcohol-related cirrhosis (x-axis) that would not be present in dependent drinkers of British and Irish ancestry in England and Wales (black line) with the 95% confidence interval of the estimate (grey outline). A range of likely alcohol-related cirrhosis incidence levels in alcohol-dependent populations are highlighted (dashed red lines).



## DISCUSSION

Alcohol-related cirrhosis and alcohol-dependence are both influenced by genetic risk factors. Hence, confounding may occur in genetic association studies without appropriate case/control comparisons, namely comparing alcohol-related cirrhosis cases with both population controls and alcohol-misusers with no-significant liver injury. Through the direct genotyping of rs738409 in a large ( $n=2,762$ ), ancestrally homogeneous and well characterized cohort this analysis demonstrates that there is no evidence for genetic association between rs738409 and alcohol dependence.

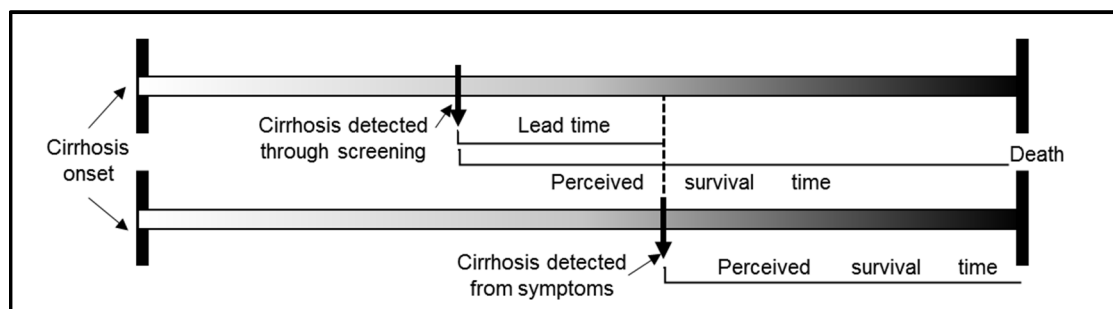
The variant rs738409 is firmly associated with alcohol-related cirrhosis risk<sup>55</sup>; this finding is typified as this was the most significantly associated variant in the GWAS of alcohol-related cirrhosis<sup>44</sup>. In the current work, the genetic association results for rs738409 from the GWAS are replicated albeit through direct genotype data rather than probabilistic imputed genotype data. The results are highly concordant with the imputed data (rs738409 UK-GWAS imputed data:  $P=2.51 \times 10^{-8}$ , OR=2.12 95% CI [1.63-2.77]; rs738409 direct genotyping data:  $P=1.55 \times 10^{-7}$ , OR=2.01, 95% CI [1.55-2.61]). The subtle differences in significance and effect size between these data likely reflect the slightly different samples used (Figure 4-5), differences in the genetic association test resulting from the use of different analytical software, and also minor differences between the empirical direct genotype data and the probabilistic imputed genotype data. The effect size of the rs738409 association with alcohol-related cirrhosis is greater in the comparison with long-term alcohol misusers with no-significant liver injury (OR=2.01, 95% CI [1.55-2.61]) than with population controls (OR=1.62, 95% CI [1.34-1.96]). This finding is likely due to increased relative statistical power, which derives from using phenotypically, screened controls and is consistent with similar comparisons in other cohorts<sup>55</sup>.

There is increasing awareness that rs738409 genotype may influence prognosis in alcohol-related liver disease<sup>46,55,443</sup>. In the current work, rs738409 was investigated for such effects in a sub-cohort of alcohol-dependent patients in which data was collected, in some, over a twenty year period from the date of enrolment or clinical presentation with alcohol-related cirrhosis until death or orthotopic liver transplant. The primary phenotypic difference between patients was the presentation with alcohol-related cirrhosis which was associated with a significantly worse prognosis than those with no-significant liver injury. Notably, in this stratified cohort the rs738409[G] allele carriers clinically present with alcohol-related cirrhosis at a younger age. This findings is in agreement with a previous study which also demonstrates that carriers of the rs738409[G] allele have an earlier age of alcohol-related cirrhosis onset<sup>46</sup>.

The primary time-to-event analysis was from the date of presentation/enrolment until death or orthotopic liver transplantation; this was performed separately in the alcohol-related cirrhosis and no-significant liver injury sub-groups.

In the cirrhosis sub-group, several variables (Pugh score, MELD score, gender and rs738409 allele status) were associated with the time from clinical presentation with cirrhosis until death or orthotopic liver transplant. A primary finding from this analysis was that of an increased hazard of death or orthotopic liver transplant (HR= 26% per G allele) for each rs738409[G] allele carried. This finding did not remain significant in the multivariate analysis and only decompensated liver disease status (as determined by a Pugh score) and male gender were independently associated with the time-to-event. As the multivariate analysis was performed with a reduced dataset this may have reduced the statistical power to detect the association between rs738409 allele status which was observed in the univariate analysis. This finding is consistent with the association between the rs738409[G] allele and reduced transplantation free survival in a population waiting for liver transplant<sup>133</sup>.

The presentation with alcohol-related cirrhosis may occur incidentally, due to referral based on abnormal LFTs, or symptomatically due to the development of liver failure. These differences in presentation result from the presence, or absence, of clinical symptoms and therefore the date of presentation is subject to a lead time bias<sup>445</sup> (Figure 4-16). In disease modalities such as cancer<sup>77</sup>, where screening is a key strategy to reduce mortality, lead time bias can result in an overestimation of survival time. However, this was unlikely to have occurred in the present study as patients with cirrhosis, at the baseline date of presentation, did not have significantly different rs738409 allele frequencies by either the MELD or CTP scores.



**Figure 4-16 Lead time bias**

The occurrence of lead time bias results from the earlier detection of a disease, in comparison to the typical symptomatic presentation, which may result in an overestimation of the perceived survival time without affecting the course of the disease

The major causes of death in the alcohol-related cirrhosis sub-group were liver failure and HCC. This is the first study to demonstrate a significant association between HCC risk developing from alcohol-related cirrhosis and the rs738409 variant in a British and

Irish ancestry cohort. This finding replicates those findings from several other European ancestry cohorts<sup>432</sup>. HCC is associated with a worse prognosis in cirrhosis and therefore, rs738409 genotype information may have clinical utility in identifying those requiring greater HCC surveillance.

Unlike the other variables, the data regarding abstinence post-diagnosis with cirrhosis was incomplete in the majority of patients with cirrhosis. In those patients in which this information was obtained, it was not associated with the time until death or liver transplant; a finding is incongruous with clinical guidelines<sup>444</sup> and the findings from several studies<sup>469</sup>. However, the quality of this data may be limited by several factors: (i) abstinence data was missing for a majority of the sub-group; (ii) abstinence data were dichotomised due to lack of information and were therefore less informative than quantitative data (e.g. amount and length of alcohol consumption after cirrhosis diagnosis); and, (iii) abstinence data was collected retrospectively and therefore has a potential to be influenced by ascertainment bias.

In patients with no-significant liver time-to-event comparisons were performed from the date of enrolment until death. Surprisingly in those with no-significant liver injury, the rs738409[G] allele, which is associated with an increased risk of alcohol-related cirrhosis, was associated with a decreased hazard ratio of a death (HR=0.52). This finding remained significantly associated in the multivariate model with both the age at enrolment and rs738409 as independent predictors of prognosis. This is the first report of an association between the rs738409[G] alcohol-related cirrhosis risk allele and a favourable prognosis in a population of alcohol-misusers with no-significant liver injury. A hypothetical mechanism through which rs738409[G] may reduce mortality in alcohol-misusers without alcohol-related cirrhosis is through lowering serum triglyceride levels and hence cardiovascular disease risk<sup>257</sup>. In studies of rs738409 in NAFLD, carriers of the rs738409[G] allele have lower mean serum triglycerides, a lower Framingham cardiovascular risk score and lower prevalence of metabolic syndrome<sup>91</sup>. Functionally the rs738409 variant may impact the secretion of triglycerides from the liver as in a cohort of overweight/obese men this variant explained 46% of the variance in VLDL triglyceride secretion from the liver into the plasma; the rs738409[G] allele was associated with a reduced VLDL secretion capacity from the liver<sup>335</sup>. A similar, so-called 'catch-22 situation', has already been suggested for such a two-way disease/allele risk relationship between the alleles of the functional variant rs58542926 in *TM6SF2* and either cardiovascular or liver disease risk<sup>209</sup>.

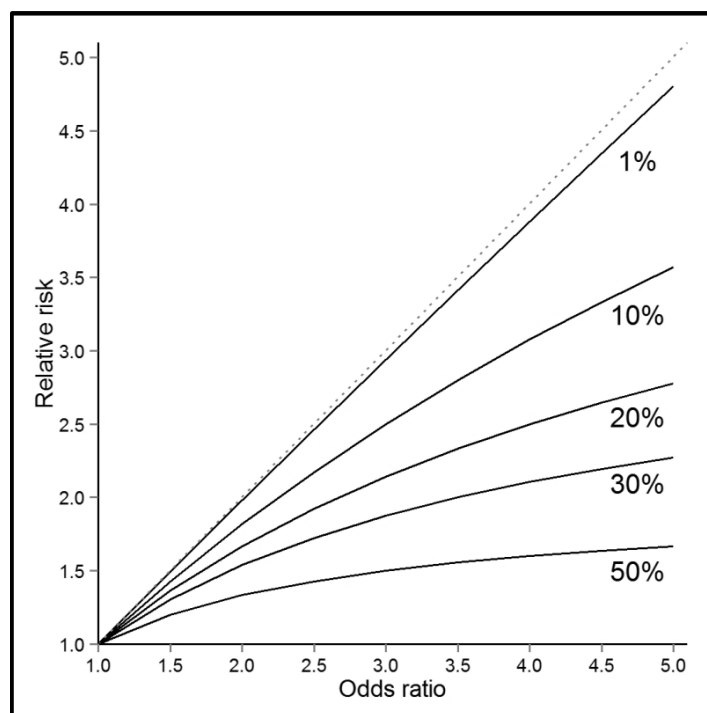
The primary time-to-event analyses were performed using the Cox proportional hazards model<sup>73</sup>. This statistical model has several advantages over other time-to-event analytical models such as the nonparametric Kaplan-Meier time-to-event

estimates<sup>215</sup> and their statistical comparison using the log-rank test or more complex regression based parametric models. The primary advantage of the Cox proportional hazards model is its ability to include data from several variables for multivariate analysis while not requiring the detailed knowledge of a survival distributions as is required for parametric models. However, there are two primary assumptions of this model: first, the non-informative censoring assumption (i.e. the cause of censoring is independent of the event); and second, the proportional hazards assumption (i.e. the hazard functions of each variable are constant over the time interval). In the current analysis all of the samples were censored at the date of last clinical observation and where this information was unclear they were excluded from analysis reducing the likelihood of non-informative censoring. Also, the proportional hazards assumption was tested using the method of Grambsch and Thernau<sup>151</sup> in which each variable is statistically tested for deviation from the proportional hazards assumption and visually described using weighted residual plots. Any variables that violated the proportional hazards assumption were stratified until there was no evidence for non-proportional hazards in the multivariate model reducing the likelihood of proportional hazards violation.

The rs738409[G] allele is common in people of British and Irish ancestry (allele frequency=22.1%) and has a relatively large effect on alcohol-related cirrhosis risk for a common genetic variant on a complex disease phenotype (OR=2.01, 95% CI [1.55-2.61]). Hence, it may have utility in identifying patients with increased risk of developing alcohol-related cirrhosis; indeed, this idea has already been suggested<sup>428</sup>. The allele frequency and OR values do not provide readily interpretable information for understanding the effect of the variant on disease incidence at a population level. This is the use of the PAR, which provides an indication of the percentage reduction in the incidence rate of a disease if the risk factor were not present in a population. When calculated from the OR, the rs738409[G] allele accounts for around one-fifth (PAR% = 20.49%, 95% CI [14.39%-26.57%]) of the population attributable risk for alcohol-related cirrhosis in this British and Irish ancestry population, a figure that is comparable to another estimate in German ancestry population<sup>413</sup>. This estimate of the population attributable risk is sizeable indicating that knowledge of rs738409 genotype may have some use in the identification of individuals who are at increased risk of developing alcohol-related cirrhosis.

In case/control genetic association studies the PAR is often calculated using the OR as a proxy for the RR<sup>413</sup>. These two measures of effect size are not equivalent; the OR nearly always overinflates the RR<sup>83,276</sup>. The scale of this over inflation is dependent on the prevalence of the disease phenotype in a population and the scale of its true effect size. For diseases with a low population frequency the RR and OR asymptotically

approach, and hence are equivalent; this is the rare disease assumption<sup>72</sup> (Figure 4-17). As the RR, cannot be empirically determined from the case/control study design employed in this analysis, several precautions were undertaken to minimize and account for PAR inflation resulting from using the OR. First, PAR estimates were derived from both the OR and an estimate of the RR using the method of Zhang and Kai<sup>478</sup> and second, 95% confidence intervals were generated for comparisons to account for uncertainty in the estimate. As expected the PAR derived from the RR estimate was lower than that derived from the OR yet nevertheless both demonstrate a substantial PAR estimate and there would likely be nearly a fifth fewer alcohol-related cirrhosis cases occurring in British and Irish ancestry populations if this variant were not present.



**Figure 4-17 The relationship between relative risk and odds ratios**

At increasing disease incidence levels (from 1% to 50%) and increasing effect sizes the relative risk and odds ratio become less equivalent. These data were calculated using the formula of Zhang and Kai, 1998<sup>478</sup>

Oxymoronically, a major strength and weakness of this analysis was the case/control study design<sup>386</sup>. Case/control studies require the retrospective selection of patients, in this instance by liver disease status, and therefore are not a representative population per se. Several features of this analysis minimized potential bias resulting from case/control selection: (i) for time-to-event analysis patients with alcohol-related cirrhosis and no-significant liver injury were analysed separately; (ii) variables were analysed using both univariate and multivariate models, including several clinical and demographic variables, to account for confounding; and (iii) phenotypic characterization and follow up was performed independent from rs738409 genotype

status. The major strength of the case/control study design is that it maximises statistical power. In the case of genetic association analysis particularly, this allows the detection of robust genetic associations, which by other study designs, namely cohort studies, could be impractical due to the overall population incidence of alcohol-related cirrhosis and the long follow up period required for it to become apparent.

In summary, this analysis validates the robust association between rs738409 and alcohol-related cirrhosis in the UCL cohort demonstrating the accuracy of imputed genotype data. Through several case/control comparison in this large and well-characterized cohort the association between rs738409 and alcohol-related cirrhosis is demonstrated to be independent from comorbid alcohol-dependence risk. A time-to-event analysis in patients with an without alcohol-related cirrhosis provides intriguing evidence that carriage of the rs738409[G] allele worsens prognosis following clinical presentation with cirrhosis yet improves prognosis in alcohol-misusers with no-significant liver injury. These data indicate that rs738409 is a significant contributor to the overall population attributable risk of alcohol-related cirrhosis in the UK as well as influencing clinical outcomes such as prognosis and the development of HCC.

---

---

**CHAPTER 5 STRUCTURAL STUDIES OF**  
***PNPLA3***

---

---

## 5.1 - BACKGROUND

### 5.1.1 - OVERVIEW

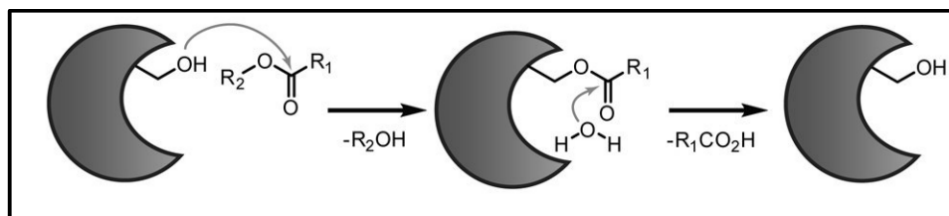
There is growing interest in the biological function of PNPLA3 due to genetic associations between the variant rs738409 and forms of liver disease. This variant results in an isoleucine to methionine amino acid substitution at the position 148 in the protein sequence and has been shown to affect the proteins function. However, the physiological role of PNPLA3 is unknown, as is the specific functional effect associated with the rs738409[G] 148Met allele. Structural biology has the potential to increase understanding of both the role of PNPLA3 and the functional effects of rs738409.

### 5.1.2 - THE PATATIN DOMAIN

*PNPLA3* is transcribed and translated into the enzyme patatin-like phospholipase domain containing 3 (*PNPLA3*)<sup>198</sup>. Other commonly used names for this enzyme include adiponutrin (*ADPN*)<sup>233</sup> and calcium independent phospholipase 2 epsilon (*iPLa2-ε*)<sup>198</sup>. *PNPLA3* is a member of the patatin-like phospholipase domain containing (*PNPLA*) family of proteins. This domains namesake is the non-specific lipid acyl hydrolase, and major constituent of the potato, the patatin glycoprotein<sup>353</sup>. The patatin glycoprotein has been structurally<sup>376</sup> and functionally<sup>178,415</sup> characterized demonstrating that this domain has unique structural features. The amino-acid sequence motifs that characterise this domain are highly conserved and ubiquitous in nature, occurring in over three-thousand species of prokaryotes and over three-hundred species of eukaryotes<sup>21</sup>.

The patatin domain is a member of the serine hydrolase superfamily of proteins. Many enzymes in this superfamily catalyse important biochemical reactions such as proteolysis and thioester bond cleavage<sup>263</sup>. Based on sequence homology information, *PNPLA3* belongs to a sub-group of this family known as the metabolic serine hydrolases. A characteristic feature this sub-group is the  $\alpha/\beta$  hydrolase fold, which juxtaposes catalytic residues forming an enzymatic active site. The active site of this superfamily is characterized by a catalytic serine residue, which functions as a nucleophile resulting from its interaction with other catalytic residues present in the active site; the core residues of serine hydrolases form either catalytic dyads or triads. In patatin domain containing proteins, this catalytic site is characterized by a Ser-Asp catalytic dyad. When performing catalysis the nucleophilic serine attacks acyl moieties, such as the ester moiety of any triacylglycerol, cleaving it into two products: a molecule with a hydroxyl moiety (e.g. glycerol) and an acyl moiety containing molecule such as a fatty acid (e.g. oleic acid) (Figure 5-1).





**Figure 5-1 Mechanism of the nucleophilic serine active-site residue**

The catalytic serine acts as a nucleophile attacking the thioester bond of a substrate molecule. This forms an acyl-enzyme intermediate at the active site serine. This is followed by water-induced saponification of the product, and regeneration of the free serine residue for entry into the next reaction cycle. Image modified from Long et al., 2011<sup>263</sup>.

## THE PNPLA FAMILY IN HUMANS

In humans, there are several proteins that contain patatin domains and these proteins have diverse and essential roles in human physiology (Figure 5-2). However, not all of these proteins have phospholipase activity as suggested by their naming convention, as some members only possess lipase activity. Based on enzymatic activity and sequence homology<sup>236,459</sup> human PNPLA proteins have been categorised into two subtypes namely the lipase type and the phospholipase A type.

All PNPLA proteins maintain the core catalytic residues in the patatin domain but otherwise vary considerably. The lipase type enzymes (PNPLA1, PNPLA2, PNPLA3, PNPLA4 and PNPLA5) mediate a diverse range of functions relating to lipid hydrolysis. The enzyme PNPLA2, which is also known as adipose triglyceride lipase has a prominent functional role as it catalyses a crucial step in triglyceride hydrolysis. It is known to be physiologically important because loss of function mutations in *PNPLA2* result in a form of neutral lipid storage disease with myopathy<sup>125</sup>. The other enzyme sub-family (PNPLA6, PNPLA7, PNPLA8 and PNPLA9) have the capacity to hydrolyse phospholipids into fatty acids and other lipophilic substances. The enzyme PNPLA6, which is also known as neuropathy target esterase, is expressed in the nervous system and exhibits significant phospholipase activity. It is known to be physiologically important because loss of function mutations in *PNPLA6* result in significant neurological deficits and neurodegeneration<sup>429</sup>.

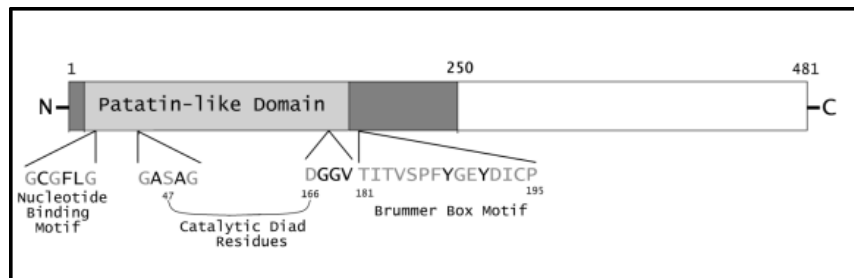


**Figure 5-2 Phylogenetic and structural comparison of human PNPLA family members**

This schematic shows a phylogenetic tree of human PNPLA family members on the left-hand side with the measure of support (0-1 scale) for the tree branching. On the right-hand side, proportional linear models of the protein domain organisation are shown with their length in amino-acids given with the location of the core patatin domain (red box), active site (vertical line) and other regions of the protein (grey) highlighted. Image from Kienesberger et al., 2009<sup>223</sup>

### 5.1.3 - PNPLA3 STRUCTURE

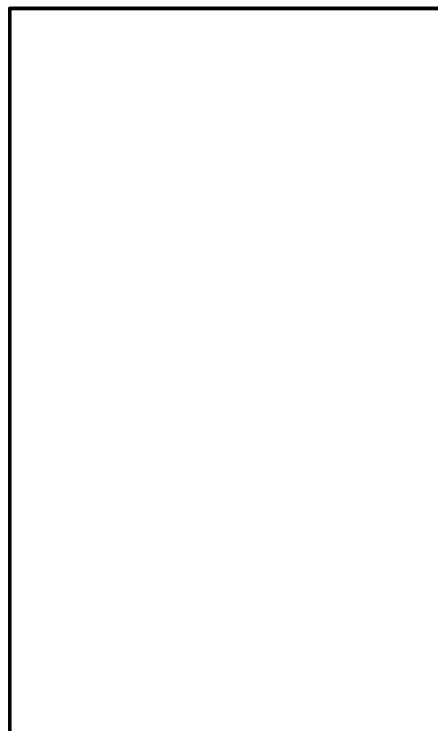
*PNPLA3* is transcribed into five putative splice variants<sup>126</sup> of which two encode sequences that are known to be translated into proteins. The reference protein transcript for *PNPLA3*, encodes a protein containing a sequence of 481 amino acids. The first 250 residues of *PNPLA3* correspond to a patatin-like domain. The motifs that characterise this domain include a catalytic dyad composed of a serine at position 47 and asparagine at position 166. Other notable motifs include a glycine rich nucleotide binding motif (Gly-X-Gly-X-X-Gly)<sup>26</sup> (Figure 5-3). *PNPLA3* also contains a less common Brummer-box motif, which may be essential for targeting the protein to intracellular lipid droplets<sup>56</sup>. The remainder of the primary sequence towards the C-terminus of the protein (residues 250-481) is of unknown function. Studies in ex vivo models have demonstrated that this region contains protein sequence (between residues 320 to 481) which localise *PNPLA3* to lipid droplets<sup>300</sup>.



**Figure 5-3 The domain architecture of PNPLA3**

The region from the N-terminal domain contains the core patatin domain. In this region there are several amino acid motifs including a nucleotide binding motif, the catalytic dyad residue motifs and a Brummer box motif.

There are no experimentally determined three-dimensional structural models of PNPLA3. As a proteins function is implicit to its structure<sup>387</sup> it is likely that significant information could be obtained if this data were available. However, there are structural models in the protein databank (PDB) of other patatin domain containing proteins that have been obtained using the technique of X-ray crystallography<sup>30,376</sup> (Figure 5-4).



**Figure 5-4 The stages of protein structure determination by X-ray crystallography**

The technique of X-ray crystallography is reliant on the formation of crystalline solids containing proteins molecules; these structures are characterized by the highly ordered repetitive arrangement of protein molecules in space. X-rays have a wavelength corresponding to the length of covalent bonds in molecules and thus constructive or deconstructive interference occurs when passed through a substance. In crystalline solids the regular lattice arrangement allows for the production of non-random diffraction patterns. The diffraction pattern produced contains information regarding the atomic structure of a protein molecule; however converting these requires complex algorithmic mathematics guided by expert crystallographers to generate an atomic model. Image obtained from Splettstoesser, 2015<sup>407</sup>

The three dimensional structure of protein horse-leaf nettle patatin (Pat17) enzyme has been used to classify the domain fold of the patatin domain<sup>376</sup>. This structure has also been used to predict the structure of the patatin domain of PNPLA3 using the in silico technique of homology modelling<sup>459</sup>. This predicted model of PNPLA3 contains several putative features of this domain such as its overall  $\alpha/\beta$  hydrolase fold, the positioning of its catalytic dyad residues and a also a 'lid' region (Figure 5-3).

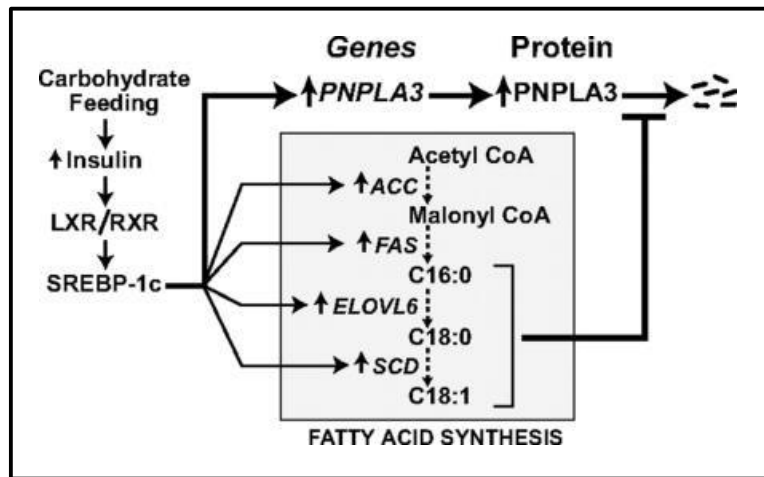
**Figure 5-5 A predicted model of the patatin domain structure in PNPLA3**

The regions predicted to fold as  $\beta$ -sheets and  $\alpha$ -helices are shaded yellow and blue respectively. Side chains of the catalytic aspartate and serine residues are rendered in stick format. The glycine-rich region is indicated by a red arrow and the putative "lid" region by a white arrow. Image from Wilson et al., 2006<sup>459</sup>

#### 5.1.4 - PNPLA3 FUNCTION

##### Nutritional Regulation

In comparison to PNPLA2 and PNPLA6, the physiological role of PNPLA3 is less well understood. The discovery of *PNPLA3* resulted from its regulation by nutritional stimuli<sup>22</sup> as high carbohydrate or protein diets up-regulate its expression and fasting down regulates its expression<sup>198</sup>. In mice<sup>188</sup> and human tissues<sup>103</sup>, *PNPLA3* expression is nutritionally regulated by a feed forward loop (Figure 5-6). *PNPLA3* transcription is observed in in the liver<sup>247</sup> as well as in the retina<sup>336</sup>, skin and adipose tissue<sup>188</sup>. The protein has been localised to these tissues as well and may also be secreted into the blood plasma in minute quantities as a multimeric protein<sup>460</sup>.



**Figure 5-6 A proposed feed forward loop nutritional regulatory mechanism of PNPLA3**  
 Following carbohydrate feeding, the postprandial increase in circulating insulin levels up regulates the liver X receptor (LXR) or the retinoid X receptor (RXR) which consequently activate the transcription factor sterol regulatory element binding protein 1C (SREBP-1C). This transcription factor binds to a sequence upstream of the first exon of PNPLA3 up-regulating gene transcription. It is also a transcription factor for several other genes involved in fatty acid synthesis. This results in an increase in free fatty acids which may also inhibit further PNPLA3 transcription forming a feedback loop. Abbreviations: LXR – Liver X receptor; RXR – Retinoid X receptor; SREBP-1c – Sterol regulatory element binding protein 1c; CoA – Coe-enzyme A; ACC – FAS – Fatty acid synthetase; ELOVL6 - Fatty Acid Elongase 6; SCD - Stearoyl-CoA Desaturase. Image from Huang et al., 2010<sup>188</sup>.

### Enzymatic studies of PNPLA3 function

The nutritional regulation of PNPLA3 and its significant sequence homology with adipose triglyceride lipase (PNPLA2) suggest a functional role for the protein in lipid metabolism. The PNPLA3 protein has been characterized in vitro by purifying the enzyme<sup>233</sup> and testing its activity through candidate approaches assaying synthetic or hydrolytic activity towards selected lipid substrates.

The earliest study to test PNPLA3 activity in vitro purified the recombinant protein from insect cells and demonstrated that it has triacylglycerol hydrolysis activity and the acyl-CoA-independent transacylation activity<sup>198</sup>. Subsequent studies, have demonstrated that it has hydrolytic substrate preference for oleic acid containing glycerolipids with only limited activity against other lipid substrates such as phospholipids, cholesteryl esters, and retinyl esters<sup>187</sup>. Notably, the triacylglycerol synthesizing transacylation activity was not observed in this experiment. Recombinant PNPLA3 has also been heterologously expressed and purified in *E. coli* from which the enzyme was shown to have lipid synthetic activities but was primarily active as an lysophosphatidic acid acyltransferase<sup>233</sup>. The phospholipids synthesised by this reaction are involved in inflammatory processes. PNPLA3 heterologously expressed and purified in yeast<sup>334</sup> from which the purified enzyme demonstrates glycerolipid hydrolysis activity, lysophosphatidic acid acyltransferase activity and retinyl palmitate hydrolase activity<sup>336</sup>. The results from in vitro studies are diverse and sometimes conflicting. The enzymatic

activity of PNPLA3 may involve the synthesis of lipids, primarily through transacylation, and also hydrolytic activity through esterase and thioesterase activity, towards a host of lipid substrates. It remains unknown whether any of these in vitro activities are physiologically relevant.

### Ex vivo studies of PNPLA3 function

PNPLA3 has been investigated in ex vivo models primarily via gene knockout or gene over expression in cultured human cell lines. The first study to characterise loss of *PNPLA3* in a human adipocyte cell-line determined that this had no obvious effects on glycerol or fatty acid release<sup>221</sup>. Studies investigating PNPLA3 function on retinol metabolism in hepatic stellate cells have demonstrated that upregulation of *PNPLA3* results in the extracellular release of retinol<sup>336</sup>. A more holistic analysis of PNPLA3 function in human hepatic stellate cells via protein overexpression followed by mass-spectrometric analysis of cellular lipid content has identified alterations in the ratios of triacylglycerol's that contain saturated or monounsaturated fatty acid moieties<sup>375</sup>. Visualization of PNPLA3 in human cell lines demonstrates that it localises to the surface of lipid droplets<sup>56,300</sup> and to a lesser extent the cell membrane<sup>170</sup>. From cellular models, it is evident that PNPLA3 is likely involved in lipid metabolism as it binds to lipid droplets and its expression levels correlate with alterations in cellular lipid content.

### In vivo studies of PNPLA3 function

The human PNPLA3 protein and its orthologues have been studied in several organisms but most extensively in zebrafish<sup>259</sup>, mice<sup>20</sup> and rats<sup>234</sup> in which gene expression has either been 'knocked down' or 'knocked-out'. In Zebrafish when *PNPLA3* expression is knocked down the effects are moderate and include reductions in hepatic progenitor cell numbers, reduced liver size and alterations in gene expression<sup>259</sup>. This loss of *PNPLA3* does not influence the Zebrafish predisposition to steatosis when exposed to alcohol or to acute liver injury when exposed to acetaminophen. In Rats when *PNPLA3* gene expression is knocked down there are several phenotypic effects<sup>234</sup> including a reduced predisposition to steatosis, reductions in the levels of diacylglycerol and acyl-CoA intermediates and alterations in the ratio of phosphatidic acid to lysophosphatidic acid. In mice when *Pnpla3* is deleted there are no obvious phenotypic effects<sup>20</sup>: the hepatic lipid content of wild-type and knock out mice are equivalent as are rates of cholesterol and fatty acid synthesis and triacylglycerol hydrolysis activity. From in vivo studies it seems evident that PNPLA3 functions in lipid metabolism, however, findings between different organisms are conflicting.

## PNPLA3 ILE148MET

The variant rs738409 leads to a non-synonymous Ile148Met amino acid substitution in PNPLA3. In orthologous PNPLA3 enzymes<sup>370</sup> the isoleucine corresponding to position 148 in the human protein sequence is highly conserved. This sequence conservation implies this residue is involved in protein function. However, it is difficult to reconcile this functional effect when the physiological role of PNPLA3 remains unknown. For this reason, the majority of the functional studies of the Ile148Met substitution have focused on either gains or losses of particular measurable functions in either in vitro, ex vivo or animal models.

### In vitro studies of PNPLA3 Ile148Met

There has been contention regarding the lipid synthetic or hydrolytic activity of PNPLA3. Following from this, in vitro studies have identified both gains and losses of lipid synthesis or lipid hydrolysis resulting from the Ile148Met substitution. In vitro studies have demonstrated that the 148Met variant of PNPLA3 has decreased ability to catalyse triacylglycerol hydrolysis<sup>188</sup>; a decreased ability to catalyse retinol hydrolysis<sup>336</sup>; but, an increased ability to catalyse lysophosphatidic acid acyl transfer<sup>233</sup>.

### Ex vivo studies of PNPLA3 Ile148Met

In cell lines, transfection and over expression of the wild-type enzyme (PNPLA3 148Ile) or the variant enzyme (PNPLA3 148Met) significantly alters the ratios of several lipids suggesting that the substitution results in a loss of triacylglycerol remodelling functionality<sup>375</sup>. In a hepatic stellate cell line, the transfection and overexpression of the 148Met mutant PNPLA3 results in the accumulation of retinol containing lipid droplets<sup>336</sup> also suggesting a loss of function.

### In vivo studies of PNPLA3 Ile148Met

In animal models The knock-out of the murine orthologue *Pnpla3* demonstrate no obvious phenotypic difference in the histology of liver tissue or the levels of several key lipids this suggesting that the complete loss of PNPLA3 is not deleterious per se<sup>20</sup>. However, when PNPLA3 148Met is transiently overexpressed in mice it induces clear histological features of liver disease and lipid deposition. It is likely that this effect results from a loss of enzymatic activity in the patatin domain as chronic overexpression of a catalytically dead PNPLA3 (Ser47Gly) result in a similar phenotype to the 148Met variant<sup>247</sup>. Recently, the 148Met variant was introduced into the murine orthologue *Pnpla3*<sup>400</sup>. Following the consumption of carbohydrate rich diet, *Pnpla3* 148Met mutant mice have increased hepatic lipid accumulation comparison to wild-type counterparts with notable accumulation of the on lipid droplets<sup>400</sup>. From this model it

appears that functional effects of PNPLA3 148Met enzyme requires both the expression of a catalytically inactive PNPLA3 protein and its accumulation on lipid droplets.

### 5.1.5 - PNPLA3 ILE148MET: STRUCTURE AND FUNCTION

A proteins structure is determined by its amino acid sequence<sup>387</sup> and this structure is implicit to biological function. Hence, alterations in the sequence of a protein alter its structure, which may alter function and ultimately influence the pathogenesis of a disease. The variant rs738409 in *PNPLA3* is non-synonymous resulting in an Ile148Met amino acid substitution in PNPLA3.

The structure of PNPLA3 has not been experimentally determined. The structure has been investigated<sup>170,233,470</sup>, however, using homology model predictions derived from the structure of the distantly related isozyme from *S. cardiophyllum* (Pat17)<sup>459</sup> (Figure 5-5). These predicted models of the patatin domain (residues 1-180) of PNPLA3 suggest that the 148Ile residue is sterically close to Ser-Asp catalytic dyad (Figure 5-7) and hence its substitution with 148Met may preclude the access of substrate to the enzymes active site<sup>170</sup>.

Structural modelling has therefore provided a hypothesis as to how the Ile148Met substitution alters PNPLA3 function and ultimately liver disease risk. The strength of this hypothesis, however, is limited by the accuracy of structural modelling. Homology models may provide an accurate overall topology of a protein structure. However, they are rarely if ever accurate at the atomic level. Hence, alterations such as the atomic positioning of residues in the active site of PNPLA3 are likely to contain substantial error. Further, existing models of PNPLA3 only predict the structure of patatin domain and provide no information about the remaining sequence (Figure 5-7). Without experimental validation, it is clear that the structural mechanisms of the effects of the Ile148Met substitution will remain hypothetical.



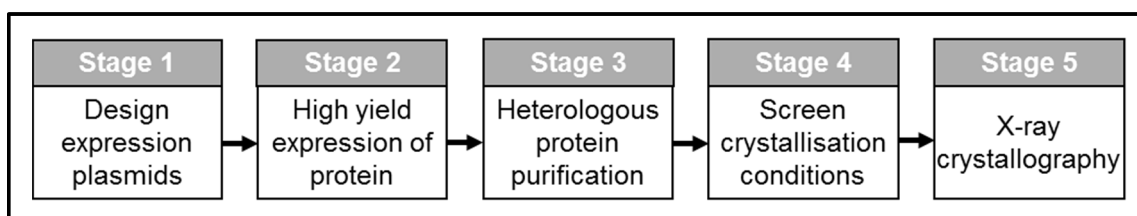


**Figure 5-7 A predicted model of the PNPLA3 active site**

This model demonstrates the hypothetical positioning of the patatin domain catalytic dyad (Ser47 and Asp166) and the position of Ile148 (left panel) and Met148 (right panel), respectively. Protein traces are rainbow-colored from N to C terminus (blue to red). The dots indicate a space-filling model corresponding to van der Waals atomic radii. Oxygen and sulphur atoms are coloured red and yellow. Image from He et al., 2010<sup>170</sup>

## 5.2 - AIMS

The overarching aim of this study was to obtain the crystal structure of PNPLA3 and study the effects of the Ile148Met amino-acid substitution on protein structure and function. The largely empirical process of protein structure determination may be split into several stages which guided the aims of this work (Figure 5-8).



**Figure 5-8 Experimental stages to obtain a protein structure**

The aims of the individual sectional work plan were:

- (i) To perform a computational analysis of the PNPLA3 protein sequence to identify likely features of the protein to aid design of heterologous protein expression system
- (ii) Create recombinant vectors for the expression of the human PNPLA3 protein, and/or its domains in a host organism/cell type
- (iii) Refine a protein expression protocol for large scale purification of PNPLA3
- (iv) Crystallise the protein
- (v) Collect X-ray diffraction data and initialise structure determination.

## 5.3 - PROTEIN SEQUENCE ANALYSIS

### 5.3.1 - OVERVIEW

The purification and crystallisation of proteins is a largely empirical process. There are many experimental strategies to maximise the chances of obtaining a protein in a suitable form for crystallographic analysis. A widely used approach for multi-domain proteins is the 'divide and conquer' strategy whereby different domains of a protein are expressed in a recombinant expression system. This follows the hypothesis that individual domains of a protein are more tightly folded and hence easier to express, purify and subsequently crystallise. Structural modelling and other computational techniques can provide information regarding the folding and likely domain organisation of proteins such as PNPLA3<sup>115</sup>. In this section, the primary protein sequence of PNPLA3 is analysed focussing on the identification of protein domains through computational analysis via sequence alignment, structural modelling and other in silico techniques. The primary aim of these analyses is the identification of protein regions for recombinant protein expression experiments and protein crystallography.

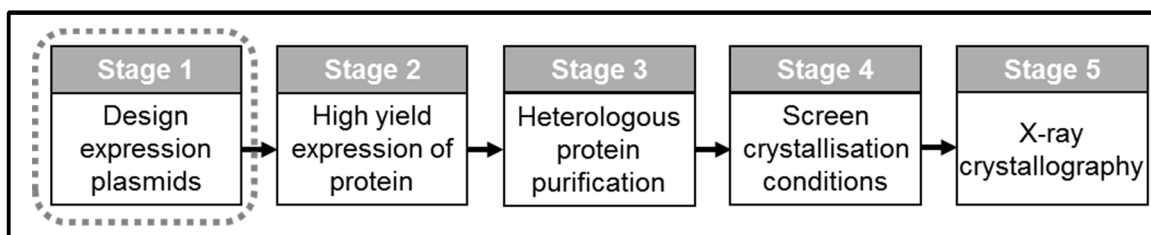


Figure 5-9 Experimental stages to obtain a protein structure

### 5.3.2 - MATERIALS AND METHODS

#### PRIMARY SEQUENCE ANALYSIS

The primary sequence of PNPLA3 that was used in all protein sequence analyses (Supplementary Sequence 1) was obtained from the National Center for Biotechnology Information (NCBI) protein database (Accession: NP\_079501). The expected physico-chemical parameters of the protein such as the molecular weight, the isoelectric point and the grand average hydrophobicity were calculated using the ProtParam tool<sup>135</sup>. The native disorder in the protein structure, was predicted using DISOPRED2<sup>448</sup> which uses an algorithm to compare, and predict intrinsic disorder, based on data from over 750 empirically determined disordered protein sequences.

## SEQUENCE ALIGNMENT

The Basic Local Alignment Search tool (BLAST) is used to interrogate the similarity of amino-acid or nucleotide sequences against reference sequence databases. The BLAST software is used as either standalone package or via a server<sup>424</sup> to interrogate several publicly available nucleotide and protein databases. BLAST was utilised to identify orthologues of PNPLA3, to evaluate PNPLA3 sequence conservation and to identify homologous protein structures from the PDB sequence database.

### Non-Redundant Sequences Database

The non-redundant sequences database is an extensive database of protein sequences containing all non-redundant 69,159,658 sequences from the GenBank, PDB, SwissProt, Protein Information Resource and Protein Research Foundation databases. The PNPLA3 protein sequence underwent a protein-protein BLAST against this database using default parameters. The primary sequences of these homologues were aligned using COBALT<sup>322</sup> and the alignment was visualized using JALVIEW<sup>449</sup>.

## PROTEIN STRUCTURE PREDICTION

There are many methods for predicting protein structure with the most common being homology modelling, threading and ab initio protein structure prediction:

- (i) Homology modelling, which is also known as comparative modelling, is a template-based method that utilises existing protein structures. Using this method the sequence of a protein is compared with the sequences of proteins in a database such as the PDB with an empirically determined structure. Identification of high sequence homology allows the residues of the protein of interest to be spatially overlaid on to the protein structure that has significant sequence homology generating a homology model.
- (ii) Protein threading, which is also known as fold recognition, is another template based method. In comparison to homology modelling this technique is less reliant on high sequence homology instead requiring conserved sequence motifs often found in conserved domains and folds. Typically threading is an iterative approach where the sequence of interest is threaded on to many hundreds of different template structures identified from a domain database search. The accuracy of these template alignments is scored algorithmically typically via estimating the free energy of the threaded structure where the best structures are used to construct a final structural model.

- (iii) Ab initio structure prediction is a template-free method aiming to mimic protein folding that occurs naturally by computational simulation. Protein folding is a complex biophysical process where an unfolded protein chain, which theoretically may take an infinite number of conformations, folds into a well-defined structure all within the millisecond time scale. Thermodynamic processes drive protein folding. Primarily, intramolecular interactions and the torsional limits of a proteins backbone drive the formation of secondary and tertiary structural elements, which consequentially results in an overall folded structure. Hence, a folded structure is thermodynamically stable relative to the unfolded protein chain and thus a folded protein structure have a low free energy. There are two primary methods to simulate protein folding process: molecular dynamics (MD) simulations and Monte-Carlo (MC) energy minimisation simulations<sup>249</sup>. The MD simulation approach involves computing the dynamic movements of all the atoms in a protein and the surrounding solvent over a biologically relevant timescale. The MC energy minimisation approach involves testing multiple conformations of a protein structure and iteratively selecting those structures with the lowest free energy.

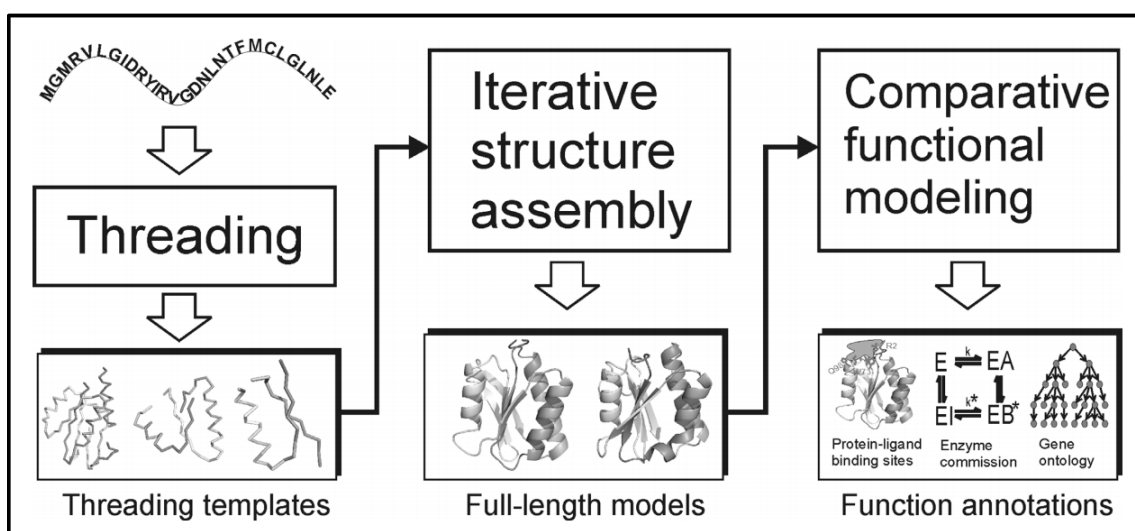
Each structural modelling method has advantages and limitations. Homology modelling may predict the structure of a protein with high accuracy, although this is reliant on the existence of protein structures with high sequence homology (>30% sequence identity). This limits this approach to a minority of proteins or only to certain domains of protein. Protein threading overcomes this issue, partially, requiring only remote sequence homology through the identification of conserved structural folds. Never the less, threading is template dependent and can only be used when a related empirically determined structure exists. The ab initio approach overcomes this limitation through simulations of protein folding from sequence information alone (i.e. template independence). However, this approach is limited by computational complexity; for perspective, the millisecond time-scales required for MD simulations of protein folding would take greater than a year to perform on a standard desktop computer<sup>485</sup>.

### ITASSER

The different structural modelling methods can be used in isolation or can be they can be combined forming a composite approach as is applied by the software ITASSER<sup>331</sup>. The approach of ITASSER involves a pipeline of these different processes split into three stages (Figure 5-10).

- (i) Threading where using the meta-threading program LOMETS<sup>467</sup> structural templates or super-secondary structure motifs are identified from the PDB database. LOMETS uses several different threading algorithms performing thousands of threading alignments; the top ten template fragments are selected for the following stage.
- (ii) Structural assembly where using the top structural fragments generated by threading undergo structural assembly into a full-length structure. Any regions of the structure that are not covered by threading undergo ab initio folding based on MC simulations. The full-length structural model undergoes a second round of MC simulation refinement.
- (iii) Atomic level structural refinement where the lowest free energy full-length structural assemblies from the MC refinement stage undergo full atomic MD simulations generating five top structural models for further analysis.

A full-length structural model prediction of PNPLA3 was generated using ITASSER where its primary sequence as the entry parameter (Supplementary Sequence 1) and other all parameters were set at the default.



**Figure 5-10 A flowchart of the stages of the ITASSER structural modelling pipeline**  
Image from Yang et al, 2015<sup>472</sup>

### Model Accuracy

The accuracy of the predicted by ITASSER structural models cannot be directly determined without an experimentally validated structural model for comparison. Never the less, the accuracy of the structure may be estimated using the confidence score function (C-score). The C-score is calculated during the threading and structural alignment stages of ITASSER structure prediction (Figure 5-10). Its calculation involves the comparison of and its value ranges from -5 to 2 where the higher the score the greater the estimated accuracy of the model.

There are other, more commonly used, measures of protein structural similarity such as the template modelling score (TM-score) and the root mean squared deviation (RMSD). The RMSD is a measure of the average distance between aligned atoms in two superimposed structures in angstroms (Å) where higher values indicates less structural similarity between two structures. However, the RMSD metric is limited by its sensitivity as to two structures can have significant structural overlap but contain small regions of significant structural dissimilarity resulting in low RMSD values. The TM-score is less influenced by small regions of dissimilarity<sup>479</sup>. As with the other metrics, a higher value indicates greater structural alignment (a score >0.5 indicates a model of correct topology and a score <0.17 means a random similarity).

The TM score and RMSD score cannot be directly calculated by ITASSER as such metrics require the comparison with a validated protein structural model. However, for comparability these metrics may be estimated from the C-score using experimentally determined correlation coefficients derived from comparisons between the C-score and RMSD and TM scores of comparisons between ITASSER predicted models and empirically determined structural models<sup>479</sup>.

### Comparison with Previous Modelling

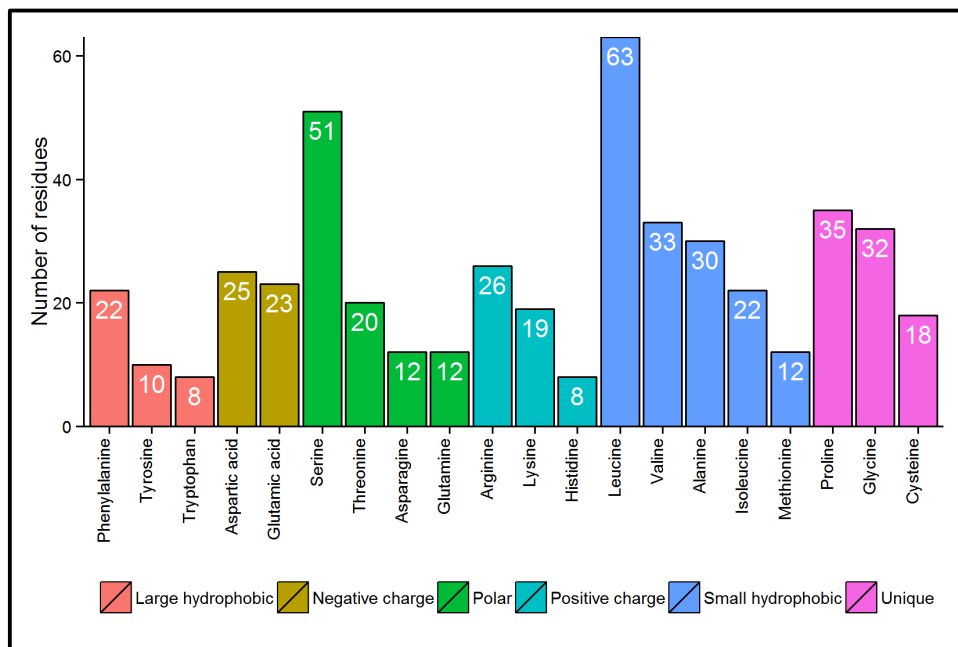
A homology model prediction based on the patatin structure 1OXW has been used in previous studies of PNPLA3<sup>170,470</sup>. This structure was downloaded from SwissProt<sup>19</sup>. Comparison between this, and the ITASSER derived structural prediction was performed using the software TM-align<sup>480</sup>. Structural model predictions were visualized and analysed using the open source educational PyMol software<sup>383</sup>.

## **5.3.1 - RESULTS**

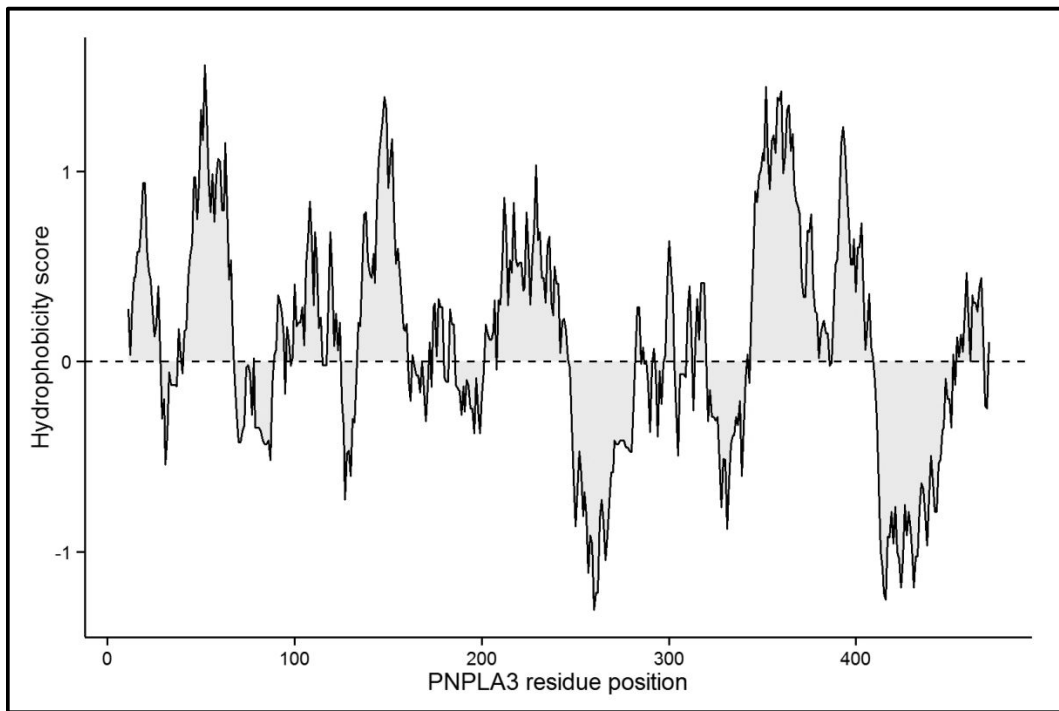
### **PRIMARY SEQUENCE ANALYSIS**

PNPLA3 has a predicted isoelectric point of 6.27 and thus at physiological pH it is likely to have a net negative charge. The acid content of the protein is distributed amongst all of the residue-types with leucine and serine being the most common (Figure 5-11). It contains several cysteine residues and therefore may have a potential to form disulphide bridges if exposed to non-reducing environments such as the lumen of the rough endoplasmic reticulum, the intermembrane space in mitochondria. The proteins grand average hydrophobicity<sup>235</sup> is low (0.10) suggesting that that it is likely to be relatively soluble in water. At a residue, level the hydrophobicity fluctuates along the length of the protein chain and there are two short regions of the sequence that contain predominantly hydrophilic residues (residues 247-297 and 410-451) (Figure 5-12). Analysis of the primary sequence of PNPLA3 for regions of intrinsic disorder suggests

two regions of predicted intrinsic disorder (Figure 5-13) which correspond to the regions with greater hydrophilicity (residues 265-298 and 401-481). Native structural disorder is a feature of many proteins and is frequently found in the linker region joining two tightly folded domains of a protein where such regions often have low sequence complexity.

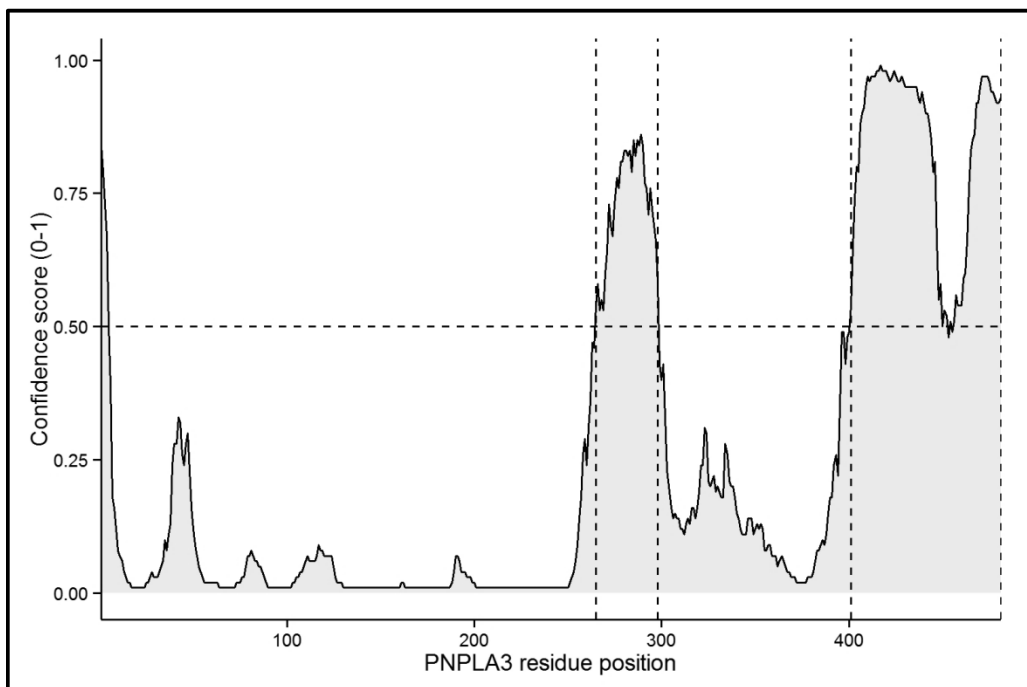


**Figure 5-11 The residue content of PNPLA3**  
Residues are grouped and coloured by their biochemical properties



**Figure 5-12 Plot of residue hydrophobicity along the sequence of PNPLA3**

The average hydrophobicity is calculated on a sliding scale for every 20 residues where a value below 0 suggests a region of hydrophobicity. Plot created using the method of Kyte and Doolittle, 1982<sup>235</sup>



**Figure 5-13 The predicted intrinsic disorder along the sequence of PNPLA3**

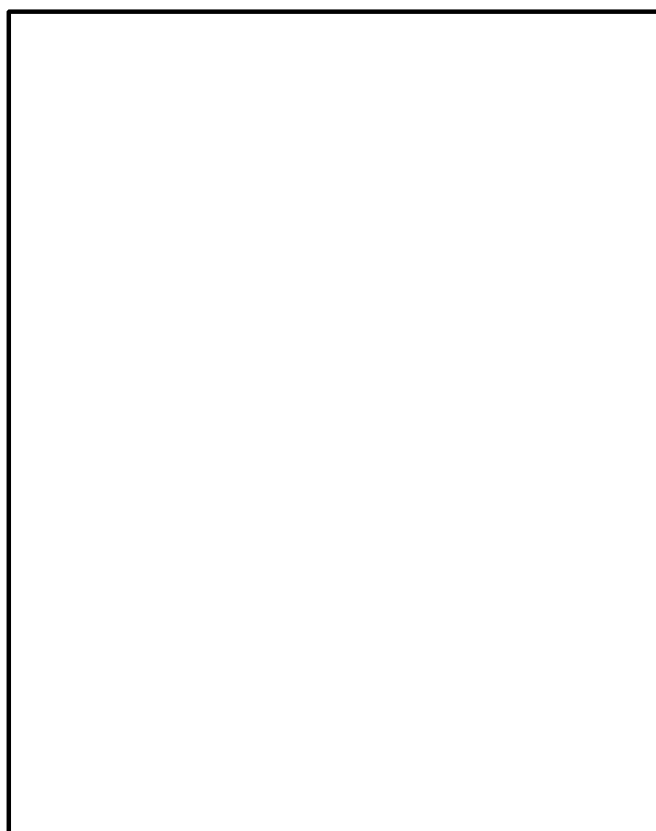
The plot shows position in the sequence against probability of disorder. Regions above the threshold confidence score (0.5) are classified as intrinsically disordered. Plot calculated using DISOPRED<sup>448</sup>



## SEQUENCE ALIGNMENT

There are 1161 protein sequences from 281 different organisms in the with high sequence homology to PNPLA3 in the non-redundant protein database (E-value <  $1 \times 10^{-30}$ ). The majority of these are likely to be orthologues or paralogues of PNPLA3 from other species. The sequences with the highest similarity to the human PNPLA3 occur in mammalian species from both placental and non-placental lineages. Of the placental mammals, which consist the majority of mammalian species there are homologous protein sequences in primates (21 species), carnivores (5 species), hoofed animals (16 species) and rodents (13 species).

A subset of these mammalian protein sequences underwent multiple sequence alignment with the human PNPLA3 sequence (Figure 5-14). Between the human and mammalian orthologues there are two regions of almost complete sequence conservations (residues 1-250 and 300–400). The regions with lower sequence conservation correspond to those regions that are predicted to be intrinsically disordered (Figure 5-13). The extreme C-terminal region of human PNPLA3 (residues 400-480) only occurs in the primate lineages in this comparison.



**Figure 5-14 Multispecies alignment of mammalian PNPLA3 orthologues**

Sequence alignments are coloured using blosum matrix colouring where residue positions of conserved homology are a more intense shade of blue.

## STRUCTURE PREDICTION

### Structure Prediction

The software ITASSER identified several structural models from the PDB sequence database using a position-specific iterative BLAST (Table 5-1). The template structures used for threading were 4AKF (VipD), 1OXW (Patatin), 3TU3, 4QMK and 4AKX (both ExoU).

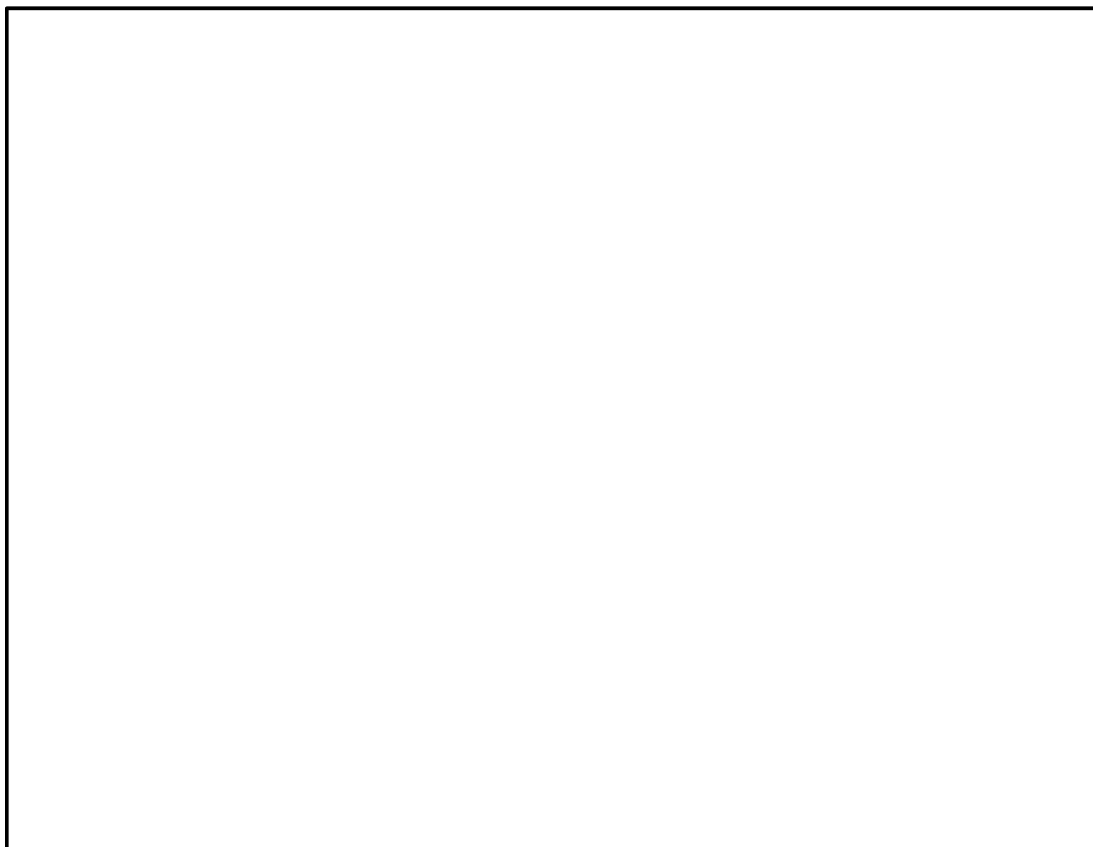
Table 5-1 The top template structures used in PNPLA3 structural model building

Rank	PDB Structure Templates	Identity	Normalised Z-score
1	4QMKA	17%	0.89
2	4AKFA	13%	1.92
3	4AKFA	13%	1.09
4	1OXWA	18%	5.87
5	4AKFA	17%	4.53
6	4QMKA	14%	1.61
7	4AKFA	17%	5.64
8	3TU3B	18%	1.72
9	4AKXB	14%	0.99
10	4AKFA	14%	1.34

The top ten threading templates used during ITASSER model building identified via the LOMETS meta-server which generates thousands of template alignments using ten different protein threading algorithms. One template of the highest Z-score is selected from each threading algorithm used on the LOMETS server. Top-alignments are scored by the Normalised Z-score where a score greater than 1 indicates a good alignment and vice versa

The level of sequence homology between human PNPLA3 and the template structures identified by ITASSER was low (Figure 5-15). The core regions of sequence homology between these sequences are those motifs, which characterise the patatin domain. These include a glycine-rich, nucleotide binding, motif (Gly-X-Gly-X-X-Gly; residues 14, 16 and 19), a serine hydrolase catalytic dyad consisting of a catalytic serine motif (Gly-X-Ser-X-Gly; residues 45, 47 and 49) and a catalytic aspartate motif (Asp-X-Gly/Ala; residues 166 and 168). There are several other conserved regions in the protein sequences including a patch of several small hydrophobic residues (residues 80-90) and two hydroxyl group containing residues (residues 145-149). The residue at

position 149 is a proline and it is conserved between all of the protein sequences. This proline is notable as it occurs a single residue along from isoleucine 148, which is the site of the Ile148Met substitution.



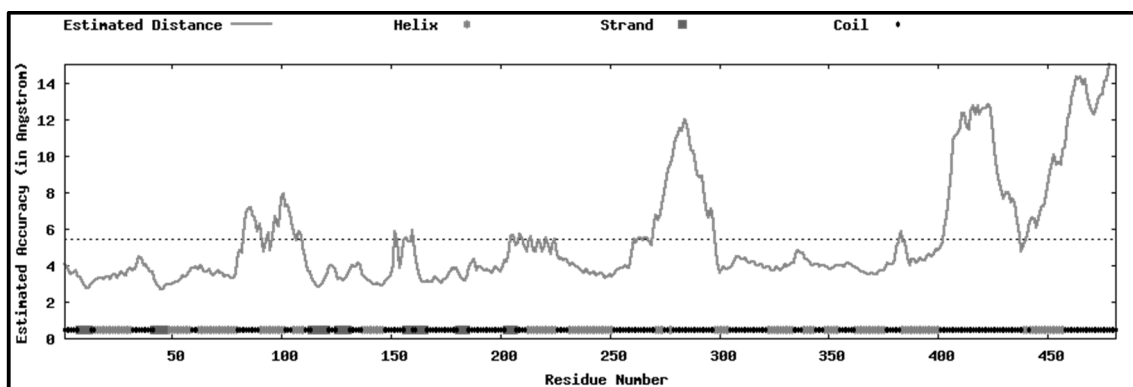
**Figure 5-15 Top ITASSER threading template sequence alignments**

Sequence alignment are coloured using blosum matrix colouring where residue positions of conserved homology are a more intense shade of blue.

### Accuracy of the Structure Prediction

Five full-length structural predictions of PNPLA3 were created using ITASSER. These had C-scores ranging from -2.18 to -3.66. The most-accurate of these had a C-score of -2.18, an estimated TM-score of  $0.46 \pm 0.15$  and an estimated RMSD of  $12.5 \pm 4.3$  Å. In the most accurate prediction, the level of estimated error differs along the sequence of amino acids (Figure 5-13). Many of the regions have a low estimated structural accuracy correspond to regions of predicted intrinsic disorder and low sequence conservation in mammalian orthologues of PNPLA3.

The low C-score, high estimated RMSD and estimated TM-score values of this model suggest that it is of low accuracy and it is therefore unsuitable for confidently inferring the atomic level position of the Ile148Met residue in reference to the active site residues. However, it does reveal a potential domain organisation of the entire protein comprising an N-terminal patatin domain joined to an uncharacterized C-terminal domain surrounded by intrinsically disordered linker regions.



**Figure 5-16 The estimated local accuracy of the PNPLA3 structural model**

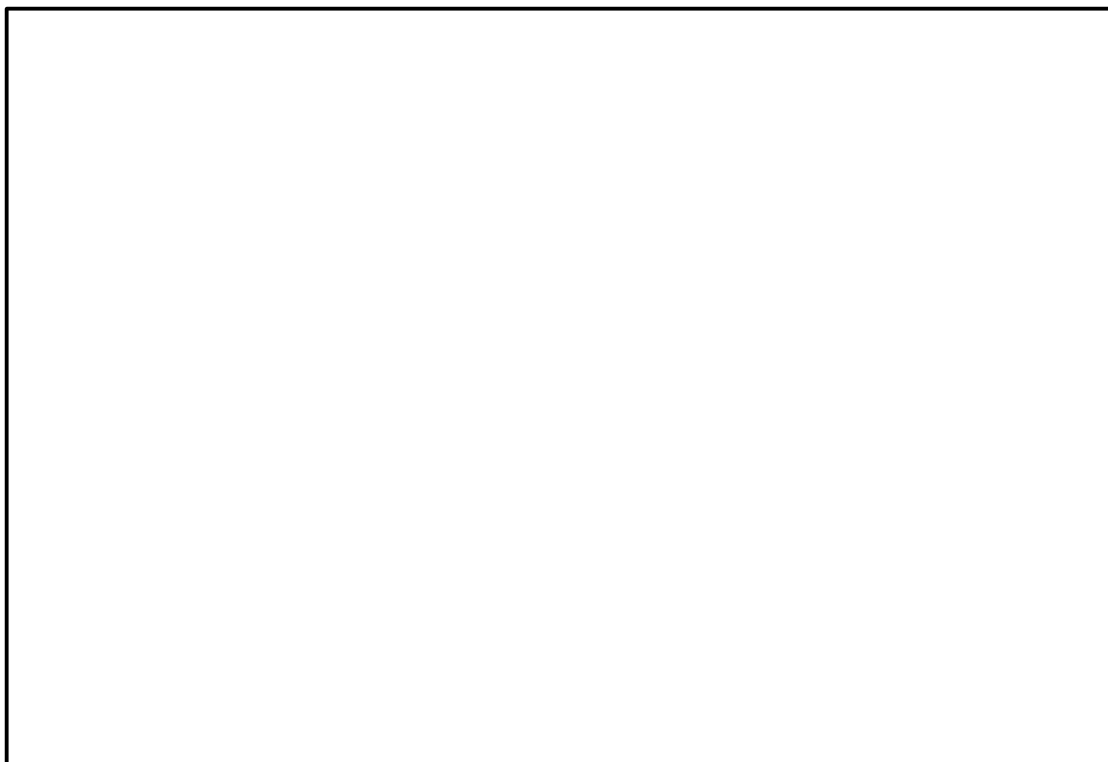
The predicted secondary structure of the model is shown along the X-axis by residue positions

### Three-dimensional Structure

Visual analysis of the three-dimensional coordinates of the predicted structural mode of PNPLA3 (Figure 5-16) demonstrates its putative domain organisation. The first N-terminal domain corresponds to the predicted patatin-like domain of PNPLA3 and contains the characteristic structural features of this domain such as the  $\alpha/\beta$  serine hydrolase fold and the close proximity of the catalytic residues, Ser47 and Asp166 (Figure 5-18). The residue Ile148, which is the site of a natural variant associated with alcohol-related cirrhosis, is sterically close to these active site-residues. The remainder of the structure contains a separate domain-like structure composed of several  $\alpha$ -helices. Two  $\alpha$ -helices from the first half of the predicted structures sequence (residues 180-250) which correspond to a functionally characterized Brummer-box motif<sup>56</sup> extend into the putative C-terminal domain of the model. The putative linker region (residues 250-320) has no secondary structure in this predicted structure.

### Comparison with a Previous Predicted Structure

A predicted structure of PNPLA3 has been analysed previously<sup>480</sup>. This predicted structure is based on a homology model built around the plant patatin structure 1OXW. There is significant structural alignment between this structure and the full-length ITASSER model. When aligned both models are structurally similar between the first 170 residues of PNPLA3 with a RMSD of 2.99Å and TM score of 0.77 (Figure 5-17 (right image)). These two predicted models are therefore structurally homologous. However, as they both use the same reference structure, albeit using different techniques, they are non-independent.



**Figure 5-17 The top structural model of full length PNPLA3**

The left image shows a cartoon representation with rainbow colouring from the C to N terminal of PNPLA3 overlaid with a transparent model of the protein surface. The right image shows a backbone of PNPLA3 (green) showing its alignment with the SwissProt structural model of PNPLA3 (cyan)



**Figure 5-18 The patatin domain of the top model**

Active site residues are highlighted in the patatin-like domain of the structural model highlighting the predicted close special proximity of the catalytic residues to the position of isoleucine 148.

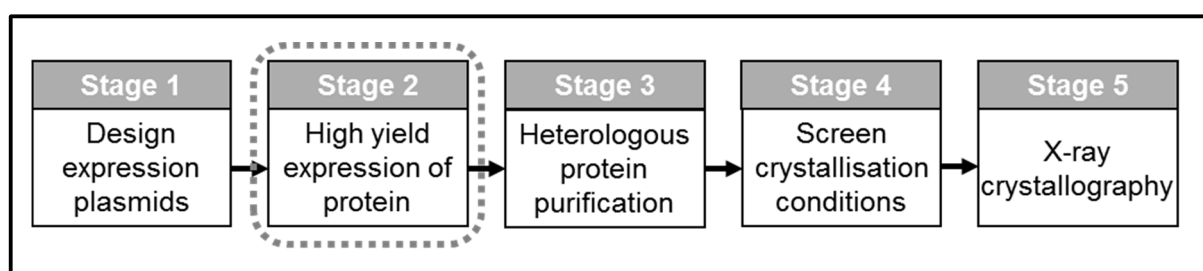
## 5.3.2 - SUMMARY

The computational analysis of the primary sequence of PNPLA3, using several independent bioinformatics tools, indicates that the protein structure of PNPLA3 may be composed of two domains joined by a disordered linker sequence. There are two regions of high sequence conservation between mammalian orthologues. The regions of low sequence conservation correlate with regions that are predicted to have high a probability of high intrinsic disorder and high hydrophobicity. A predicted structure of PNPLA3 highlights this putative domain organisation. The low estimated accuracy of this model means that inferences on structure and function are limited. These domain boundaries guided the design of recombinant plasmids for heterologous protein expression trials.

## 5.4 - PROTEIN EXPRESSION VECTORS

### 5.4.1 - OVERVIEW

The computational analyses of the PNPLA3 protein sequence guided the design of plasmids for the expression of PNPLA3 that were created and tested at the Oxford Protein Production Facility (OPPF)<sup>440</sup> (Figure 5-19). The OPPF is set up as a structural proteomics facility to enable the development of plasmids for the heterologous expression of protein for structural biology purposes. The protocols<sup>31,36,37</sup> employed and developed at the OPPF, focus on high throughput cloning and protein expression techniques with the aim of maximizing success through creating and testing a number of plasmids in parallel.



**Figure 5-19 Experimental stages to obtain a protein structure**

At the second stage the primary aim is to develop a method for obtaining high yields of the protein of interest, typically using recombinant DNA to express the gene of interest in a host protein expression system

## 5.4.2 - MATERIALS AND METHODS

### TEMPLATE DNA

A template cDNA sequence of the human *PNPLA3* gene was purchased (SourceBioscience, Nottingham, UK) (Supplementary Sequence 2). This template contains the rs738409[G] allele and hence codes for the 148Met variant of the enzyme.

### POPIN SUITE VECTORS

The pOPIN vectors, which are used at the OPPF are derived from<sup>31</sup> the commercial pTriEx2 plasmid (Novagen, Nottingham, UK)<sup>31</sup>. There are several pOPIN vector types, which largely differ in nucleotide sequence around their translated region. They all share core features for recombinant protein expression such as the nucleotide sequence encoding a polyhistidine tag and a 3C protease consensus cleavage and the ability to express recombinant protein in multiple hosts (Figure 5-20).

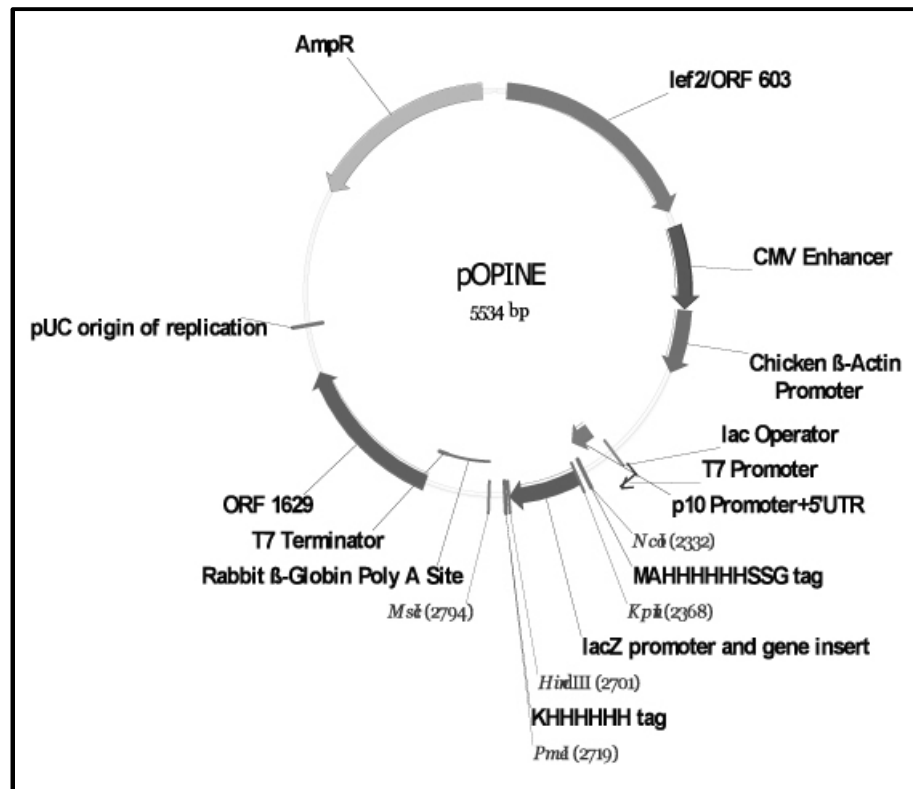
Table 5-2 pOPIN suite vectors chosen for heterologous PNPLA3 expression

Backbone	Fusion tag	Restriction site for linearisation	Forward primer extension	Reverse primer extension
pOPINE	POI-3C-HIS	NcoI, PmeI	AGGAGATATAACCATG	GTGATGGTGATGTTT
pOPINE-3C-eGFP	POI-3C- eGFP-3C-HIS	NcoI, PmeI	AGGAGATATAACCATG	CAGAACTTCCAGTTT
pOPINE-3C-HALO7	POI-3C- Halo7-HIS	NcoI, PmeI	AGGAGATATAACCATG	CAGAACTTCCAGTTT
pOPINM	HIS-MBP-3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA
pOPINO	SS-POI-HIS	KpnI, PmeI	CTACCGTAGCGCAAGCT	GTGATGGTGATGTTT
pOPINS3C	HIS-SUMO- 3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA
pPOPINTRX	HIS-TRX- 3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA
pOPINF	HIS-3C-POI	KpnI, HindIII	AAGTTCTGTTTCAGGGCCCG	ATGGTCTAGAAAGCTTTA

Abbreviations: POI – Protein of interest; HIS – Polyhistidine tag; MBP – Maltose binding protein; eGFP – enhanced green fluorescent protein; TRX - thioredoxin reductase; SUMO - small ubiquitin-like modifier; 3C - Rhinovirus 3C protease site; SS – signal sequence

Eight pOPIN vector types were selected for recombinant protein expression trials of PNPLA3 (Table 5-2). The pOPINE and pOPINF vector types were the simplest vector backbone selected, as these allow the expression of a protein with either an N- or C-terminal polyhistidine tag. The others plasmids (pOPINE-3C-eGFP, pOPINE-3C-HALO7, pOPINM, pOPINO, pOPINS3C and pPOPINTRX) were selected a priori for their unique features. These features include nucleotide sequence which allows the co-expression of an N or C-terminal fusion protein tag protein such as maltose binding protein (MBP), green fluorescent protein (GFP), the small ubiquitin like modifier

(SUMO) protein or thioredoxin reductase (TXR). The inclusion of fusion protein tags may increase protein solubility, aid protein folding, and allow additional stages of protein purification. The pOPINO vector differed from the other vectors as it produces a fusion protein containing a signal sequence that secretes the protein into the periplasmic space of *E. coli*. The periplasmic space is a reducing environment and may aid the folding of proteins that contain native disulphide bonds.



**Figure 5-20 The backbone of a pOPINE plasmid**

Core features of pOPIN plasmids include: pUC origins of replication for high-copy replication in *E. coli*; the ampicillin resistance gene (AmpR) for positive selection in ampicillin containing media; T7 promoter, transcription terminator and a lac operator sequence for high-level transcription in *E. coli*; a CMV enhancer, Chicken  $\beta$ -actin promoter and  $\beta$ -globin polyA signal sequences for transcription in HEK293T; recombination sites for integration into the baculovirus genome for expression in insect cells; the p10 promoter/5'UTR for protein expression in insect cell lines; and, a poly Histidine tag adjoining a 3C protease cleavage site for affinity purification.

## CONSTRUCT DESIGN

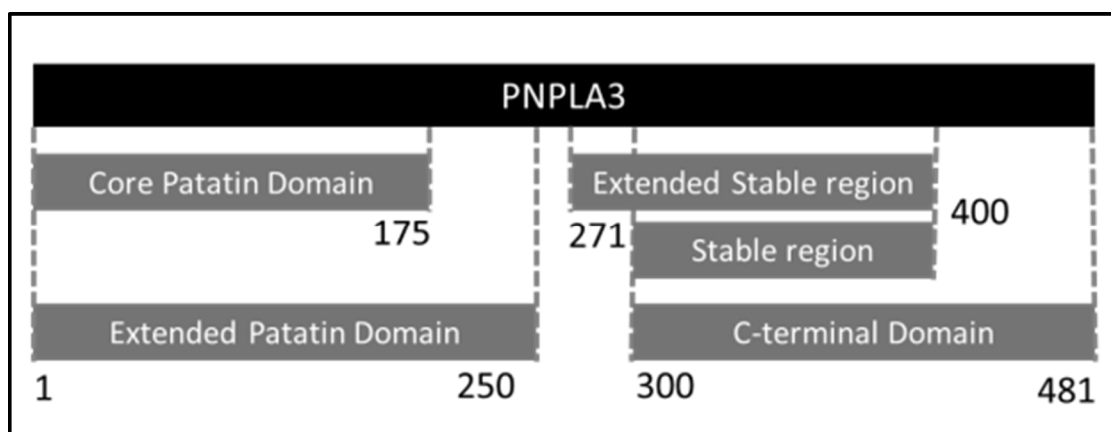
The design of PNPLA3 expression plasmids was guided by three parameters:

- The 96-well plate layout employed for high-throughput cloning at the OPPF
- The selection of pOPIN suite vector backbones available at the OPPF
- The selection of PNPLA3 inserts guided by computational analysis of domain boundaries

The high-throughput cloning protocol employed at the OPPF utilises a 96-well plate on which every cloning reaction is performed in duplicate. This limits the design to 48



unique plasmids. A total of 8 pOPIN vector backbone types were selected for plasmid design (Table 5-2) therefore allowing the trial of 6 different *PNPLA3* inserts. The selection of these *PNPLA3* inserts were made corresponding to putative domains of *PNPLA3* identified by computational analysis (Figure 5-21). Hence, plasmids were designed to express the full-length protein, the N-terminal patatin-like domain or the putative C-terminal domain. These include the entire gene (from residues 2-481); two insert fragment containing the patatin-like domain (from residues 2-175 and 2-250); and, three insert fragments containing the predicted C-terminal domain (from residues 271-400, 300-400 and 300-481) (Figure 5-21).

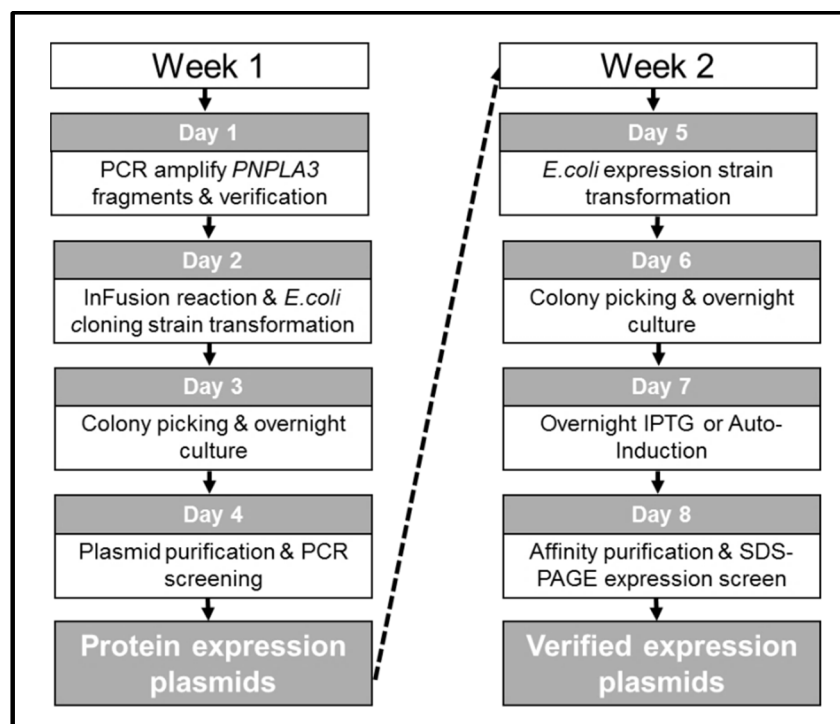


**Figure 5-21 PNPLA3 inserts selected for plasmid construction**

The entire *PNPLA3* protein and five regions of it were selected for recombinant protein expression trials. The same *PNPLA3* cDNA nucleotide sequence underwent PCR amplification to amplify template nucleotide sequences for insertion into the backbone of pOPIN plasmids.

## PLASMID CONSTRUCTION

The construction of plasmids was performed using high-throughput techniques (Figure 5-22). During the first stage different fragment of *PNPLA3* were inserted into the different pOPIN plasmid vectors creating constructs. In the second stage, the constructs were tested for their efficacy in recombinant protein expression in *E. coli*.



**Figure 5-22 A schematic of the stages of high-throughput plasmid construction**

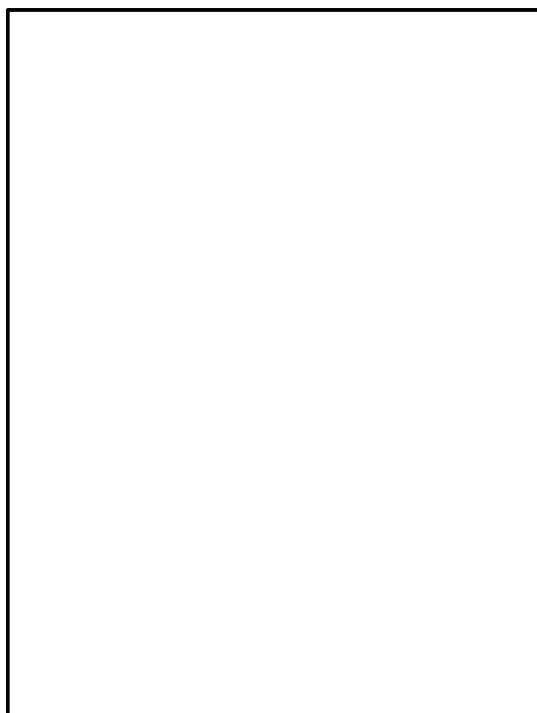
High-throughput plasmid construction and validation was performed at the OPPF over a two week period. During the first week, plasmids are constructed via the insertion of regions of the into pOPIN plasmid. During the second week the plasmids undergo protein expression trials in *E.coli* to explore whether they are suitable for recombinant protein expression.

The *PNPLA3* template fragments were inserted into the pOPIN vector suite plasmids using In-Fusion cloning<sup>36</sup>. This cloning method is DNA ligase-independent cloning technology, which requires fewer stages than traditional cloning and is therefore suited for high-throughput methods. The experimental process may be split into several stages (Figure 5-23). This starts with PCR amplification of the template DNA (Table 5-3) using primers that have vector specific 15 base pair overhangs on the forward and reverse primer sequences (Table 5-2). The success of PCR is determined via agarose gel electrophoresis and comparison of the molecular weights of the PCR amplicons.

Table 5-3 The PCR amplification conditions used on the *PNPLA3* insert

Number of Cycles	Condition
1X	94°C 2 minutes
29X	98°C 10 seconds
	60°C 30 seconds
	68°C x 1min/kilo base-pair length of template
1X	68°C 2 minutes
1X	4°C Hold

Subsequently, each amplicons is incubated with a restriction endonuclease linearized pOPIN plasmid backbone and the InFusion enzyme. The 96-well plate layout is maintained and the several different pOPIN plasmid vectors are added to their pre-determined positions on the plate (Figure 5-26). The InFusion enzyme has proof-reading exonuclease activity and generates cohesive 15 base pair complementary single stranded ends of both the amplicon and the linearized pOPIN vector. This allows the cohesive strands of vector and template DNA anneal forming circular yet nicked plasmids. These nicked plasmids were used to transform *E. coli*, which were grown overnight on antibiotic selective LB-Agar covered with the galactose analogue X-gal. During this period, transformants repair and replicate the nicked plasmid. Transformants containing the inserted plasmid were selected by blue-white screening and were grown overnight in antibiotic selective media. Finally, the construct plasmid DNA is purified from *E. coli* by performing an alkaline lysis based miniprep. The experimental methods used are given in greater detail in the published OPPF protocol<sup>37</sup>.



**Figure 5-23 Stages of InFusion cloning**

The stages of InFusion cloning: 1 – A template vector is linearised with a restriction enzyme generating overhanging DNA sequences; 2 – PCR amplification of template DNA generating in-frame insert DNA for protein/fragment with vector complementary overhanging DNA sequence; 3 – Incubation of linearized vector with InFusion enzyme generating a circular plasmid constructs; 4 – Spin column purification of plasmid constructs from InFusion assay mix; 5 – Transformation of competent *E. coli* with constructs; 6 – Blue/white colony screening of *E. coli* for positive transformation in antibiotic selective agar

Table 5-4 PNPLA3 expression constructs created at the Oxford Protein Production Facility

Position		PNPLA3		Backbone	Fusion Tag	Expected molecular weight (kDa)
1 <sup>st</sup> replicate	2 <sup>nd</sup> replicate	First residue	Last residue			
A1	A7	2	481	pOPINE	C-His	52.8
B1	B7	2	175	pOPINE	C-His	19.14
C1	C7	2	250	pOPINE	C-His	27.39
D1	D7	300	481	pOPINE	C-His	20.01
E1	E7	300	400	pOPINE	C-His	11.11
F1	F7	271	400	pOPINE	C-His	14.3
G1	G7	2	481	pOPINE-3C-eGFP	POI-3C-eGFP-KHIS6	81.8
H1	H7	2	175	pOPINE-3C-eGFP	POI-3C-eGFP-KHIS6	48.14
A2	A8	2	250	pOPINE-3C-eGFP	POI-3C-eGFP-KHIS6	56.39
B2	B8	300	481	pOPINE-3C-eGFP	POI-3C-eGFP-KHIS6	49.02
C2	C8	300	400	pOPINE-3C-eGFP	POI-3C-eGFP-KHIS6	40.11
D2	D8	271	400	pOPINE-3C-eGFP	POI-3C-eGFP-KHIS6	43.3
E2	E8	2	481	pOPINE-3C-HALO7	POI-3C-HALO7-KHIS6	87.8
F2	F8	2	175	pOPINE-3C-HALO7	POI-3C-HALO7-KHIS6	54.14
G2	G8	2	250	pOPINE-3C-HALO7	POI-3C-HALO7-KHIS6	62.39
H2	H8	300	481	pOPINE-3C-HALO7	POI-3C-HALO7-KHIS6	55.02
A3	A9	300	400	pOPINE-3C-HALO7	POI-3C-HALO7-KHIS6	46.11
B3	B9	271	400	pOPINE-3C-HALO7	POI-3C-HALO7-KHIS6	49.3
C3	C9	1	481	pOPINF	N-His	52.91
D3	D9	1	175	pOPINF	N-His	19.25
E3	E9	1	250	pOPINF	N-His	27.5
F3	F9	300	481	pOPINF	N-His	20.02
G3	G9	300	400	pOPINF	N-His	11.11
H3	H9	271	400	pOPINF	N-His	14.3
A4	A10	1	481	pOPINM	HIS6-MBP-3C-POI	97.91
B4	B10	1	175	pOPINM	HIS6-MBP-3C-POI	64.25
C4	C10	1	250	pOPINM	HIS6-MBP-3C-POI	72.5
D4	D10	300	481	pOPINM	HIS6-MBP-3C-POI	65.02
E4	E10	300	400	pOPINM	HIS6-MBP-3C-POI	56.11
F4	F10	271	400	pOPINM	HIS6-MBP-3C-POI	59.3
G4	G10	2	481	pOPINO	SS[OmpA]-POI-KHIS6	52.8
H4	H10	2	175	pOPINO	SS[OmpA]-POI-KHIS6	19.14
A5	A11	2	250	pOPINO	SS[OmpA]-POI-KHIS6	27.39
B5	B11	300	481	pOPINO	SS[OmpA]-POI-KHIS6	20.02
C5	C11	300	400	pOPINO	SS[OmpA]-POI-KHIS6	11.11
D5	D11	271	400	pOPINO	SS[OmpA]-POI-KHIS6	14.3
E5*	E11*	1	481	pOPINM	HIS6-MBP-3C-POI	97.91
F5	F11	1	175	pOPINS3C	N-His SUMO-3C-POI	19.25
G5	G11	1	250	pOPINS3C	N-His SUMO-3C-POI	27.5
H5	H11	300	481	pOPINS3C	N-His SUMO-3C-POI	20.02
A6	A12	300	400	pOPINS3C	N-His SUMO-3C-POI	11.11
B6	B12	271	400	pOPINS3C	N-His SUMO-3C-POI	14.3
C6	C12	1	481	pPOPINTRX	HIS6-TRX-3C-POI	52.91
D6	D12	1	175	pPOPINTRX	HIS6-TRX-3C-POI	19.25
E6	E12	1	250	pPOPINTRX	HIS6-TRX-3C-POI	27.5
F6	F12	300	481	pPOPINTRX	HIS6-TRX-3C-POI	20.02
G6	G12	300	400	pPOPINTRX	HIS6-TRX-3C-POI	11.11
H6	H12	271	400	pPOPINTRX	HIS6-TRX-3C-POI	14.3

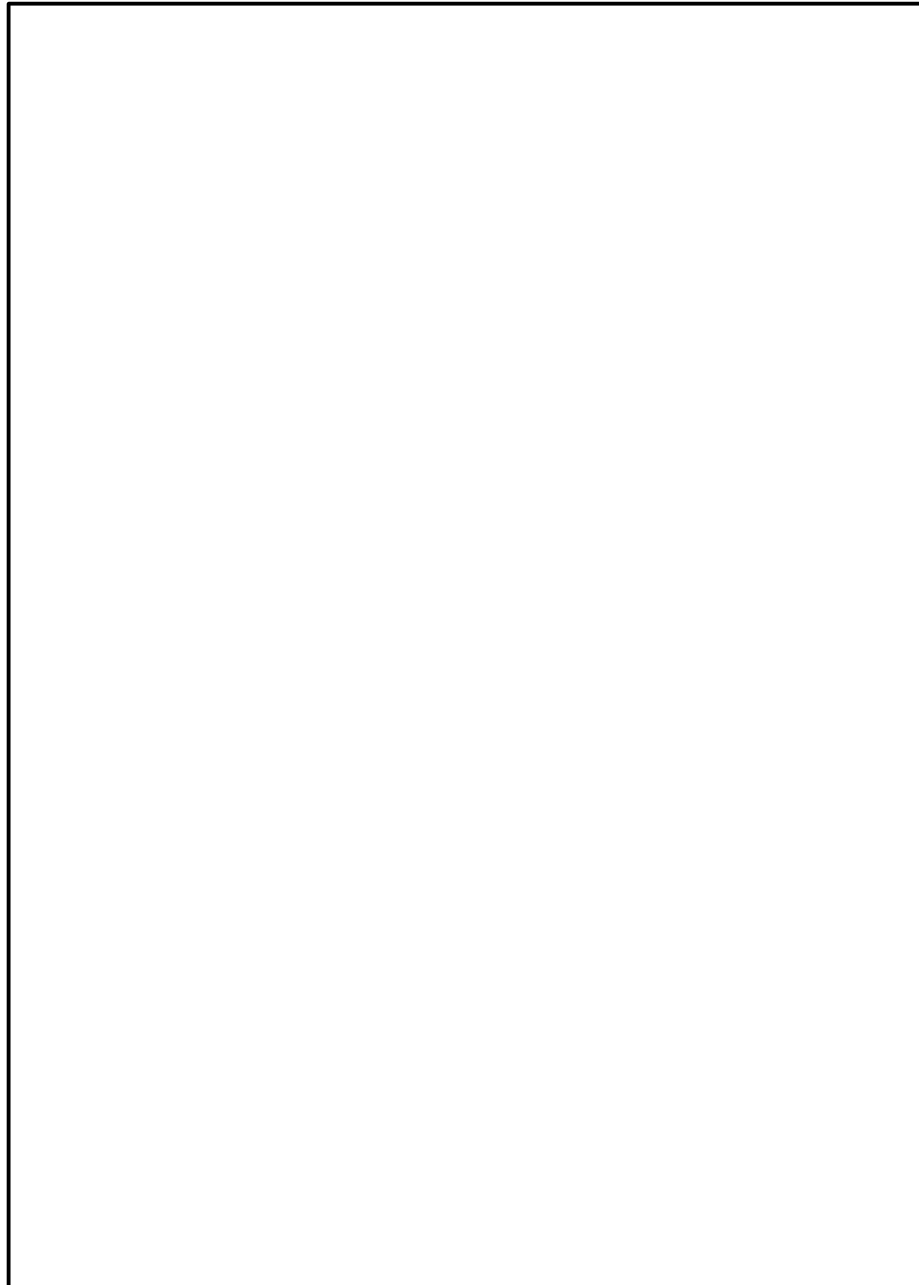
\*Sample E5 contained a pOPINM vector backbone rather than pOPINS3C due to an error during the primer design stage.

Abbreviations: C – carboxyl terminus; GFP – green fluorescence protein; HIS – histidine; MBP – maltose binding protein; POI – protein of interest; SUMO - Small Ubiquitin-like Modifier; TRX - thioredoxin

## ***E. COLI* EXPRESSION TRIALS**

High throughput protein expression trials in *E. coli* were performed at the OPPF<sup>37</sup> (Figure 5-24). Four different expression conditions were tested for every construct using two different *E. coli* strains and two different induction conditions. The two *E. coli* strains, Lemo21(DE3) and Rosetta(DE3) were transformed with the pOPIN-PNPLA3 constructs via heat-shock. Both of these *E. coli* strains have been optimised for recombinant protein expression. The Lemo21(DE3) strain contains an additional plasmid which allows for the tightly controlled induction of recombinant protein induction<sup>382</sup>. The Rosetta(DE3) strain contains an additional pRARE plasmid that expresses human tRNA genes, which reduces codon usage bias when inducing proteins that are non-native to *E. coli*<sup>311</sup>. The transformed *E. coli* were streaked on to the individual wells of antibiotic containing LB-agar 6-well plates, containing X-gal and incubated overnight.

Positive transformants were selected from the overnight plates via blue/white screening. The transformants were grown overnight in antibiotic containing LB in 96-well deep-well blocks. These cultures were used to inoculate 24-well deep well plates containing either terrific broth or auto-induction-media for overnight auto-induction or IPTG induction respectively. The growth of *E. coli* was monitored via visual inspection of LB transparency and when at a sufficient cell density IPTG was added to each well of the cultures grown in terrific broth. Both auto-induction and IPTG induction cultures of both Lemo21 and Rosetta were grown overnight at a reduced temperature (20°C). On the final day, the *E. coli* were harvested from these cultures via centrifugation. Cell pellets were lysed by freeze-thaw and resuspension in lysis buffer (supplemented with 1mg/mL Lysozyme and either 3 units/mL of Benzonase). Any recombinant polyhistidine tagged proteins were purified from the cell lysate using a magnetic nickel bead system on a 96-well plate. Protein expression was determined by separating proteins using sodium dodecyl sulphate (SDS)- poly acrylamide gel electrophoresis (PAGE) and visualizing them using Coomassie staining. Protein sizes were estimated the according to a reference protein sample where the molecular weight of protein bands in the reference was known in kilodaltons (kDa).



**Figure 5-24 Stages of *E. coli* protein expression trials**

The stages of protein expression trials employed at the OPPF included: 1 – Transformation of the *E. coli* recombinant protein expression strains, Lemo21 and Rosetta; 2 – Blue/white screening of colonies for positive transformants; 3 – Overnight induction of recombinant protein expression via the addition of IPTG or by auto-induction; 4 – Affinity purification of induced *E. coli* cell lysates; 5 – SDS-PAGE screen of high yield recombinant protein expression.

## **EUKARYOTIC EXPRESSION TRIALS**

The pOPIN vectors may also be used for protein expression in Eukaryotic cell lines, primarily the Sf9 insect cell line and the HEK293T human cell line. However, the experimental methods required for these host systems are of greater technical complexity than for *E.coli*. All high-throughput protein expression trials in the Eukaryotic expression systems were performed by OPPF staff members using published protocols<sup>307</sup>.

### **HEK293T**

The HEK293T cell line (Life Technologies) is a derivation of human-embryonic kidney cell line (HEK) that have been optimized for recombinant protein expression. Their properties include neomycin resistance, higher transfection efficiency, vector production and increased transduction efficiency. This cell line was transfected with the purified plasmid constructs. Both soluble and insoluble fractions of cell-lysates were analysed for recombinant protein expression via Western blotting using an anti-histidine tag antibody.

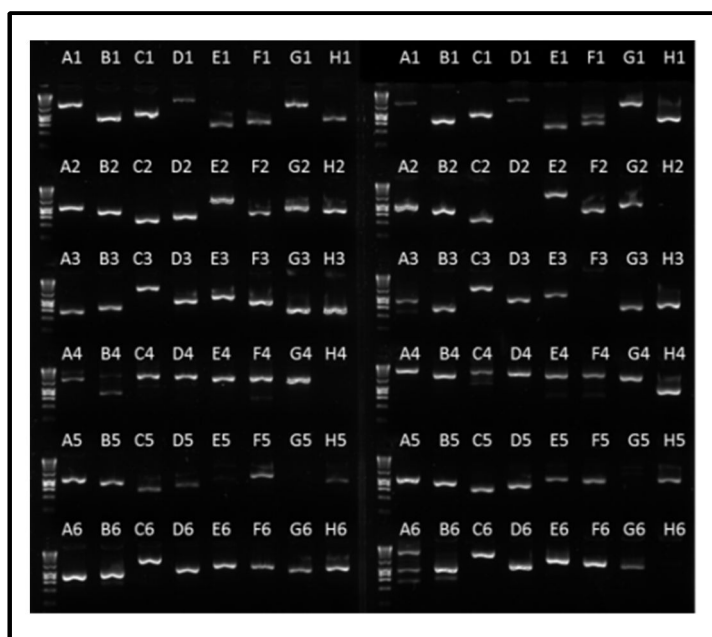
### **Sf9**

Sf9 cells are a clonal isolate from the insect *S. frugiperda*. In contrast to the other cell types used for recombinant protein expression, these cells do not directly uptake the plasmid vector. Instead, the recombinant genetic material is inserted into the Sf9 genome using the Baculovirus Expression system. The recombinant material is first incorporated into a recombinant viral vector known as a bacmid. These high-molecular weight nucleotide sequences contain the genome of the baculovirus as well as other nucleotide sequences, which allow the insertion of template DNA. The Bac-to-Bac™ system (Invitrogen) was used for *PNPLA3* construct insertion using the *E. coli* strain DH10BAC, which have been genetically modified to contain a modified bacmid which allows for the site-specific transposition of the pOPIN plasmid DNA into its nucleotide sequence. This occurs due to two the presence of the p10 baculovirus promoter and the flanking lef2 (ORF 603) and ORF1629 baculovirus recombination sites (Figure 5-20). The purified bacmids are used to infect sf9 cells resulting in the incorporation of the bacmid genome as well as the construct nucleotide sequence into the genomes of Sf9 cells. This infection requires several rounds of optimisation for high-yield protein expression<sup>306</sup>. The soluble fractions of cell-lysates were analysed for recombinant protein expression via Western blotting using an anti-histidine tag antibody.

## 5.4.3 - RESULTS

### PLASMID CONSTRUCTION

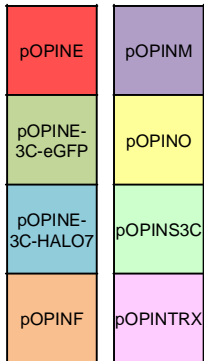
The majority of PNPLA3 template amplification PCR assays were successful, although a minority failed (Figure 5-25). As these assays were performed in duplicate, there were suitable inserts for all constructs except for the sample in position G5 (pOPINS3C-PNPLA3-1-250). These PCR amplified inserts, underwent InFusion cloning with linearised pOPIN vector suite plasmid backbones. All of the plasmid constructs, except for the sample in position D1 (pOPINE-PNPLA3-300-481), passed through the stages of InFusion cloning, transformation and plasmid purification. The remaining plasmid constructs were purified from having an average concentration of 433 ng/ $\mu$ L. From each duplicate well, a single plasmid construct was selected for protein expression trials. The same 48-well plate layout was maintained, although well positions G5 and D1 were empty and were henceforth used as positions for positive or negative controls (Figure 5-26).



**Figure 5-25 A comparison of correct *PNPLA3* insert size for In-Fusion cloning**

Different fragments of *PNPLA3* underwent PCR amplification with primers containing pOPIN backbone specific overhangs creating template DNA sequences suitable for insertion using InFusion cloning. The success of PCR was confirmed on an agarose gel (in duplicate) of the template DNA amplicons created using PCR amplification. Successful PCR amplification was judged based on the presence of a single, well-defined band, which correlated with the expected size of the template.



	1	2	3	4	5	6	
<b>A</b>	PNPLA3 2-481	PNPLA3 2-250	PNPLA3 300-400	PNPLA3 2-481	PNPLA3 2-250	PNPLA3 300-400	<b>Plasmid backbone</b> 
<b>B</b>	PNPLA3 2-175	PNPLA3 300-481	PNPLA3 271-400	PNPLA3 2-175	PNPLA3 300-481	PNPLA3 271-400	
<b>C</b>	PNPLA3 2-250	PNPLA3 300-400	PNPLA3 2-481	PNPLA3 2-250	PNPLA3 300-400	PNPLA3 2-481	
<b>D</b>	EMPTY	PNPLA3 271-400	PNPLA3 2-175	PNPLA3 300-481	PNPLA3 271-400	PNPLA3 2-175	
<b>E</b>	PNPLA3 300-400	PNPLA3 2-481	PNPLA3 2-250	PNPLA3 300-400	PNPLA3 2-481	PNPLA3 2-250	
<b>F</b>	PNPLA3 271-400	PNPLA3 2-175	PNPLA3 300-481	PNPLA3 271-400	PNPLA3 2-175	PNPLA3 300-481	
<b>G</b>	PNPLA3 2-481	PNPLA3 2-250	PNPLA3 300-400	PNPLA3 2-481	EMPTY	PNPLA3 300-400	
<b>H</b>	PNPLA3 2-175	PNPLA3 300-481	PNPLA3 271-400	PNPLA3 2-175	PNPLA3 300-481	PNPLA3 271-400	

**Figure 5-26 Plate layout of pOPIN-PNPLA3 constructs**

The plate layout of the 48 plasmid constructs that were suitable for further heterologous protein expression trials in *E.coli*, Sf9 and HEK293T cells.

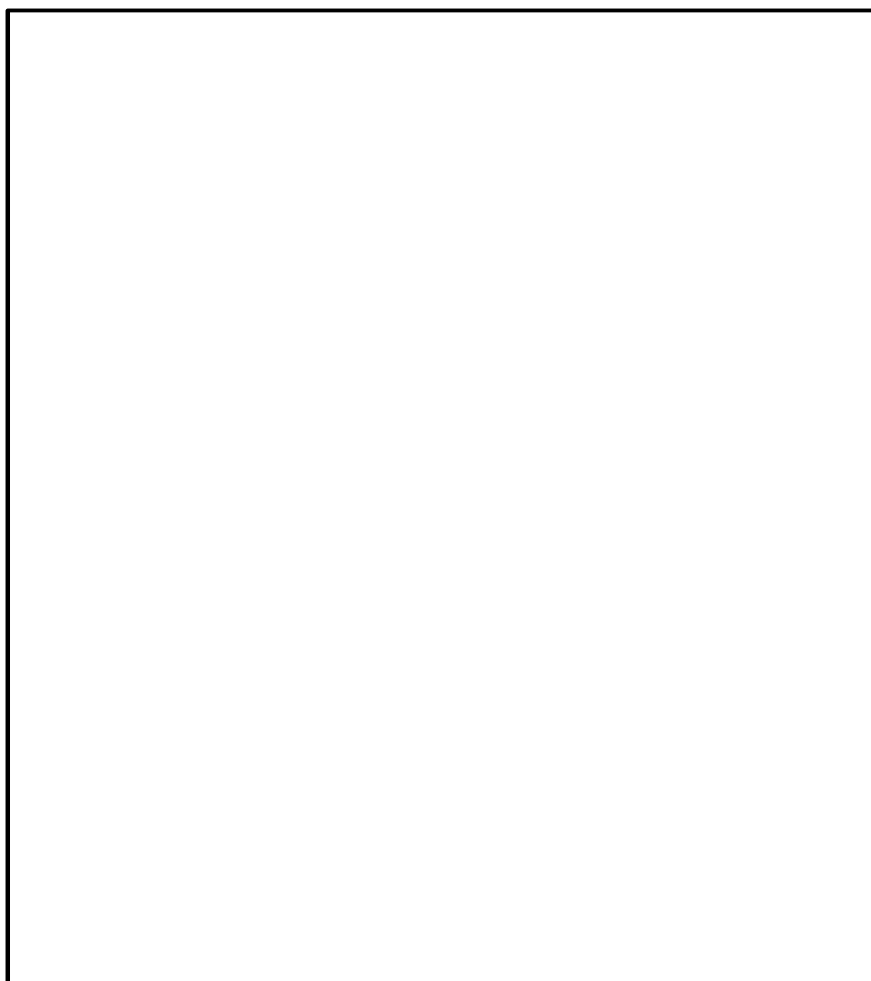
## ***E. COLI* EXPRESSION TRIALS**

### Overview

Two different strains of *E.coli* (Lemo21 and Rosetta) underwent expression trials under two different protein induction conditions (IPTG induction and auto induction). In all experiments, a positive controls plasmid (pOPINE-eGFP) was included in an empty position (G5) on the plate. The positive control sample demonstrated high expression of recombinant GFP under all conditions. In contrast, only a minority of the PNPLA3 pOPIN construct demonstrate visually discernible protein expression. The only PNPLA3 construct that had consistent protein expression under all experimental conditions (B5) produced bands that were at twice the expected molecular weight of the recombinant protein. Several constructs (B3, B1, E5, B2, C2, H2, B4 and C4) consistently expressed in the Rosetta strain under both the IPTG and auto-induction conditions. Several constructs (A4, B3, B6, A6 and E5) also consistently expressed in the Lemo21 strain under both IPTG induction and auto-induction conditions. However, none of the bands produced by these constructs was of the expected molecular weight.

## Rosetta Auto-induction

The affinity purified lysates from the Rosetta strain following auto-induction reveals 6 distinct protein bands when analysed by SDS-PAGE (Figure 5-27). Aside from the positive control (G5) none of the protein bands correspond with the estimated molecular weight of the recombinant fusion protein. Two of the samples (B1 and B5) show a distinct uniform band at approximately twice the size of the expected fusion protein. Another position (E5) shows multiple bands between 35-50 kDa suggesting partial cleavage of the recombinant protein. There are several faint bands on certain lanes, which may correspond to the recombinant fusion protein without the protein of interest attached. For example, the bands at around 40 kDa on lanes B4, C4 and D4 may correspond to the MBP fusion tag.

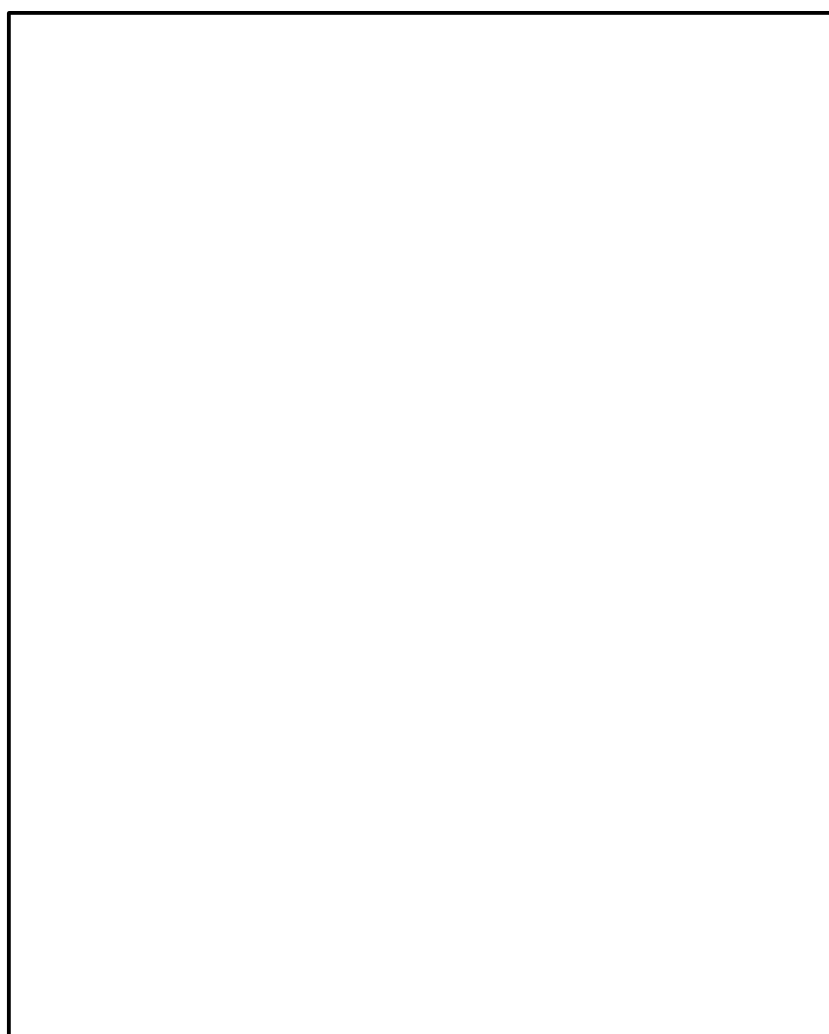


**Figure 5-27 SDS-PAGE gels of affinity purified, auto-induced Rosetta lysates**

The two images show Coomassie stained gels for the affinity purified samples from each construct. Red arrows highlight bands with evident expression of significant quantities of protein. The table below these images, gives the estimated molecular weight of these bands in comparison to the expected molecular weight of the expressed protein.

## Rosetta IPTG Induction

The affinity purified lysates from the Rosetta strain following IPTG induction of recombinant protein expression reveals 10 distinct protein bands when analysed by SDS-PAGE (Figure 5-28). Aside from the positive control (G5) none of the bands are at the expected molecular weight of the recombinant fusion proteins. Two positions (B1 and B5) show clear bands at approximately double the expected molecular weight of the expected fusion protein. Other positions (B2, C4 and E5) demonstrate multiple distinct bands suggesting partial cleavage of the expressed protein. Two of the distinct bands (C2 and F5) have similar molecular weights as the expected fusion protein tags GFP and SUMO.

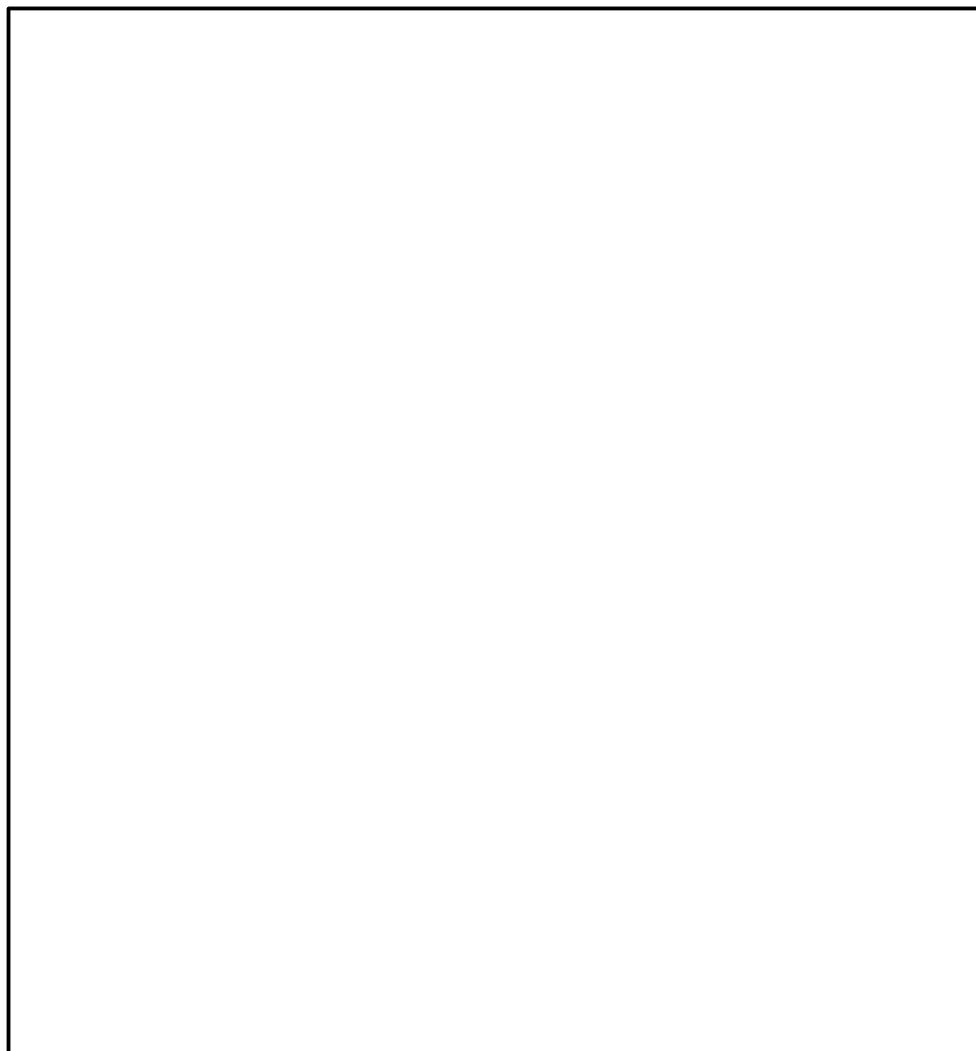


**Figure 5-28 SDS-PAGE gels of affinity purified, IPTG-induced Rosetta lysates**

The two images show Coomassie stained gels for the affinity purified samples from each construct. Red arrows highlight bands with evident expression of significant quantities of protein. The table below these images, gives the estimated molecular weight of these bands in comparison to the expected molecular weight of the expressed protein.

## Lemo21 Auto-induction

The affinity purified lysates from the Lemo21 strain following auto-induction of recombinant protein expression reveal 5 distinct recombinant protein bands when analysed by SDS-PAGE (Figure 5-29). Aside from the positive control (G5), none of the bands are at the expected molecular weight of the recombinant fusion proteins. Lane B5 shows a clear band at approximately double the expected molecular weight of the expected fusion protein. Lane B6 shows a clear band which is 10 kDa larger than the expected fusion protein. Position A4 shows a predominant band at ~70 kDa and multiple bands at around 36-45 kDa suggesting partial cleavage of the recombinant protein. The distinct band in lane A6 has the same molecular weights as the expected fusion protein tag SUMO.

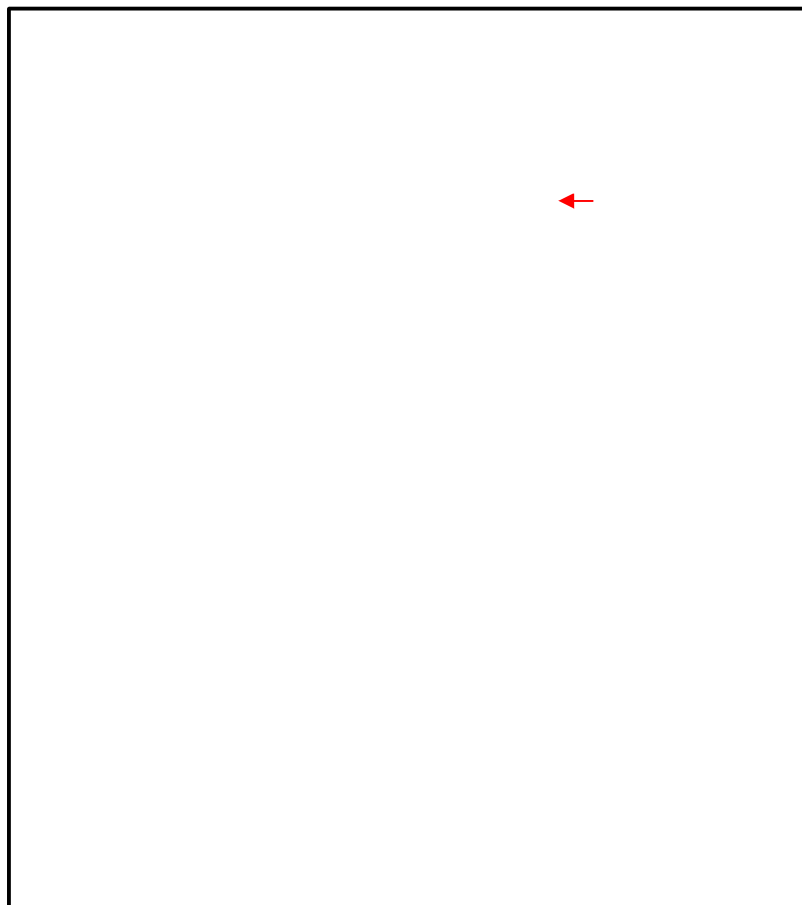


**Figure 5-29 SDS-PAGE gels of affinity purified, auto-induced Lemo21 lysates**

The two images show Coomassie stained gels for the affinity purified samples from each construct. Red arrows highlight bands with evident expression of significant quantities of protein. The table below these images, gives the estimated molecular weight of these bands in comparison to the expected molecular weight of the expressed protein.

## Lemo21 IPTG Induction

The affinity purified lysates from the Lemo21 strain following IPTG induction of recombinant protein expression reveal 6 distinct recombinant protein bands (Figure 5-30). The intense bands correspond to the well positions: Aside from the positive control (G5) none of the bands are at the expected molecular weight of the recombinant fusion proteins. Lane B3 shows a clear band at, which is 10 kDa smaller than the expected fusion protein. Lane B5 shows a clear band at approximately double the expected molecular weight of the expected fusion protein. Lane B6 shows a clear band which is 10 kDa larger than the expected fusion protein. Position A4 shows a predominant band at ~70 kDa and multiple bands at around 36-45 kDa. The lane E5 also shows multiple bands between 35-45 kDa; both suggesting partial cleavage of the recombinant protein. The distinct band in lane A6 has the same molecular weights as the expected fusion protein tag SUMO.



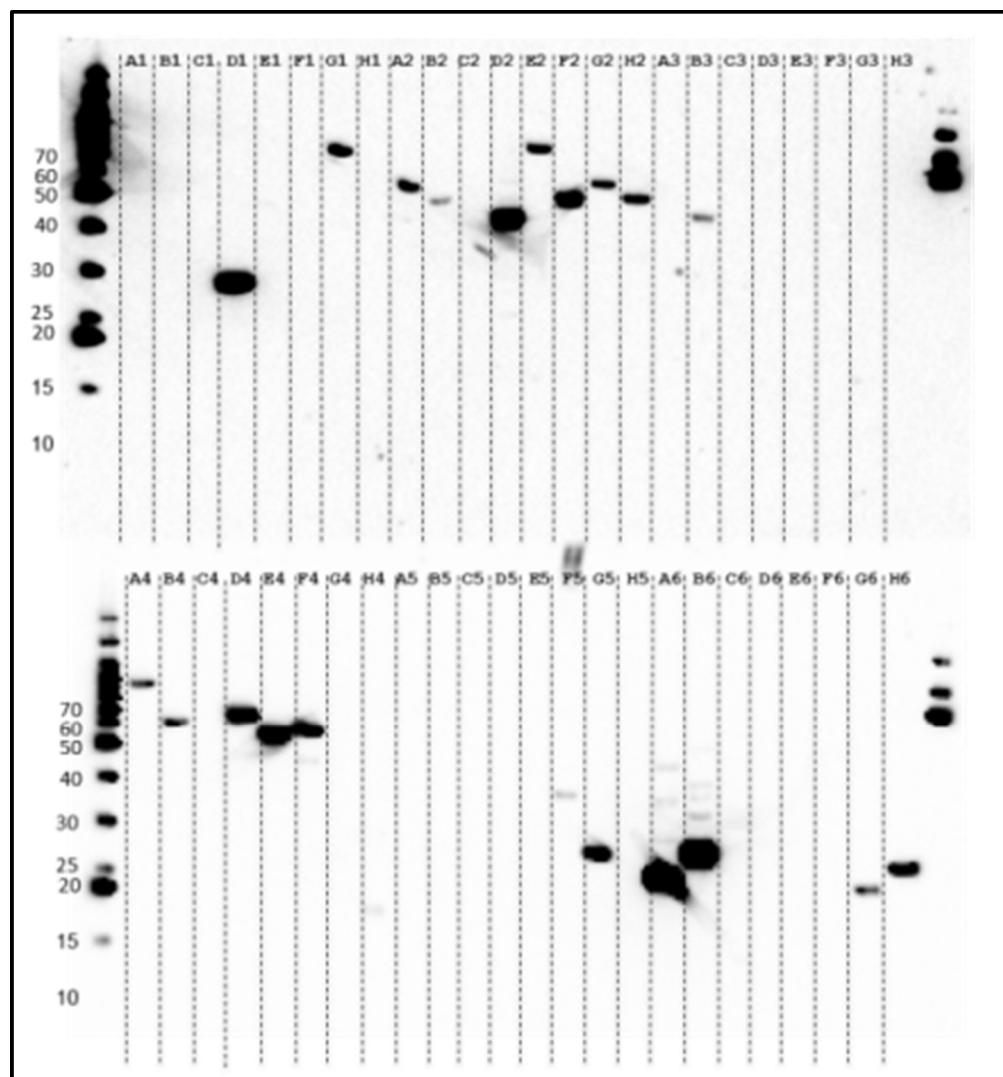
**Figure 5-30 SDS-PAGE gels of affinity purified, IPTG-induced Lemo21 lysates**

The two images show Coomassie stained gels for the affinity purified samples from each construct. Red arrows highlight bands with evident expression of significant quantities of protein. The table below these images, gives the estimated molecular weight of these bands in comparison to the expected molecular weight of the expressed protein.

## EUKARYOTIC EXPRESSION

### HEK293T

The expression of protein in HEK293T cells was determined using Western blotting (Figure 5-31). Due to the higher sensitivity of this technique over Coomassie staining, it provides little information regarding levels of protein expression. Twenty of the constructs expressed recombinant protein in HEK293T cells. Nearly all of the constructs with a pOPINE-3C-HALO7 and pOPINE-3C-GFP backbones expressed a recombinant protein at the expected molecular weight. There was a trend for the expression constructs to contain a fusion protein tag.

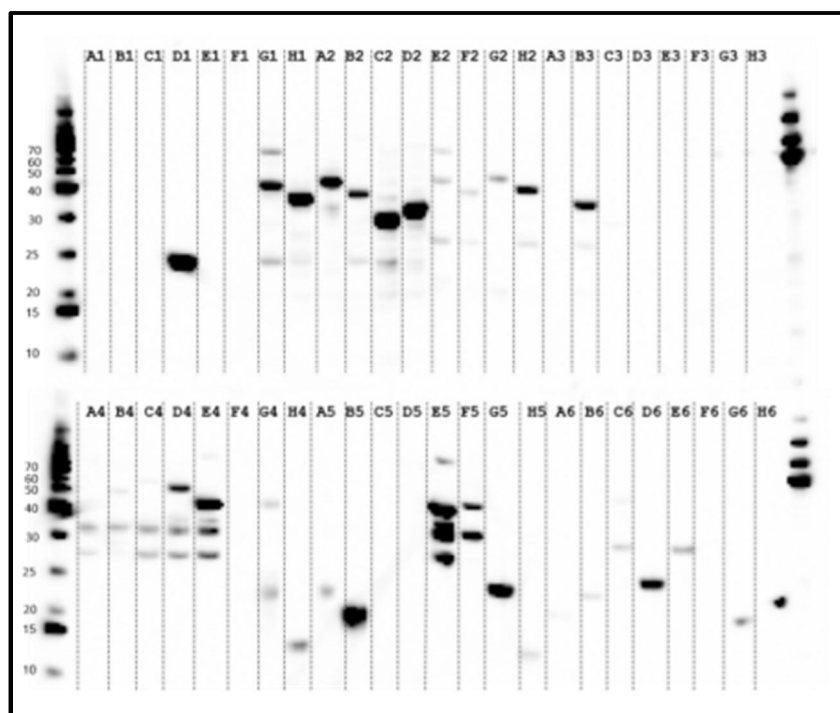


**Figure 5-31 PNPLA3 expression trials in HEK293T cells**

Images show the results of a Western-blot on the crude lysate purified from HEK293T cells. The black bands on these images represent protein samples containing a polyhistidine tag motif and thus represent heterologous protein. Standardized molecular weight markers on the extreme left-hand and right-hand lanes.

## SF9

The expression screen in Sf9 cells largely mirrored the results of the screen in HEK293T cells (Figure 5-32). The majority of samples with discernible protein expression contained a fusion partner tag. There was evident proteolytic degradation in the Sf9 cells and in some instances, only the fusion protein tag was detected.



**Figure 5-32 PNPLA3 expression trials in Sf9 cells**

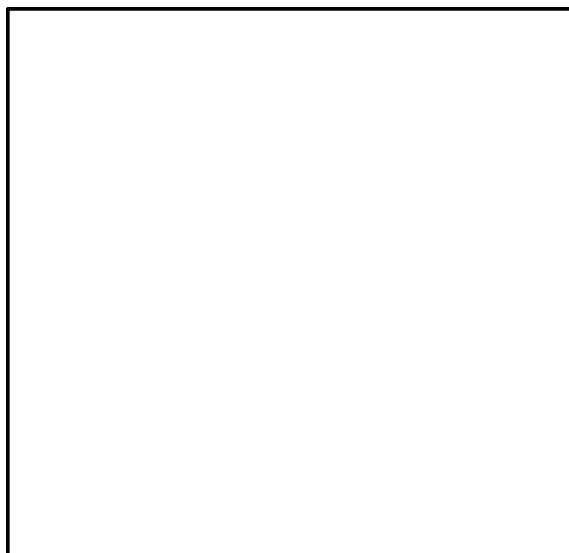
Images show the results of a Western-blot on the crude lysate purified from Sf9 cells. The black bands on these images represent protein samples containing a polyhistidine tag motif and thus represent heterologous protein. Standardized molecular weight markers on the extreme left-hand and right-hand lanes. **SUMMARY**

A total of 46 plasmids were purified at a sufficient yield, for PNPLA3 expression trials. The protein expression trials of these construct demonstrates some variability and inconsistency between eukaryotic (HEK293T and Sf9) and prokaryotic (*E.coli*) expression systems. In *E. coli*, several constructs demonstrated significant protein expression, however, none of these protein samples were of the expected molecular weight. In the Eukaryotic cell lines, a number of constructs also demonstrated protein expression although these were mostly at the expected molecular weight. In the HEK293T cells, most of the recombinant proteins were of their expected molecular weight. In the Sf9 cells, results were largely similar albeit with evident proteolytic degradation in several cases. It seems apparent that a number of plasmids have the potential to express in PNPLA3 in *E.coli*, HEK293T and Sf9 cells.

The follow up of these constructs for protein crystallography purposes was guided by practical as well as experimental factors. First, the UCL crystallography laboratory is

optimised for heterologous protein expression and purification in *E. coli*. This laboratory setup is common as *E.coli* are tractable to genetic manipulation and rapidly attain a sufficiently large biomass required for obtaining milligram amounts of protein. This is one of the reasons why over 90% of the protein structures in the protein databank are purified using an *E.coli* expression system<sup>306</sup>. Therefore, plasmid constructs were selected for follow up if they demonstrated protein expression in *E.coli*. Second, this research is aimed at characterizing the effect of the Ile148Met amino-acid substitution. For this reason, characterizing the patatin domain in which this substitution resides is of a greater priority than the C-terminal region of PNPLA3.

In the constructs with discernible protein expression in *E.coli*, only three contained the patatin domain (A4, E5, B1). Two of these wells contain identical constructs (A4 and E5). The protein expressed by the E5 construct demonstrates distinct protein expression yet significant proteolysis. The protein expressed by the A4 construct is of a larger molecular weight (closer to the expected weight of PNPLA3) and appears to be less degraded, however, there is also less protein expression. Based on these findings and empirical decision making the plasmid in well A4 plasmid (Figure 5-33) was selected for further characterization. This plasmid was selected as: (i) it demonstrates heterologous protein expression in *E.coli*; (ii) it appears to express of the full length PNPLA3 protein containing the patatin domain; and, (iii) the expressed protein demonstrates less evident proteolytic degradation than similar constructs.



**Figure 5-33 A plasmid map of the A4 plasmid construct**

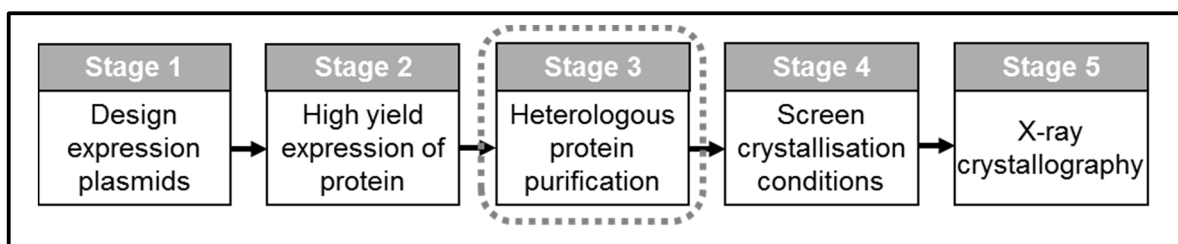
In this construct the gene sequence for the entire PNPLA3 protein is inserted into a pOPINM vector backbone. The open reading frame should express an N-terminal polyhistidine tag motif followed by a maltose binding protein fusion partner, a short linker region and then PNPLA3.



## 5.5 - PROTEIN PURIFICATION

### 5.5.1 - OVERVIEW

The purification of proteins for crystallography purposes often requires an optimised protocol. Determining the optimal conditions for such protocols is largely empirical. The focus of this section is the verification and optimisation of a protein purification protocol for a selected plasmid construct. This effort results in the verification of recombinant PNPLA3 purification from an *E.coli* expression system with testing of the proteins suitability for crystallography screening (Figure 5-34).



**Figure 5-34 Experimental stages to obtain a protein structure**

At the third stage, the primary aim is to purify the protein at a sufficient yield, in a suitable form for further stages crystallography

### 5.5.2 - MATERIALS AND METHODS

#### PLASMID AMPLIFICATION & PURIFICATION

*E. coli* were transformed (MAX Efficiency® DH5α™, 18258-012, Invitrogen) using the manufacturer stated protocol using routine growth media (Table 5-5). The plasmid DNA was purified from overnight cultures using an Axygen Mini Prep kit (CORNING AXYGEM APMNP50, 14-223-009, Fisher Scientific) using the manufacturers protocol.

#### Plasmid Sequencing

Sanger sequencing was performed externally (SourceBioscience, Nottingham, UK) on selected plasmid constructs<sup>378</sup>. Sequencing primers were designed for complementarity to nucleotide sequence regions near the overlap between the plasmid backbone and the *PNPLA3* insert (Table 5-6). Sequencing results were visualized and underwent alignment the predicted plasmid sequence using the software SNAPGene<sup>159</sup>.

Table 5-5 Media used for propagating *E. coli*

Solution	Recipe/Contents
1000X Ampicillin	100 mg/mL
LB (autoclaved)	10 mg/mL NaCl, 10 mg/mL Tryptone, 5 mg/mL yeast extract
LB Agar (autoclaved)	LB + 35 mg/mL Agar

All buffers made in double-deionized H<sub>2</sub>O except where otherwise explicitly stated  
Abbreviations: LB – Luria Bertani

Table 5-6 Plasmid sequencing primers

Primer Name	Sequence (5'3')
T7 Forward	TAATACGACTCACTATAGGG
Rabbit Beta Actin Reverse	TTTTGGCAGAGGAAAAAGA
Rs2076212	GAAGGTGACCAAGTTCATGCTCCAGCTCATCTCCGGCAAATAG
Rs2076213	GAAGGTCGGAGTCAACGGATTGTTCTCCGACAGGGTCTCG

## RECOMBINANT PROTEIN EXPRESSION IN *E. COLI*

### Transformation and Induction

*E. coli* (Lemo21(DE3), C2528H, New England BioLabs®) were transformed using the manufacturer stated guidelines and grown in an overnight culture in 5 mL of LB with antibiotics (0.5X Ampicillin and 1X Chloramphenicol). A 500 µL aliquot of this overnight culture was mixed with 500 µL of glycerol, flash frozen, and kept for long-term storage in a -80°C freezer. Another aliquot of the overnight culture was used to inoculate 6 litres of Terrific Broth (TB) media containing antibiotics (0.5X Ampicillin, 1X Chloramphenicol) and grown for several hours (37°C, 150 RPM) until reaching an optical density at 600 nm. At this cell density, recombinant protein expression was induced via the addition of IPTG and the induction regulator L-rhamnose. Induction was performed overnight 12-16 hours (21°C, with agitation at 150 RPM).

Table 5-7 *E. coli* growth and induction media

Solution	Constituents
TB (Autoclaved)	Tryptone 12 mg/mL, 24 mg/mL Yeast extract, 0.4% glycerol, 17 mM KH <sub>2</sub> PO <sub>4</sub> , 72 mM K <sub>2</sub> HPO <sub>4</sub>
1000X Chloramphenicol	34 mg/mL chloramphenicol in 100% ethanol
IPTG solution	0.8 M IPTG
L-rhamnose	0.8 M L-rhamnose

All buffers made in double-deionized H<sub>2</sub>O except where otherwise explicitly stated  
 .Abbreviations: TB –Terrific Broth, IPTG - Isopropyl β-D-1-thiogalactopyranoside; mM – millimolar

### Harvesting & Lysis

*E. coli* cells were separated from overnight growth media via ultra-centrifugation (4<sup>0</sup>C, 3000 RPM, 30 minutes, Beckman Coulter rotor JA-25 550). The cell pellets obtained following centrifugation were weighed and either directly lysed or stored in a -20<sup>0</sup>C freezer. At the time of use, additional protease inhibitor tablets (cOmplete, Mini, EDTA-free; 11836153001; Roche Life Science) and lysozyme (100 ng/μL, Sigma, L7651) and, DNase1 (40 μg/mL) were added to all lysis buffers and were mixed with *E. coli* cell pellets (~2.5 mL of lysis buffer per gram of cell pellet). The lysis mixture underwent sonication (9 s pulse, 10 minutes, 10 kHz) on ice followed by separation in soluble and insoluble fractions by ultracentrifugation (Beckman coulter, rotor JA 17), (4<sup>0</sup>C, 30,000g, 45 minutes). The soluble fraction underwent syringe filtration (0.45 μm pore size), and were stored on ice before further purification while insoluble fractions were stored at -20<sup>0</sup>C.

### **NICKEL AFFINITY PURIFICATION**

Nickel affinity purification relies on the formation of specific coordinate covalent bonds between charged nickel (Ni<sup>2+</sup>) ions and the polyhistidine tag motif commonly engineered on to recombinant proteins. In this work, HisTrap FF Crude columns of both 1 mL (product code: 17-5319-01) and 5 mL (product code: 17-5255-01) column volumes were used for affinity purification<sup>137</sup>. The process of affinity purification using HisTrap is controlled either manually using a syringe or, with finer control, using a fast-protein liquid chromatography machine. This allows the precise control of flow rate and the use of elution gradients using different media as input. Elution of proteins bound to columns was effected with the addition of a high imidazole elution buffer. Protein concentrations in the eluted fractions were estimated by measuring the ultraviolet absorbance at 280 nm using a nanodrop spectrophotometer.

## GEL ELECTROPHORESIS

### SDS-PAGE

SDS-PAGE allows the estimation of protein molecular weight, protein concentration and protein purity. SDS-PAGE gels were cast manually a maximum of 1 week before use. Samples were prepared for SDS-PAGE by mixing with 2X Laemmli buffer (1:1) followed by heat denaturation (95°C for 5 minutes). Gels were loaded into a cassette and then into a tank and loaded with Tris-Glycine running buffer. No greater than 10 µL of the samples were loaded into each lane of any gel and these were run at constant amps (25 mA, 1-2 hours). All gels were run with either Broad Range (Promega, V8491) or Rainbow™ (GE Healthcare Biosciences RPN755E) protein molecular weight markers.

### Coomassie Staining

Protein band visualization was effected via immersion of SDS-PAGE gels in Coomassie stain (with agitation at room-temperature, 1-2 hours) followed by immersion in de-stain solution.

Table 5-8 The constituents of polyacrylamide gels, electrophoresis running buffers and gel staining media

Solution/Media/Buffer	Constituents
Separating gel	12% Acrylamide; 150 mM Tris-HCl (pH 8.8); 0.4% w/v SDS
Stacking gel	4% Acrylamide; 50 mM Tris-HCl (pH 6.8); 0.4% w/v SDS
Tris-Glycine running buffer	25 mM Tris-HCl pH 8.3; 192 mM glycine; 0.1% w/v SDS
SDS-PAGE loading buffer (Laemmli Buffer)	100mM Tris (pH 6.8); 4% w/v SDS; 0.2% w/v Bromophenol blue; 20% v/v Glycerol
Coomassie gel de-stain	40% Methanol; 10% Acetic Acid
Coomassie gel stain	0.125% w/v Coomassie blue; 50% v/v Methanol; 0% v/v Acetic acid

All buffers made in double-deionized H<sub>2</sub>O.

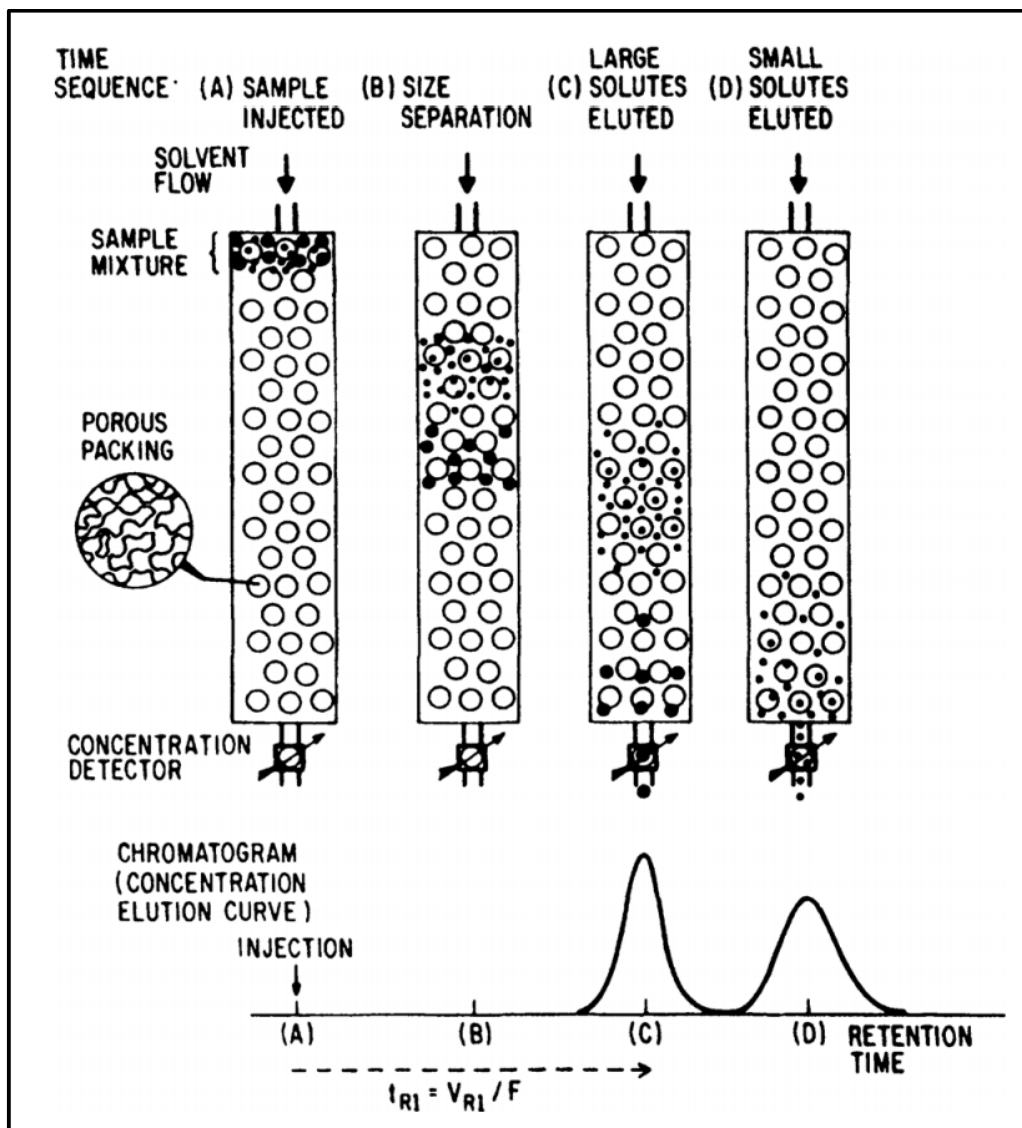
Abbreviations: SDS – Sodium dodecyl sulphate; Tris - Tris(hydroxymethyl)aminomethane; w/v – weight per volume ratio; v/v –volume per volume ratio

## MASS SPECTROMETRY

Protein mass spectrometry (MS) allows the precise determination of the different molecules present in a sample. Protein bands were cut from Coomassie stained SDS-PAGE gels and underwent preparation for MS, involving: the removal of Coomassie stain; reduction and alkylation to enhance sample coverage via the removal of disulphide bonds; and, tryptic digestion of the protein into peptide fragments. MS analysis of peptide fragments present in samples was performed on a Velos Orbitrap Mass spectrometer. From the MS spectra data, the proteins present in each sample were determined using Proteome Discoverer 1.3 via searches of the UniProt human and *E. coli* database with the Mascot search engine. Carbamidomethylation was set as a fixed modification with methionine oxidation as variable. The preparation of samples for MS were performed using a standard protocol<sup>392</sup>.

## SIZE EXCLUSION CHROMATOGRAPHY

Size-exclusion chromatography (SEC) separates molecules based on size. It is often the final stage during protein purification through which the protein of interest is separated from co-purified contaminants and its level of homogeneity determined. A SEC column contains beads of either agarose or polyacrylamide; based on the column design these beads have different molecular pore size and thus alter the passage of molecules through the column effecting separation (Figure 5-35). All SEC on PNPLA3 protein samples were performed on a Varian Prostar high performance liquid chromatography machine (HPLC) with monitoring the elution with an ultra-violet/visible light detector (Model 345) and data were analysed using Varian Star 5.5 software. A Superdex 200 10/300 GL column (Figure 5-35) with a volume of 24 mL was used in all experiments; this column has an optimum separation range of for proteins with a molecular weight ranging from 10-600 kDa. Before use, the column was primed with a SEC running buffer via passing the media through the column for several column volumes. Before use, all SEC buffers were degassed. To perform SEC, aliquots (50-100  $\mu$ L) of concentrated protein sample were injected into the HPLC machine at a flow rate of 0.5 mL/min. Sample passage through the column was monitored via an elution chromatogram (Figure 5-35).



**Figure 5-35 Principle of size-exclusion chromatography**

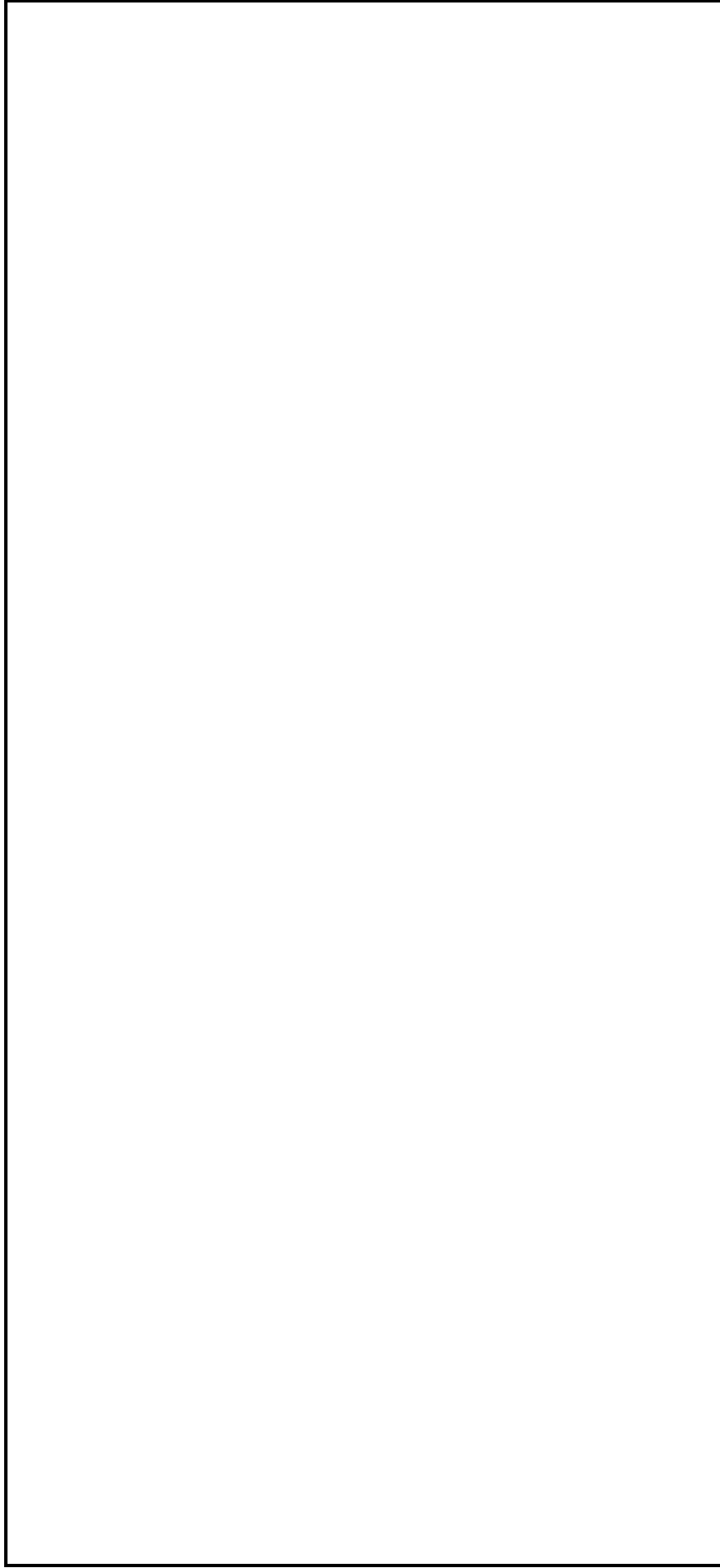
This sequence an example of a sample flowing through a SEC column and the profile measured on a UV detector based on elution where: A – a sample is injected into a column and a flow of solvent is applied; B – Samples are separated by size when run through the column; C – larger solutes pass more quickly through the column and are eluted first as measured by a peak on the chromatogram; D – Smaller solutes pass more slowly through the column and are eluted later as measured by a peak on the chromatogram.

Image from Striegel et al., 2009<sup>416</sup>

### 5.5.3 - RESULTS

#### PLASMID SEQUENCING

The sequence off the A4 plasmid was confirmed using insert specific primers (Table 5-6) and Sanger sequencing demonstrating that both MBP and *PNPLA3* nucleotide sequences were present in the A4 construct. This plasmid is expected to express a fusion protein containing a N-terminal polyhistidine tag motif followed by a maltose binding protein tag, a short linker region and the entire PNPLA3 protein. The sequence of this fusion protein is expected to contain 869 amino acids and have a molecular weight of 95.4 kDa.



**Figure 5-36 Sequencing confirmation of the PNPLA3-MBP construct**

The sequencing results were aligned with the predicted nucleotide sequence of the plasmid. The three sequencing chromatograms obtained align to the MBP gene and the expected *PNPLA3* insert sequence

## PROTEIN PURIFICATION

Two protein purifications experiments were performed independently at different time-points:

1. The first purification experiment was a scale-up of the purification protocol performed at the OPPF
2. The second purification experiment was optimised for the purification of a high-yield protein suitable for crystallization experiments.

Between both of these experiments, the transformation procedure, strain of *E. coli* used and over-night induction protocol used were identical, and were as given in the earlier materials and methods. The primary difference between these purifications was the choice of buffers (e.g. presence of reducing agents in buffers), the equipment used during the purification (e.g. fast protein liquid chromatography) and characterization of protein samples following purification (e.g. mass spectrometry).

### First Protein Purification

In total, the 6 litres of overnight culture produced an *E. coli* cell pellet weighing 19.0 grams. The *E. coli* cell lysate was passed through a 1 mL FF HisTrap manually with a syringe followed by the wash and elution buffer (Table 5-9). All flow-through and washes were collected in 1 mL fractions and protein concentration determined. The pooled sample was concentrated using 4 mL centrifugal concentrators (Vivaspin®, Massachusetts, USA) until a protein concentration of >20 mg/mL in a volume of 500 µL. Analysis of the affinity purified eluent by SDS-PAGE and Coomassie staining revealed three major bands corresponding to molecular weight of around ~40 kDa, ~45 kDa and ~66 kDa (lane 10: Figure 5-38).

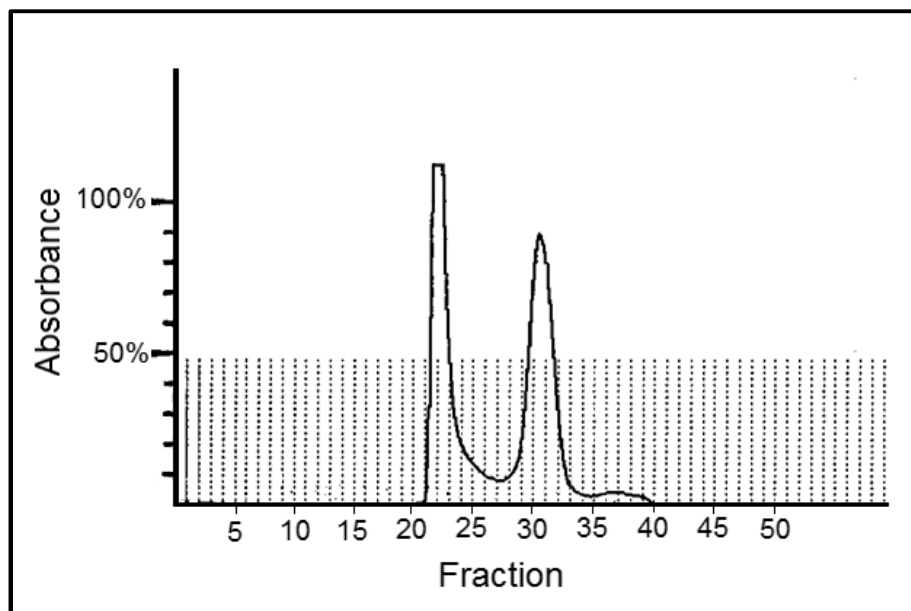
Table 5-9 Buffers used in the first purification

Buffer	Constituents
Wash Buffer	50 mM NaH <sub>2</sub> PO <sub>4</sub> (pH 8.0), 300 mM NaCl, 20 mM Imidazole, 0.05% Tween-20
Elution Buffer	50 mM NaH <sub>2</sub> PO <sub>4</sub> (pH 8.0), 300 mM NaCl, 250 mM Imidazole, 0.05% Tween-20
Lysis Buffer	50 mM NaH <sub>2</sub> PO <sub>4</sub> (pH 8.0), 300 mM NaCl, 10 mM Imidazole, 1% Tween-20
SEC buffer	20 mM Tris-HCl (pH 7.5), 300 mM NaCl, 5% Glycerol, 0.02% Sodium Azide

All buffers made in double-deionized H<sub>2</sub>O.  
Abbreviations: SEC – Size exclusion buffer



This protein sample underwent size exclusion chromatography (Figure 5-37) where it eluted as two major peaks. The first sample to elute, corresponding to the first peak on the chromatogram (Figure 5-37: fractions 20-25), eluted in the void volume of the column. This void volume is essentially the volume of the liquid in a SEC column which does not enter the permeable matrix of the beads (i.e. the mobile phase) showing that the protein sample is of a greater molecular mass than the maximum pore size of the SEC column (>600 kDa). The second sample to elute, corresponding to the second peak on the chromatogram (Figure 5-37: fractions 30-35) corresponds to a protein sample that has entered the permeable matrix of the beads (i.e. the stationary phase). The peak fractions from each major peak were combined, spin concentrated and visualized by SDS-PAGE (Figure 5-38). The fractions which eluted during the first peak on the SEC chromatogram (fractions: 22, 23, 24) correspond predominantly to the ~66 kDa band protein sample; there are numerous other less intense protein bands of varying size. Although the protein band on an SDS-PAGE gel runs at a mass of ~66kDa, it runs at a mass greater than It is likely that these proteins are forming a soluble compound in solution with a mass greater than 600 kDa. The fractions which eluted during the second peak on the chromatogram (fractions: 30, 31) correspond to two bands with a mass of ~40 kDa and ~45 kDa respectively (lanes 6 and 7: Figure 5-38); these samples are relatively pure with little co-purification of other protein.



**Figure 5-37 Size exclusion purification elution profile from the first purification**

This chromatogram shows the UV absorption percentage (y-axis) versus the elution of the sample into numbered fractions (x-axis). This demonstrates the elution of two peaks, the first between fractions 21-23 and the second between fractions 29-32.



**Figure 5-38 The protein bands present in the peak elution fractions**

The protein samples present in the peak fractions eluted during size exclusion chromatography were visualized using SDS-PAGE. Lanes: (1) Marker; (2) Fraction 22; (3) Fraction 23; (4) Fraction 24; (5) Fraction 28; (6) Fraction 30; (7) Fraction 31; (8) Fraction 38; (10) Affinity purified sample

The four bands selected for MS analysis included the three most distinct bands and a more faint band corresponding to the expected molecular weight of the PNPLA3-MBP fusion protein (Figure 5-39). These were of predicted molecular weights 40 kDa (band 1: Figure 5-39), ~45 kDa (band 2: Figure 5-39), ~66 kDa (band 3: Figure 5-39) and ~100 kDa (band 4: Figure 5-39). MS analysis confirmed that bands 1 and 2 contain the recombinant MBP fusion tag alone (Supplementary Table 7) whereas bands 3 and 4 contain both MBP and PNPLA3 (Supplementary Table 7). It is evident that the majority of the PNPLA3-MBP fusion protein does not run at its expected molecular weight on an SDS-PAGE gel.



**Figure 5-39 The protein bands selected for mass-spectrometry analysis**

This SDS-PAGE gel image is highlighted with the four bands that were cut from the gel and underwent MS analysis. The proteins present in these bands were: band 1 = maltose binding protein; band 2 = maltose binding protein; band 3 = maltose binding protein and PNPLA3; band 4 = maltose binding protein and PNPLA3

## Second Protein Purification

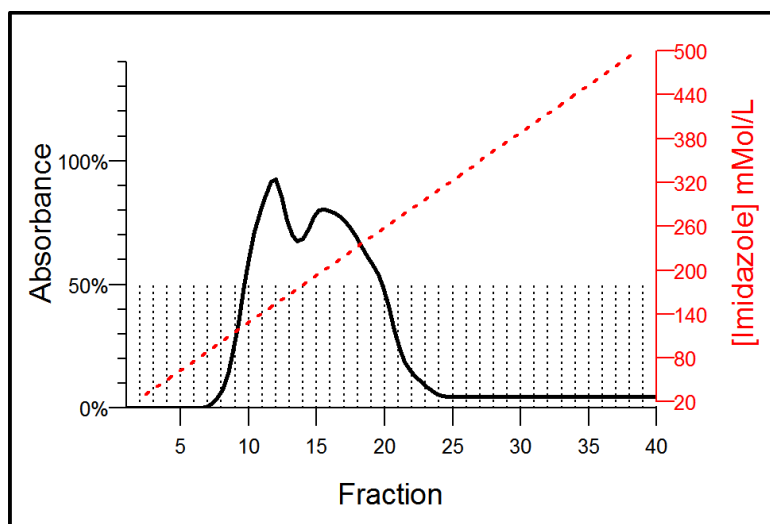
In total, the 6 litres of overnight culture produced an *E. coli* cell pellet weighing 28.97 grams. The soluble fraction of the *E. coli* cell pellet lysate was passed through a 5 mL HisTrap affinity column using a fast protein liquid chromatography machine (Äkta, Sweden) (flow rate: 5 mL/min, fraction size: 2.5 mL), followed by 3 passages of wash buffer followed by the gradient addition of the elution buffer over 100 mL (Table 5-10). This lysis buffer differed from that used in the first purification as it contained a higher salt concentration and the reducing agent DTT.

Table 5-10 Buffers used in the second purification

Buffer	Constituents
Wash Buffer	50 mM Tris-HCl (pH 7.5), 500mM NaCl, 30 mM Imidazole, 5% Glycerol
Elution Buffer	50 mM Tris-HCl (pH 7.5), 500mM NaCl, 500 mM Imidazole, 5% Glycerol
Lysis Buffer	50 mM Tris (pH 7.5), 500 mM NaCl, 30 mM Imidazole, 0.2% Tween-20, 2 mM DTT, 5% glycerol
SEC buffer	20 mM Tris-HCl (pH 7.5), 200 mM NaCl, 2mM DTT, 0.02% Sodium Azide

All buffers made in double-deionized H<sub>2</sub>O  
Abbreviations: SEC – Size exclusion chromatography

The protein sample eluted as two indistinct peaks across the imidazole gradient (Figure 5-41). SDS-PAGE visualization of the protein bands present in these peak fractions (Figure 5-41) demonstrates that the imidazole gradient effected the separation of the higher molecular weight protein samples, which correspond to the known band size of the PNPLA3-MBP fusion protein from the lower molecular weight cleaved MBP tag bands. The peak fractions of the higher molecular weight PNPLA3-MBP samples were pooled (Figure 5-41: fractions 16 to 21) and concentrated, producing approximately 25 mg of protein.



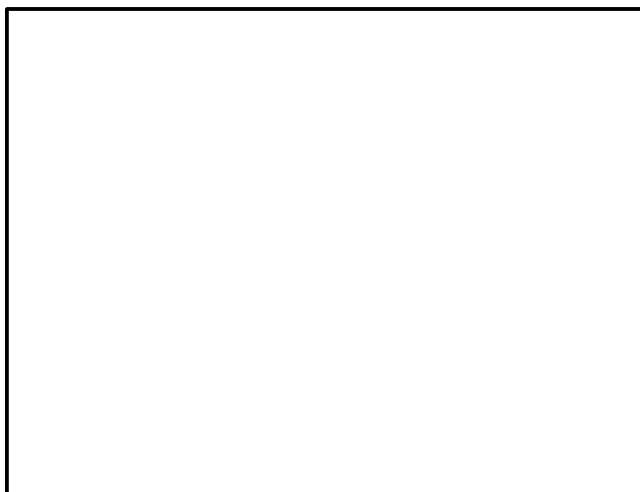
**Figure 5-40 Fast liquid protein chromatography elution profile from an affinity column**  
 Fast liquid protein chromatography was used to elute the bound protein samples from the nickel affinity column during the second purification. The UV absorbance was measured (left-hand y-axis, normalised to the maximum value) and plotted relative to the imidazole concentration (right-hand y-axis) for each fraction that was eluted (x-axis). \*The elution profile curve was derived from kernel mean smoothing line fit to the UV absorbance values recorded for each fraction



**Figure 5-41 Visualization of protein samples eluted from affinity column**  
 Aliquots of the cell lysate and different fractions from the fast liquid protein chromatography elution of the sample from the second purification. The elution gradient effects the separation of the cleaved MBP tag from the full-length protein sample. Lanes: (1) Marker; (2) Lysate; (3) Flow-through; (4) Fraction 10; (5) Fraction 12; (6) Fraction 16; (7) Fraction 19; (8) Fraction 21

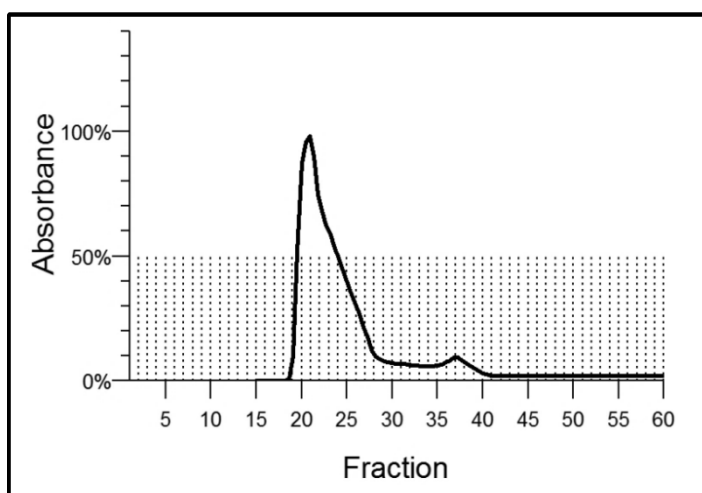
The concentrated sample was visualized at different concentrations to determine the sample purity. At lower concentrations (Figure 5-42: lane 1) the sample appears relatively pure with little co-purification of other proteins. However, when concentrated (Figure 5-42: lane 6), other protein bands become visible. The sample was also tested under semi-native PAGE conditions (i.e. without the presence of reducing agents or detergent in loading buffer) (Figure 5-42: lane 9). Under this condition, the protein sample fails to enter the separating polyacrylamide gel suggesting that the recombinant

protein is forming a high molecular weight compound. When passed through a size exclusion column, nearly all of the protein present elutes in the void volume of the column (Figure 5-43), consistent with the protein forming a high molecular weight compound in solution.



**Figure 5-42 Visualization of sample at several dilutions and denaturation conditions**

The concentrated protein sample purified from affinity purification was visualized at different concentrations and under different denaturation conditions. The predominant bands correlate with those bands from the previous purification, which were verified by mass spectrometry. With increasing sample concentration (lane 2 to lane 6) it is evident that there are multiple protein bands. When tested using different combinations of detergent and reducing agents (lane 7 to lane 9) the sample run under semi-native conditions (lane 9) fails to enter the gel. Lanes: (1) Marker; (2) 1:40 dilution of sample ; (3) 1:20 dilution of sample; (4) 1:10 dilution of sample; (5) 1:5 dilution of sample; (6) 1:2 dilution of sample; (7) 1:10 dilution of sample + Laemelli; (8) 1:10 dilution of sample + Laemelli + 8M Urea; (9) 1:10 dilution of sample + Loading buffer (No BME or SDS)



**Figure 5-43 Size exclusion chromatogram elution profile from the second purification**

This schematic representation of an elution profile of the concentrated protein sample from the second purification experiment. The UV absorbance was measured (left-hand y-axis, normalised to the maximum value) and plotted for each fraction that was eluted (x-axis). \*The elution profile curve was derived from kernel mean smoothing line fit to the UV absorbance values recorded at each fraction

#### 5.5.4 - SUMMARY

A protocol was developed and verified for the heterologous purification of PNPLA3 from an *E. coli* expression system. Despite purifying milligrams of the protein, the protein is not eluting in a homogeneous form, instead forming a high molecular weight compound. Before crystallography trials, this sample will require greater characterization and the optimisation of a protocol resulting in a stable and correctly folded protein.

#### 5.6 - DISCUSSION

In silico analysis of the sequence conservation, intrinsic disorder and three dimensional structure of PNPLA3 suggest the presence of two domains. The first amino-terminus domain of PNPLA3 comprises the already characterized<sup>459</sup> patatin domain, while the second putative carboxyl-terminal domain of PNPLA3 shares little to no sequence homology with any other known domains. Between these two domains, there is a peptide sequence with low sequence conservation and high predicted intrinsic disorder and hydrophobicity. This may be a domain linker region, although it shares similarities with lipid droplet binding regions of a protein<sup>332</sup> (e.g. predicted intrinsic disorder and hydrophobicity) and thus may play other roles in protein function. Lipid droplet binding is a known activity of PNPLA3, which may involve residues in this region of the protein<sup>300</sup>.

Others have investigated structural models of PNPLA3<sup>170,198,233,470</sup>. The predicted models of previous analyses share consistent features with model from the current analysis, in particular, the close spatial positioning of the catalytic dyad residues to the position of isoleucine 148. This similarity likely results, in part, from both models relying on the same structural information for model determination from the PDB, namely the structure of horse leaf nettle patatin (pat17). A unique feature of the structural model generated in this work, is that it comprises the entire protein rather than the patatin domain alone. Despite this, current predicted structural models of PNPLA3 are of low accuracy and therefore experimental structural determination is required to validate these models of PNPLA3.

Over forty plasmid constructs were created for the purpose of the heterologous expression of PNPLA3 for crystallographic purposes. Of the constructs that demonstrated significant protein expression, most generally contained a fusion protein tag (e.g. MBP, GFP, SUMO or HALO). Fusion partner tags generally enhance the solubility of proteins and therefore this finding may be consistent with insolubility of PNPLA3 as it is a known lipid droplet binding protein<sup>56</sup>. Others have expressed and purified heterologous PNPLA3<sup>170,187,233,334</sup>, with the purification protocols sharing in

many instances, features identified in the current purification such as insolubility<sup>144,157,334</sup>, proteolytic degradation<sup>233</sup> and the formation of higher than expected molecular weight compounds<sup>334</sup>. It seems likely that PNPLA3 may be a challenging protein to purify in a suitable form for crystallographic studies.

The purity and homogeneity of a protein sample are key parameters for the successful crystallisation and the subsequent determination of a protein's structure using X-ray crystallography<sup>140</sup>. Hence, the high molecular weight form of PNPLA3 purified in this work was unsuitable for further crystallographic analysis. There are several mechanisms which could explain the formation of this high molecular weight compound. The recombinant protein may be forming a soluble aggregate as MBP fusion proteins are known to solubilise unfolded tag allowing aggregate formation<sup>476</sup>. PNPLA3 is also a lipid droplet binding protein<sup>56</sup> and thus the purified protein may interact with, and co-purify with lipids forming a high molecular weight compound. Alternately, this may be the native state of PNPLA3 as it has been observed to form oligomers physiologically<sup>460</sup>. If ongoing studies determine that PNPLA3, or domains/regions of it, are unsuitable for traditional protein crystallography due to its biochemical and biophysical properties then there are a range of other techniques that could be used to structurally characterise it. Circular dichroism could be used to determine the level of folding of protein and its secondary structure features<sup>61</sup>. Solution state NMR also has the potential to determine the structure of regions of the protein such as lipid droplet binding motifs<sup>39</sup> that are intractable to crystallisation.

In summary, a computational analysis of PNPLA3 guided design of plasmid constructs, which led to the development of a high-yield protocol for the purification of PNPLA3. This analysis highlights several features of the PNPLA3 protein including a putative carboxyl terminal domain. Despite producing milligram quantities of protein, the sample was not homogeneous and hampered further crystallographic experiments. Work is ongoing to characterise this protein sample and those from the other PNPLA3 expression constructs for utility in protein crystallisation assays, functional assays and structure determination.

---

---

## CHAPTER 6 GENERAL DISCUSSION

---

---

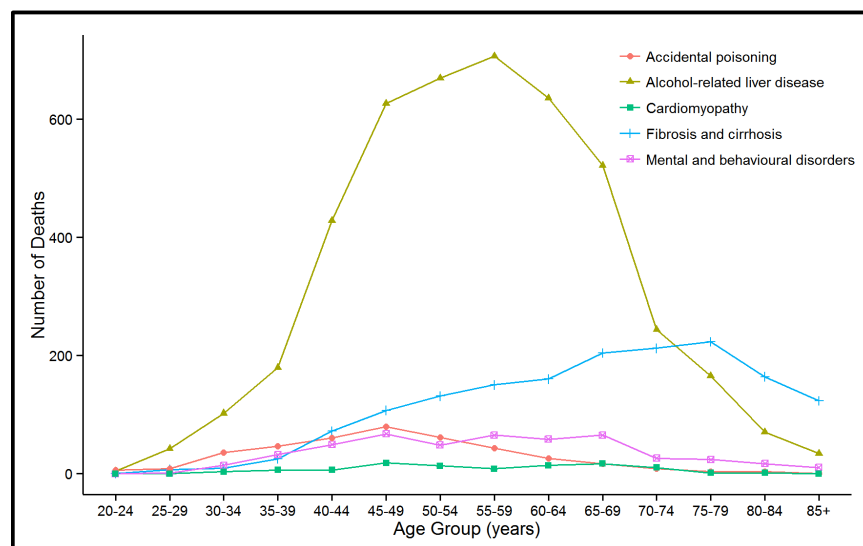


Liver disease is the third most common form of preventable death in the UK and in contrast to most other nations in Europe, its mortality rates are increasing. Liver disease is the only major cause of death in which such significant increases have occurred (Figure 6-1). The majority of liver disease deaths in the UK are attributable to alcohol-related liver disease. In 2012 in England and Wales, alcohol-related liver disease accounted for 63% (4,425) of alcohol-related deaths in whom the majority (31%) of deaths were among those aged 50-59 years<sup>315</sup> (Figure 6-2). Morbidity and mortality resulting from alcohol-related liver disease is a huge public health problem in the UK, which is only likely to worsen in the near future<sup>458</sup>.



**Figure 6-1 Standardized UK mortality rate data**  
100% in 1970. Image from Williams et al., 2014<sup>458</sup>

\*Data were normalised to



**Figure 6-2 Alcohol-related causes of death by age-group in England and Wales in 2012**  
Data from Office for National Statistics, 2014<sup>315</sup>

Only 10-20% of alcohol misusers develop the advanced stage alcohol-related liver disease, namely cirrhosis<sup>241</sup>. It unknown why there is this difference in inter-individual risk but it is clear that much of it is driven by complex interactions between the host and the environment<sup>411</sup>. Factors relating to alcohol consumption play a large role in risk<sup>28</sup> as do co-occurring causes of liver injury<sup>298</sup>, excess body-mass<sup>304</sup>, nutritional factors<sup>373</sup>, gender<sup>24</sup> and ethnicity<sup>102</sup>. It also seems that deprivation index impacts liver disease mortality<sup>29</sup>, in part explaining the differences in mortality across the UK (Figure 6-3).



**Figure 6-3 Directly standardized years of life lost due to chronic liver disease including cirrhosis per population by primary care trust**

There were significant increase in premature death from chronic liver disease between 1993 and 2010 in England and there is a 9-fold variation in deaths between primary care trusts. Image modified from the NHS Atlas of variation in healthcare<sup>50</sup>

For a long-time it has been known that heritable, or genetic factors contribute to alcohol-related cirrhosis risk<sup>185</sup>. However, the majority of genetic studies of alcohol-related cirrhosis are candidate gene studies, which cover a tiny fraction of known genetic variation<sup>411</sup>. Such studies are limited, at best, for characterizing the contribution of genetic variation to a complex disease phenotype when there are around 88 million genetic variants and around 24,000 genes in the human genome<sup>1</sup>.

There is now considerable evidence that *PNPLA3* and in particular carriage of the rs738409[G] allele confers significant risk towards the development of alcohol-related cirrhosis<sup>55,377,388,413,428,430</sup>. It has also been shown to be a significant risk factor for the development of hepatocellular carcinoma in patients with established alcohol-related cirrhosis<sup>122,164,309,430,432</sup>. It also seems that the rs738409[G] allele influences several aspects of alcohol-related liver disease progression and outcome. Thus carriage of the

rs738409[G] allele is associated with an earlier development of cirrhosis<sup>46</sup>; a reduction in transplantation free survival<sup>133</sup> and a poorer outcome following the development of hepatocellular carcinoma<sup>443</sup>.

## 6.1 - REVIEW OF FINDINGS

This thesis has covered several areas of study, which although thematic are disparate. Thus, a summary of the findings is provided to set the scene for a broader discussion.

### 6.1.1 - GWAS

This work contributes to the first GWAS<sup>44</sup> of alcohol-related cirrhosis through which variants in the genes *PNPLA3*, *TM6SF2* and *MBOAT7* were associated at genome-wide significance levels. The UCL cohort comprised a significant proportion of the alcohol-related cirrhosis cases (42.4%) and the phenotypically well-characterized no-significant liver injury controls (24.2%). A major finding from this study was the primacy of the genetic association between rs738409 in *PNPLA3* and alcohol-related cirrhosis risk. This variant has the largest effect on risk, when measured using odds ratios, of any of the variants in validated loci. This analysis also demonstrates the similarity between the genetic risk factors for alcohol-related cirrhosis and the NAFLD phenotype as both *PNPLA3*<sup>370</sup> and *TM6SF2*<sup>229</sup> are already known NAFLD risk loci.

### 6.1.2 - EXTENDED GWAS

An extended GWAS was performed via the inclusion of an additional number of severe alcoholic hepatitis cases recruited from the STOPAH trial. The harmonisation process, merging and re-analysis of this data revealed several novel variants nearing genome-wide significance levels while variants in *PNPLA3*, *TM6SF2* and *MBOAT7* also remained significantly associated. The validity of imputed genotype was confirmed via direct genotyping in the UCL and STOPAH cohorts, however, despite showing the same directionality in all instances none of these variants replicated. One of the loci identified via this analysis was between the genes *LIPG* and *ACAA2*, is also a validated cholesterol level quantitative trait locus<sup>11,52,410,416</sup>, a feature that it shares with the variant rs58542926 in *TM6SF2*.

### 6.1.3 - *PNPLA3* AND ALCOHOL-RELATED CIRRHOSIS RISK

The variant rs738409 was directly genotyped in the entire UCL cohort allowing comparisons between alcohol dependent cases, population controls and patients with and without alcohol-related cirrhosis. These comparisons demonstrate the primacy of the genetic association between rs738409 and alcohol-related cirrhosis and its independence from alcohol-dependence risk. In patients with alcohol-related cirrhosis

carriage of the rs738409[G] allele associated with several aspects of disease progression including time to presentation, time to death or liver transplant and the development of HCC. In patients with no-significant liver injury, carriage of the rs738409[G] allele was paradoxically associated with increased lifespan. A population attributable risk estimate for carriage of rs738409 on alcohol-related cirrhosis demonstrates that it is likely a substantial contributor to the overall incidence of cirrhosis in individuals of British and Irish ancestry who misuse alcohol.

#### **6.1.4 - PNPLA3: STRUCTURAL AND FUNCTIONAL STUDIES**

The PNPLA3 protein was investigated primarily for the purposes of structural characterization. Through in silico analysis of the protein sequence two domains were predicted, one of which is of unknown structural fold. This structural analysis guided the design of over forty different recombinant plasmids, which created for the purposes of high-yield protein expression in *E.coli*. High levels of protein expression were detected for several plasmid constructs, however, there was significant heterogeneity between the expected and experimentally determined molecular weight. One of these plasmid, which should express the entire PNPLA3 protein bound to an MBP tag, was selected for detailed characterization. This plasmid construct underwent DNA sequencing and the expressed protein mass spectrometry validating efficacious PNPLA3 expression from recombinant gene expression in *E.coli*. Milligram amounts of the PNPLA3 protein were purified, yet despite this, the protein did not purify in a homogeneous form suitable for further structural characterization.

### **6.2 - WHERE ARE WE GOING**

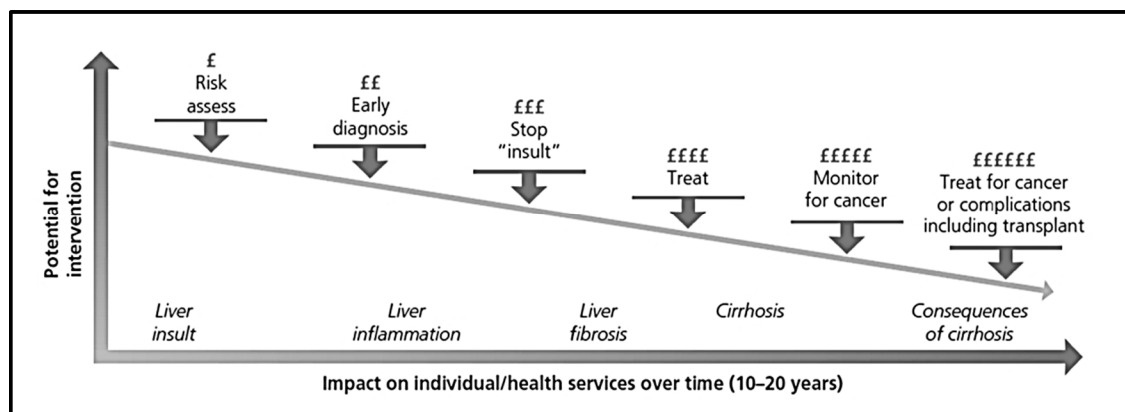
#### **6.2.1 - DETERMINING AT RISK GROUPS**

The major causes of liver-disease related deaths in the UK are alcohol, obesity and viral hepatitis. The UK now has the ability to eliminate the morbidity and mortality attributable to viral hepatitis, namely hepatitis B and hepatitis C<sup>458</sup>. However, due to the current alcohol misuse and obesity epidemics the prevalence of both alcohol-related and NAFLD related cirrhosis are expected to increase in the near future. Thus, beyond prevention strategies to reduce excess alcohol consumption, there is a need for other strategies to minimize morbidity and mortality attributable to alcohol-related liver disease.

The director of the National Health Service has recently published a strategy in which extensive plans are entailed for the use of genomic technologies in personalised medicine in the near future<sup>220</sup>. The genotyping of validated risk variants has a potential to contribute to this plan, either through use of genotype information in 'at risk' tests for

the prediction of disease and targeted prevention and early identification of a disease or for guiding an evidence based targeted intervention strategy. A paradigm for this already exists in the clinical management of hepatitis C viral infection. The variant rs8099917 in *IL28B* is associated with sustained virological response to hepatitis C virus, type 1 infection<sup>136</sup>. Although this information is not entirely predictive, and hence should not guide treatment entirely, the European Association for the Study of the Liver recommends that the genotyping of rs8099917 in *IL28B* 'in selected cases with genotype 1 [HCV], it may assist the physician and patient in management decisions'<sup>118</sup>.

There are two points in the natural history of alcohol-related liver disease in which rs738409 genotype could prove a useful biomarker: (i) the detection of liver disease before the development of cirrhosis; and, (ii) prognosis relating to complications of liver failure or the development of HCC following presentation with cirrhosis. As the progression of alcohol-related liver disease is associated with increasing treatment costs and decreasing potential for intervention, clearly the detection of liver damage before the development of cirrhosis is preferable (Figure 6-4).



**Figure 6-4 The potential for, and cost of, intervention during the course of liver disease**  
As liver disease progresses the opportunities for intervention diminish, whereas the relative costs of interventions that can be applied increase. Image modified from the NHS Atlas of variation in healthcare<sup>50</sup>

Liver function tests are often used in a primary care setting, although they are inefficient for the detection of early stage alcohol-related liver disease: up to 90% of people with early alcohol-related fibrosis and 75% of people with severe fibrosis have normal liver function test results<sup>166,391,458</sup>. In alcohol misusing patients, intervention is a cost effective means to reduce drinking and halt progression of alcohol-related liver disease and up to 50% of patients will stop drinking when informed by a specialist that they have cirrhosis<sup>444</sup>. For this reason there is a growing recognition for the role for the low cost non-invasive detection of the early stages of alcohol-related liver disease. In the UK, two models have been tested for this purpose: the Southampton Traffic Light (STL) model<sup>390,391</sup> in which the levels of serum biomarkers (hyaluronic acid, amino

terminal type III procollagen peptide and platelet count) are used to stratify and inform patients of their likelihood of developing cirrhosis followed by one year follow up for drinking cessation; and, the Nottingham study<sup>166</sup>, in which a serum AST:ALT ratio was used to stratify at risk groups (AST:ALT ratio  $\geq$  0.8) to undergo transient elastography<sup>330</sup> and, if significant liver stiffness was detected referral to a specialist hepatology clinic. These studies demonstrate that screening and intervention can reduce alcohol misuse in at risk groups and increase detection rates of alcohol-related liver disease. However, there are limitations, in particular the STL model and the ALT/AST ratio have moderate sensitivities for the detection of significant fibrosis, and could therefore waste screening resources due to false positives. For these purposes, rs738409 genotype information could be used to improve the performance of these detection algorithms, or develop new algorithms with the inclusion of other variables such as self-reported current alcohol intake, age, gender<sup>24</sup> or deprivation index<sup>29</sup>.

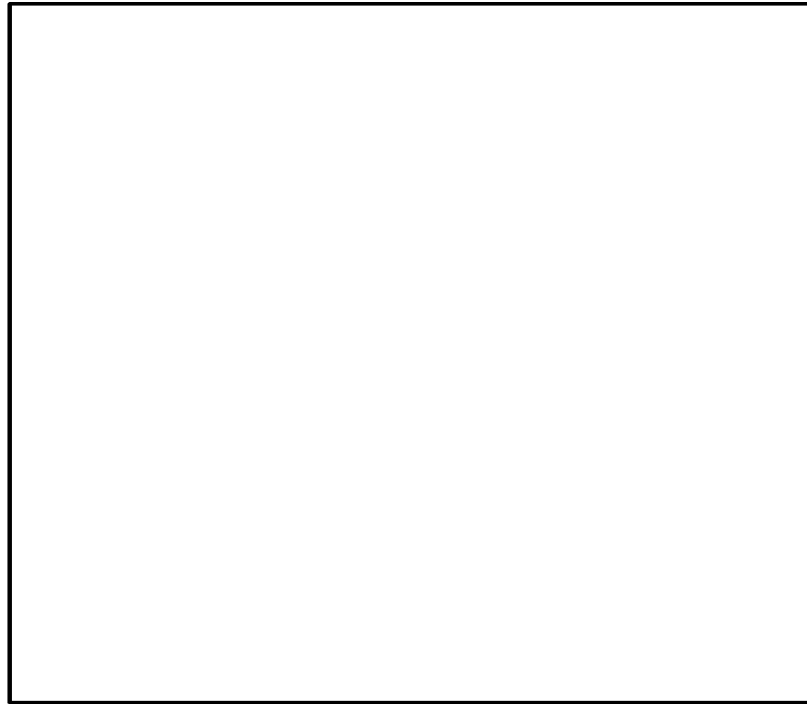
Cirrhosis is a primary risk factor for the development of HCC, which has an approximate annual incidence of 2-4% in patients with established cirrhosis<sup>272,458</sup>. The screening for HCC typically involves abdominal ultrasound scanning and measuring the serum biomarker  $\alpha$ -fetoprotein. This surveillance increases the chance of curative therapy and survival and in the UK NICE<sup>308</sup> will recommend screening at 6 monthly intervals, for every patient with cirrhosis. However, this screening has modest sensitivity and thus false negative results will occur<sup>397</sup>, which would be particularly detrimental to patients with rapidly growing tumours. As there is considerable evidence that rs738409[GG] homozygotes are more likely to develop HCC<sup>161,432</sup>, a case could be made for either more frequent screening of rs738409[GG] homozygotes or more costly, yet more accurate<sup>401</sup>, screening of homozygotes with magnetic resonance imaging.

Beyond rs738409 genotype per se there are many other genetic variants, which contribute to alcohol-related liver disease risk. For example, variants in *TM6SF2* and *MBOAT7* are associated with cirrhosis risk<sup>44</sup> and the variant rs1800562 in *HFE* is associated with HCC risk in alcohol-related cirrhosis<sup>200</sup>. Due to the moderate effect size of most genetic variants, they are unlikely to have significant predictive value by themselves. For this reason, it would be of greater value for their effects to be incorporated into multivariate models, potentially with the inclusion of validated gene-gene or gene-environment interaction effects. There is also growing enthusiasm for polygenic risk score analysis<sup>104</sup>, in which the contribution of genetic variants that do not meet traditional thresholds are incorporated into a polygenic model, which can be used to predict risk. There is a clear opportunity for further evaluation of genetic data to either stratify at-risk groups for targeted intervention or the monitoring of prognosis in alcohol-related liver disease.

## 6.2.2 - UNDERSTANDING DISEASE MECHANISMS

A feature of the GWAS of alcohol-related cirrhosis<sup>44</sup> presented in this work is the similarity in genetic risk factors with those identified in NAFLD such as rs738409 in *PNPLA3*<sup>370</sup> and rs58542926 in *TM6SF2*<sup>229,261</sup>. Conversely, the variant rs641738 in the *MBOAT7/TMC4* locus<sup>44</sup>, which is a novel alcohol-related cirrhosis risk locus, has recently been associated with steatosis grade, necroinflammation and fibrosis stage in NAFLD<sup>273</sup>. Based on the functions annotated to these genes, it seems likely that lipid metabolic pathways are important in both disease modalities. From a clinical perspective, it has long been recognised that both conditions share histopathological similarities<sup>4</sup>, with clinical differentiation based on excessive alcohol consumption, although the quantifiable extent to which genetic risk factors between these conditions is unknown. Polygenic risk score analysis could be used estimate the overlap in shared genetic risk factors using a bivariate linear mixed model as has been successfully applied between GWAS of psychiatric illness<sup>71</sup>. This type of analysis could also be used to compare shared genetic risk with phenotypes such as alcohol dependence<sup>139</sup> or other liver-related phenotypes<sup>53</sup>.

Due to their similarity, the paradigms that have emerged from genome-wide studies of NAFLD are also likely relevant in alcohol-related liver disease. From this perspective, the identification of the *TM6SF2* variant rs5854296 (Glu167Lys) provides an example through which a locus containing multiple non-independent genetic associations has been validated, refined and characterized. Originally, genetic associations at a broad region of linkage disequilibrium (chromosome 19p13.11) were ascribed to variants in genes such as *NCAN*<sup>149,405</sup> and *PPP1R3B*<sup>174</sup>. It was only following an exome-wide association study of liver fat content<sup>229</sup> and the functional characterization of genes in this region<sup>269</sup> that strong evidence was found implicating rs5854296 and *TM6SF2* as the most likely causal variant in this region. Subsequently, it has been demonstrated that this variant is strongly associated with fibrosis stage and moderately associated with HCC risk<sup>261</sup> and intriguingly, it has been demonstrated that this variant has opposite roles in cardiovascular disease and NAFLD risk<sup>101,337</sup> (Figure 6-5). However, it remains unknown whether all phenotypes associated with this chromosomal region are attributable to the *TM6SF2* rs5854296 variant or whether there are multiple independent genetic effects<sup>209</sup>. Unanswered questions regarding the association at *TM6SF2* may apply to other alcohol-related cirrhosis risk loci: are multiple functional variants contributing to risk in a single locus? Do the alleles, which protect against liver disease, increase risk for other disease phenotypes? Does the variant influence clinically relevant variables such as progression and the development of HCC?



**Figure 6-5 The opposing effects of the rs58542926 in *TM6SF2***

*TM6SF2* is functionally involved in the transport of VLDL from the liver. The alternate alleles of the variant rs58542926 have opposing effects on *TM6SF2* function. Abbreviations: chol - cholesterol; LDL - low-density lipoprotein cholesterol; IHTG - intrahepatic triglyceride; NASH - nonalcoholic steatohepatitis; TG - triglyceride; VLDL - very low-density lipoprotein. Image from Kahali et al., 2015<sup>209</sup>

With new cohorts undergoing GWAS, the identification of novel alcohol-related cirrhosis risk loci is almost inevitable. The progression from genetic association to understanding disease pathogenesis will require functional characterization of risk loci<sup>131</sup>. Following the example of other disease phenotypes, the functional studies that succeed GWAS will largely be empirical, and dependent on the context in which a variant occurs (e.g. protein coding region, intergenic, promoter region etc.). In silico techniques (e.g. fine mapping, expression quantitative trait loci, transcription factor binding site, epigenetic modifications etc.) and molecular biological techniques (e.g. tissue gene expression profiling, enzymatic assays, gene knockouts etc.) will probably be used to explore and validate functional hypotheses. The ease and utility of CRISPR/Cas9 genome editing<sup>70,270</sup> system has clear potential for characterizing the functional effects of genetic variants in model organisms and cell lines and will likely an increasingly important role in functional studies of variants implicated in alcohol-related cirrhosis risk.

### 6.2.3 - DEVELOPING NEW TREATMENTS

A number of pharmaceutical/nutritional therapies have been tested for efficacy in treating aspects of alcohol related liver disease<sup>264</sup>. These therapies include antibiotics, probiotics, antioxidants, anti-inflammatory compounds, stem cell progenitors and,



biliary acids. Despite the promise of these new treatments, there is an urgent need to develop new pathophysiology oriented therapies<sup>134</sup>. The genes and systems implicated via GWAS are a potential therapeutic target due to their robust association with the disease phenotype.

PNPLA3 is an obvious candidate for therapeutic intervention based on the functional evidence implicating it in lipid metabolism and robust genetic evidence implicating it with both NAFLD and alcohol-related cirrhosis risk. Accordingly, there are two patents pending regarding PNPLA3-related therapies: one for the genotyping of variants for predicting liver fat levels<sup>181</sup> and another for small molecule inhibitors and weight loss in obesity<sup>155</sup>. From a pharmaceutical perspective, it has been long known that certain organophosphorus compounds are potent inhibitors of serine hydrolase superfamily members<sup>203</sup> such as PNPLA3. These compounds inhibit patatin domain enzymatic activity via the formation of a covalent bond with the serine residue of the Ser-Asp catalytic dyad<sup>455</sup>. However, as the 148Met variant most likely results in a loss of enzymatic activity in the patatin domain<sup>400</sup> enzymatic inhibition per se is unlikely to alter pathogenic effects. Other therapeutic strategies could therefore aim to reduce its accumulation on lipid droplets or simply abrogate its gene expression. For example, liposome mediated RNA silencing of *PNPLA3* expression could hold potential<sup>78</sup>.

The function of PNPLA3 remains enigmatic, particularly the mechanisms through which the Ile148Met substitution effects fibrogenesis and the development of cirrhosis. Thus greater structural and functional characterization is still required to guide the development of therapeutics. Approaches to reduce the timespan and costs of drug development<sup>329</sup> such as structure based drug design and drug repurposing<sup>14</sup> could all hold potential for the development of PNPLA3-related therapies.

---

---

## REFERENCES

---

---

1. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491:56-65.
2. © The Trustees of the British Museum. Clay model of a sheep's liver. In; 2016.
3. Addison T. Observations on fatty degeneration of the liver. *Guy's Hospital Repository* 1836;1:476.
4. Adler M, Schaffner F. Fatty liver hepatitis and cirrhosis in obese patients. *Am J Med* 1979;67:811-6.
5. Agrawal A, Lynskey MT. Are there genetic influences on addiction: evidence from family, adoption and twin studies. *Addiction* 2008;103:1069-81.
6. Alderman J, Kato S, Lieber CS. The microsomal ethanol oxidizing system mediates metabolic tolerance to ethanol in deermice lacking alcohol dehydrogenase. *Arch Biochem Biophys* 1989;271:33-9.
7. Ali MA, Way MJ, Marks M, Guerrini I, Thomson AD, Strang J, McQuillin A, et al. Phenotypic heterogeneity in study populations may significantly confound the results of genetic association studies on alcohol dependence. *Psychiatr Genet* 2015;25:234-40.
8. Allgulander C, Nowak J, Rice JP. Psychopathology and treatment of 30,344 twins in Sweden. II. Heritability estimates of psychiatric diagnosis and treatment in 12,884 twin pairs. *Acta Psychiatr Scand* 1991;83:12-5.
9. Amato L, Minozzi S, Vecchi S, Davoli M. Benzodiazepines for alcohol withdrawal. *Cochrane Database Syst Rev* 2010;3:CD005063.
10. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*: American Psychiatric Association, 1994.
11. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* 2010;5:1564-73.
12. Anstee QM, Day CP. The genetics of NAFLD. *Nat Rev Gastroenterol Hepatol* 2013;10:645-55.
13. Anstee QM, Knapp S, Maguire EP, Hosie AM, Thomas P, Mortensen M, Bhome R, et al. Mutations in the *Gabrb1* gene promote alcohol consumption through increased tonic inhibition. *Nat Commun* 2013;4.
14. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;3:673-83.
15. Askgaard G, Grønbaek M, Kjær MS, Tjønneland A, Tolstrup JS. Alcohol drinking pattern and risk of alcoholic liver cirrhosis: A prospective cohort study. *J Hepatol* 2015;62:1061-67.

16. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BW, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 2009;41:47-55.
17. Baik I, Cho NH, Kim SH, Han BG, Shin C. Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men. *Am J Clin Nutr* 2011;93:809-16.
18. Baille M. *The Morbid Anatomy of some of the most important parts of the Human Body*. London: G. W. Nichols, 1808.
19. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res* 1999;27:49-54.
20. Basantani MK, Sitnick MT, Cai LZ, Brenner DS, Gardner NP, Li JZ, Schoiswohl G, et al. Pnpla3/Adiponutrin deficiency in mice does not contribute to fatty liver disease or metabolic syndrome. *J Lipid Res* 2011;52:318-29.
21. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, et al. The Pfam protein families database. *Nucleic Acids Res* 2004;32:D138-41.
22. Baulande S, Lasnier F, Lucas M, Pairault J. Adiponutrin, a transmembrane protein corresponding to a novel dietary- and obesity-linked mRNA specifically expressed in the adipose lineage. *J Biol Chem* 2001;276:33336-44.
23. Becker U, Grønbaek M, Johansen D, Sørensen TI. Lower risk for alcohol-induced cirrhosis in wine drinkers. *Hepatology* 2002;35:868-75.
24. Becker U, Deis A, Sørensen TI, Grønbaek M, Borch-Johnsen K, Müller CF, Schnohr P, et al. Prediction of risk of liver disease by alcohol intake, sex, and age: a prospective population study. *Hepatology* 1996;23:1025-9.
25. Bell H, Jahnsen J, Kittang E, Raknerud N, Sandvik L. Long-term prognosis of patients with alcoholic liver cirrhosis: a 15-year follow-up study of 100 Norwegian patients admitted to one unit. *Scand J Gastroenterol* 2004;39:858-63.
26. Bellamacina CR. The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins. *FASEB J* 1996;10:1257-69.
27. Bellentani S, Tiribelli C. The spectrum of liver disease in the general population: lesson from the Dionysos study. *J Hepatol* 2001;35:531-7.
28. Bellentani S, Saccoccio G, Costa G, Tiribelli C, Manenti F, Sodde M, Croce LS, et al. Drinking habits as cofactors of risk for alcohol induced liver damage. *Gut* 1997;41:845-50.
29. Bellis MA, Hughes K, Nicholls J, Sheron N, Gilmore I, Jones L. The alcohol harm paradox: using a national survey to explore how alcohol may disproportionately impact health in deprived individuals. *Bmc Public Health* 2016;16:111.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, et al. The protein data bank. *Nucleic Acids Res* 2000;28:235-42.

31. Berrow NS, Alderton D, Owens RJ: The precise engineering of expression vectors using high-throughput in-fusion™ PCR cloning. In: High Throughput Protein Expression and Purification: Springer, 2009; 75-90.
32. Bierut LJ, Dinwiddie SH, Begleiter H, Crowe RR, Hesselbrock V, Nurnberger JI, Jr., Porjesz B, et al. Familial transmission of substance dependence: alcohol, marijuana, cocaine, and habitual smoking: a report from the Collaborative Study on the Genetics of Alcoholism. *Arch Gen Psychiatry* 1998;55:982-8.
33. Bierut LJ, Goate AM, Breslau N, Johnson EO, Bertelsen S, Fox L, Agrawal A, et al. ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. *Mol Psychiatry* 2012;17:445-50.
34. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, et al. A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci U S A* 2010;107:5082-7.
35. Bird GL, O'Grady JG, Harvey FA, Calne RY, Williams R. Liver transplantation in patients with alcoholic cirrhosis: selection criteria and rates of survival and relapse. *BMJ* 1990;301:15-7.
36. Bird LE. High throughput construction and small scale expression screening of multi-tag vectors in *Escherichia coli*. *Methods* 2011;55:29-37.
37. Bird LE, Rada H, Flanagan J, Diprose JM, Gilbert RJC, Owens RJ: Application of In-Fusion™ cloning for the parallel construction of *E. coli* expression vectors. In: *DNA Cloning and Assembly Methods*: Springer, 2014; 209-34.
38. Bode C, Kugler V, Bode J. Endotoxemia in patients with alcoholic and non-alcoholic cirrhosis and in subjects with no evidence of chronic liver disease following acute alcohol excess. *J Hepatol* 1987;4:8-14.
39. Boeszoermenyi A, Nagy HM, Arthanari H, Phillip CJ, Lindermuth H, Luna RE, Wagner G, et al. Structure of a CGI-58 motif provides the molecular basis of lipid droplet anchoring. *J Biol Chem* 2015;290:26361-72.
40. Bohman M. Some genetic aspects of alcoholism and criminality: A population of adoptees. *Arch Gen Psychiatry* 1978;35:269-76.
41. Bohman M, Sigvardsson S, Cloninger CR. Maternal inheritance of alcohol abuse. Cross-fostering analysis of adopted women. *Arch Gen Psychiatry* 1981;38:965-9.
42. Brennan PL, Moos RH, Mertens JR. Personal and environmental risk factors as predictors of alcohol use, depression, and treatment-seeking: A longitudinal analysis of late-life problem drinkers. *J Subst Abuse* 1994;6:191-208.
43. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* 2009;85:847-61.

44. Buch S, Stickel F, Trepo E, Way M, Herrmann A, Nischalke HD, Brosch M, et al. A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. *Nat Genet* 2015;47:1443-8.
45. Buchner E. Alkoholische Gahrung ohne Hefezellen. In: *Berichte der deutschen chemischen Gesellschaft: WILEY-VCH Verlag*; 1897. p. 117-24.
46. Burza MA, Molinaro A, Attilia ML, Rotondo C, Attilia F, Ceccanti M, Ferri F, et al. PNPLA3 I148M (rs738409) genetic variant and age at onset of at-risk alcohol consumption are independent risk factors for alcoholic cirrhosis. *Liver Int* 2014;34:514-20.
47. Cadoret RJ, Cain CA, Grove WM. Development of alcoholism in adoptees raised apart from alcoholic biologic relatives. *Arch Gen Psychiatry* 1980;37:561-3.
48. Cadoret RJ, Troughton E, Woodworth G. Evidence of heterogeneity of genetic effect in Iowa adoption studies. *Ann N Y Acad Sci* 1994;708:59-71.
49. Cadoret RJ, Troughton E, O'Gorman TW, Heywood E. An adoption study of genetic and environmental factors in drug abuse. *Arch Gen Psychiatry* 1986;43:1131-6.
50. Care. NR. NHS Atlas of variation in healthcare. In: *Department of Health*; 2011.
51. Carter EA, Drummey GD, Isselbacher KJ. Ethanol stimulates triglyceride synthesis by the intestine. *Science* 1971;174:1245-7.
52. Chalasani N, Guo X, Loomba R, Goodarzi MO, Haritunians T, Kwon S, Cui J, et al. Genome-wide association study identifies variants associated with histologic features of nonalcoholic Fatty liver disease. *Gastroenterology* 2010;139:1567-76.
53. Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, Van der Harst P, Holm H, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* 2011;43:1131-8.
54. Chamorro AJ, Marcos M, Mir3n-Canelo JA, Pastor I, Gonzalez-Sarmiento R, Laso FJ. Association of  $\mu$ -opioid receptor (OPRM1) gene polymorphism with response to naltrexone in alcohol dependence: a systematic review and meta-analysis. *Addict Biol* 2012;17:505-12.
55. Chamorro AJ, Torres JL, Miron-Canelo JA, Gonzalez-Sarmiento R, Laso FJ, Marcos M. Systematic review with meta-analysis: the I148M variant of patatin-like phospholipase domain-containing 3 gene (PNPLA3) is significantly associated with alcoholic liver cirrhosis. *Aliment Pharmacol Ther* 2014;40:571-81.
56. Chamoun Z, Vacca F, Parton RG, Gruenberg J. PNPLA3/adiponutrin functions in lipid droplet formation. *Biol Cell* 2013;105:219-33.
57. Chan KY, Wong CM, Kwan JS, Lee JM, Cheung KW, Yuen MF, Lai CL, et al. Genome-wide association study of hepatocellular carcinoma in Southern Chinese patients with chronic hepatitis B virus infection. *PLoS One* 2011;6:e28798.

58. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015.
59. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, et al. Replicating genotype-phenotype associations. *Nature* 2007;447:655-60.
60. Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, et al. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet* 2009;5:e1000730.
61. Chemes LB, Alonso LG, Noval MG, de Prat-Gay G: Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains. In: *Intrinsically Disordered Protein Analysis*: Springer, 2012; 387-404.
62. Chen K, Shi W, Xin Z, Wang H, Zhu X, Wu X, Li Z, et al. Replication of genome wide association studies on hepatocellular carcinoma susceptibility loci in a Chinese population. *PLoS One* 2013;8:e77315.
63. Chick J, Gough K, Falkowski W, Kershaw P, Hore B, Mehta B, Ritson B, et al. Disulfiram treatment of alcoholism. *Br J Psychiatry* 1992;161:84-9.
64. Chick J, Anton R, Checinski K, Croop R, Drummond DC, Farmer R, Labriola D, et al. A multicentre, randomized, double-blind, placebo-controlled trial of naltrexone in the treatment of alcohol dependence or abuse. *Alcohol Alcohol* 2000;35:587-93.
65. Child CG, Turcotte JG. Surgery and portal hypertension. *Major Probl Clin Surg* 1964;1:1-85.
66. Christoffersen P, Nielsen K. Histological changes in human liver biopsies from chronic alcoholics. *Acta Pathol Microbiol Scand A* 1972;80:557-65.
67. Cloninger CR, Bohman M, Sigvardsson S. Inheritance of alcohol abuse. Cross-fostering analysis of adopted men. *Arch Gen Psychiatry* 1981;38:861-8.
68. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;10:101.
69. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, Butte NF. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS One* 2012;7:e51954.
70. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013;339:819-23.
71. Consortium. IS. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009;460:748-52.
72. Cornfield J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst* 1951;11:1269-75.

73. Cox DR. Regression Models and Life-Tables. *J. R. Stat. Soc.* 1972;34:187-220.
74. Crabb DW. Pathogenesis of alcoholic liver disease: newer mechanisms of injury. *Keio J Med* 1999;48:184-8.
75. Crabb DW, Edenberg HJ, Bosron WF, Li T-K. Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive ALDH2 (2) allele is dominant. *J Clin Invest* 1989;83:314.
76. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;39:D691-7.
77. Cucchetti A, Trevisani F, Pecorelli A, Erroi V, Farinati F, Ciccarese F, Rapaccini GL, et al. Estimation of lead-time bias and its impact on the outcome of surveillance for the early diagnosis of hepatocellular carcinoma. *J Hepatol* 2014;61:333-41.
78. Czech MP, Aouadi M, Tesz GJ. RNAi-based therapeutic strategies for metabolic disease. *Nat Rev Endocrinol* 2011;7:473-84.
79. Dadaev T. Concordance Typed and Imputed SNPs. In. GIT hub; 2016.
80. Dam-Larsen S, Franzmann M, Andersen IB, Christoffersen P, Jensen LB, Sørensen TI, Becker U, et al. Long term prognosis of fatty liver: risk of chronic liver disease and death. *Gut* 2004;53:750-5.
81. Dam MK, Flensburg-Madsen T, Eliassen M, Becker U, Tolstrup JS. Smoking and risk of liver cirrhosis: a population-based cohort study. *Scand J Gastroenterol* 2013;48:585-91.
82. Dancygier H. Clinical hepatology: Principles and practice of hepatobiliary diseases: Springer Science & Business Media, 2009.
83. Davies HTO, Crombie IK, Tavakoli M. When can odds ratios mislead? *BMJ* 1998;316:989-91.
84. Dawe F. O. Drinking (General Lifestyle Survey Overview-a report on the 2011 General Lifestyle Survey). In; 2013.
85. Dawson DA, Li TK, Grant BF. A prospective study of risk drinking: at risk for what? *Drug Alcohol Depend* 2008.
86. Day CP, Bassendine MF. Genetic predisposition to alcoholic liver disease. *Gut* 1992;33:1444-7.
87. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008;17:R122-8.
88. de Boer YS, van Gerven NM, Zwijs A, Verwer BJ, van Hoek B, van Erpecum KJ, Beuers U, et al. Genome-wide association study identifies variants associated with autoimmune hepatitis type 1. *Gastroenterology* 2014;147:443-52.

89. Deelen P, Bonder MJ, van der Velde KJ, Westra HJ, Winder E, Hendriksen D, Franke L, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res Notes* 2014;7:901.
90. Degoul F, Sutton A, Mansouri A, Capanec C, Degott C, Fromenty B, Beaugrand M, et al. Homozygosity for alanine in the mitochondrial targeting sequence of superoxide dismutase and risk for severe alcoholic liver disease. *Gastroenterology* 2001;120:1468-74.
91. Del Ben M, Polimeni L, Brancorsini M, Di Costanzo A, D'Erasmus L, Baratta F, Loffredo L, et al. Non-alcoholic fatty liver disease, metabolic syndrome and patatin-like phospholipase domain-containing protein3 gene variants. *Eur J Intern Med* 2014;25:566-70.
92. Delhaye M, Louis H, Degraef C, Le Moine O, Devière J, Gulbis B, Jacobovitz D, et al. Relationship between hepatocyte proliferative activity and liver functional reserve in human cirrhosis. *Hepatology* 1996;23:1003-11.
93. Dellarco VL. A mutagenicity assessment of acetaldehyde. *Mutat Res* 1988;195:1-20.
94. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, Wilson JF, et al. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet* 2012;8:e1002490.
95. Department of Health. Sensible drinking: Report of an inter-departmental working group. In; 1995. p. 89.
96. Dettling A, Fischer F, Bohler S, Ulrichs F, Skopp G, Graw M, Haffner HT. Ethanol elimination rates in men and women in consideration of the calculated liver weight. *Alcohol* 2007;41:415-20.
97. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;60:155-66.
98. DeWit DJ, Adlaf EM, Offord DR, Ogborne AC. Age at first alcohol use: a risk factor for the development of alcohol disorders. *Am J Psychiatry* 2014.
99. Dietler M. *Annu. Rev. Anthropol.* 2006;35:229-49.
100. DiStefano JK, Kingsley C, Craig Wood G, Chu X, Argyropoulos G, Still CD, Done SC, et al. Genome-wide analysis of hepatic lipid content in extreme obesity. *Acta Diabetol* 2015;52:373-82.
101. Dongiovanni P, Petta S, Maglio C, Fracanzani AL, Pipitone R, Mozzi E, Motta BM, et al. Transmembrane 6 superfamily member 2 gene variant disentangles nonalcoholic steatohepatitis from cardiovascular disease. *Hepatology* 2015;61:506-14.
102. Douds AC, Cox MA, Iqbal TH, Cooper BT. Ethnic differences in cirrhosis of the liver in a British city: alcoholic cirrhosis in South Asian men. *Alcohol Alcohol* 2003;38:148-50.



103. Dubuquoy C, Robichon C, Lasnier F, Langlois C, Dugail I, Fougelle F, Girard J, et al. Distinct regulation of adiponutrin/PNPLA3 gene expression by the transcription factors ChREBP and SREBP1c in mouse and human hepatocytes. *J Hepatol* 2011;55:145-53.
104. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013;9:e1003348.
105. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 2008;32:227-34.
106. Dutta AK. Genetic factors affecting susceptibility to alcoholic liver disease in an Indian population. *Ann Hepatol* 2013;12:901-7.
107. Edenberg HJ, Dick DM, Xuei X, Tian H, Almasy L. Variations in GABRA2, encoding the  $\alpha 2$  subunit of the GABA A receptor, are associated with alcohol dependence and with brain oscillations. *Am J Hum Genet* 2004;74:705-14.
108. Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintick JN, Almasy L, Bierut LJ, et al. Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcohol Clin Exp Res* 2010;34:840-52.
109. Edmondson HA, Peters RL, Frankel HH, Borowsky S. The early stage of liver injury in the alcoholic. *Medicine (Baltimore)* 1967;46:119-29.
110. Edwards G. Thomas Trotter's 'Essay on Drunkenness' appraised. *Addiction* 2012;107:1562-79.
111. Edwards G, Gross MM. Alcohol dependence: provisional description of a clinical syndrome. *Br Med J* 1976;1:1058-61.
112. Ehlers CL, Gilder DA, Wall TL, Phillips E, Feiler H, Wilhelmsen KC. Genomic screen for loci associated with alcohol dependence in Mission Indians. *Am J Med Genet B Neuropsychiatr Genet* 2004;129B:110-5.
113. El-Serag HB, Rudolph KL. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* 2007;132:2557-76.
114. Ellinghaus D, Folseraas T, Holm K, Ellinghaus E, Melum E, Balschun T, Laerdahl JK, et al. Genome-wide association analysis in primary sclerosing cholangitis and ulcerative colitis identifies risk loci at GPR35 and TCF4. *Hepatology* 2013;58:1074-83.
115. EMBL. Selecting the range of sequence to clone and express. In; 2015.
116. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57-74.
117. Erskine S, Maheswaran R, Pearson T, Gleeson D. Socioeconomic deprivation, urban-rural location and alcohol-related mortality in England and Wales. *Bmc Public Health* 2010;10:99.

118. European Association For The Study Of The Liver. EASL Clinical Practice Guidelines: management of hepatitis C virus infection. *J Hepatol* 2011;55:245-64.
119. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* 2013;14:379-89.
120. Evans DM, Purcell S. Power calculations in genetic studies. *Cold Spring Harb Protoc* 2012;2012:664-74.
121. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 2014;13:397-406.
122. Falletti E, Fabris C, Cmet S, Cussigh A, Bitetto D, Fontanini E, Fornasiere E, et al. PNPLA3 rs738409C/G polymorphism in cirrhosis: relationship with the aetiology of liver disease and hepatocellular carcinoma occurrence. *Liver Int* 2011;31:1137-43.
123. Fattovich G, Giustina G, Degos F, Tremolada F, Diodati G, Almasio P, Nevens F, et al. Morbidity and mortality in compensated cirrhosis type C: a retrospective follow-up study of 384 patients. *Gastroenterology* 1997;112:463-72.
124. Feitosa MF, Wojczynski MK, North KE, Zhang Q, Province MA, Carr JJ, Borecki IB. The ERLIN1-CHUK-CWF19L1 gene cluster influences liver fat deposition and hepatic inflammation in the NHLBI Family Heart Study. *Atherosclerosis* 2013;228:175-80.
125. Fischer J, Lefevre C, Morava E, Mussini JM, Laforet P, Negre-Salvayre A, Lathrop M, et al. The gene encoding adipose triglyceride lipase (PNPLA2) is mutated in neutral lipid storage disease with myopathy. *Nat Genet* 2007;39:28-30.
126. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, et al. Ensembl 2014. *Nucleic Acids Res* 2014;42:D749-55.
127. Folseraas T, Melum E, Rausch P, Juran BD, Ellinghaus E, Shiryaev A, Laerdahl JK, et al. Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J Hepatol* 2012;57:366-75.
128. Forbes RJ. *A short history of the art of distillation: from the beginnings up to the death of Cellier Blumenthal*: Brill, 1970.
129. Forrest E, Mellor J, Stanton L, Bowers M, Ryder P, Austin A, Day C, et al. Steroids or pentoxifylline for alcoholic hepatitis (STOPAH): study protocol for a randomised controlled trial. *Trials* 2013;14:262.
130. Frank J, Cichon S, Treutlein J, Ridinger M, Mattheisen M, Hoffmann P, Herms S, et al. Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. *Addict Biol* 2012;17:171-80.

131. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* 2011;43:513-8.
132. Freund G, Anderson KJ. Glutamate receptors in the cingulate cortex, hippocampus, and cerebellar vermis of alcoholics. *Alcohol Clin Exp Res* 1999;23:1-6.
133. Friedrich K, Wannhoff A, Kattner S, Brune M, Hov J, Weiss K, Antoni C, et al. *PNPLA3* in end-stage liver disease: Alcohol consumption, hepatocellular carcinoma development, and transplantation-free survival. *J Gastroenterol Hepatol* 2014;29:1477-84.
134. Gao B, Bataller R. Alcoholic liver disease: pathogenesis and new therapeutic targets. *Gastroenterology* 2011;141:1572-85.
135. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A: Protein Identification and Analysis Tools on the ExPASy Server. In: *Handbook TPP*, ed.: Humana Press Inc., 2005; 571-607.
136. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, et al. Genetic variation in *IL28B* predicts hepatitis C treatment-induced viral clearance. *Nature* 2009;461:399-401.
137. GE Healthcare. GE Healthcare. In; 2015.
138. Gelernter J, Kranzler HR, Panhuysen C, Weiss RD, Brady K, Poling J, Farrer L. Dense genomewide linkage scan for alcohol dependence in African Americans: significant linkage on chromosome 10. *Biol Psychiatry* 2009;65:111-5.
139. Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, Anton R, et al. Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci. *Mol Psychiatry* 2014;19:41-49.
140. Giegé R, Dock A, Kern D, Lorber B, Thierry J, Moras D. The role of purification in the crystallization of proteins and nucleic acids. *J Cryst Growth* 1986;76:554-61.
141. Gijón MA, Riekhof WR, Zarini S, Murphy RC, Voelker DR. Lysophospholipid acyltransferases and arachidonate recycling in human neutrophils. *J Biol Chem* 2008;283:30235-45.
142. Ginés P, Quintero E, Arroyo V, Terés J, Bruguera M, Rimola A, Caballería J, et al. Compensated cirrhosis: natural history and prognostic factors. *Hepatology* 1987;7:122-28.
143. Gizer IR, Ehlers CL, Vieten C, Seaton-Smith KL, Feiler HS, Lee JV, Segall SK, et al. Linkage scan of alcohol dependence in the UCSF Family Alcoholism Study. *Drug Alcohol Depend* 2011;113:125-32.
144. Glisson F. *Anatomia hepatis*. London, England: Du-Gardianis, 1654.

145. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;45:1274-83.
146. Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, O'Dushlaine C, et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* 2012;28:2543-5.
147. Goodwin DW, Schulsinger F, Hermansen L, Guze SB, Winokur G. Alcohol problems in adoptees raised apart from alcoholic biological parents. *Arch Gen Psychiatry* 1973;28:238-43.
148. Goodwin DW, Schulsinger F, Knop J, Mednick S, Guze SB. Alcoholism and depression in adopted-out daughters of alcoholics. *Arch Gen Psychiatry* 1977;34:751-5.
149. Gorden A, Yang R, Yerges-Armstrong LM, Ryan KA, Speliotes E, Borecki IB, Harris TB, et al. Genetic variation at NCAN locus is associated with inflammation and fibrosis in non-alcoholic fatty liver disease in morbid obesity. *Hum Hered* 2013;75:34-43.
150. Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA. Non-replication of association studies: "pseudo-failures" to replicate? *Genet Med* 2007;9:325-31.
151. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81:515-26.
152. Grant BF. ICD-10 and proposed DSM-IV harmful use of alcohol alcohol-abuse and dependence, United-States 1988 - a nosological comparison. *Alcohol Clin Exp Res* 1993;17:1093-101.
153. Greene CS, Penrod NM, Williams SM, Moore JH. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One* 2009;4:e5639.
154. Grittner U, Kuntsche S, Graham K, Bloomfield K. Social inequalities and gender differences in the experience of alcohol-related problems. *Alcohol Alcohol* 2012;47:597-605.
155. Gross RW, Jenkins CM, inventors; Human phospholipase A2 epsilon. 2013.
156. Grove J, Daly AK, Bassendine MF, Day CP. Association of a tumor necrosis factor promoter polymorphism with susceptibility to alcoholic steatohepatitis. *Hepatology* 1997;26:143-6.
157. Grove J, Daly AK, Bassendine MF, Gilvarry E, Day CP. Interleukin 10 promoter region polymorphisms and susceptibility to advanced alcoholic liver disease. *Gut* 2000;46:540-5.

158. Grove J, Daly AK, Burt AD, Guzail M, James OFW, Bassendine MF, Day CP. Heterozygotes for HFE mutations have no increased risk of advanced alcoholic liver disease. *Gut* 1998;43:262-66.
159. GSL Biotech. SnapGene software. In; 2015.
160. Gurling HM, Murray RM, Clifford CA. Investigations into the genetics of alcohol dependence and into its effects on brain function. *Prog Clin Biol Res.* 1981;69:77.
161. Guyot E, Sutton A, Rufat P, Laguillier C, Mansouri A, Moreau R, Ganne-Carrie N, et al. PNPLA3 rs738409, hepatocellular carcinoma occurrence and risk model prediction in patients with cirrhosis. *J Hepatol* 2013;58:312-8.
162. Halsted CH, Robles EA, Mezey E. Distribution of ethanol in the human gastrointestinal tract. *Am J Clin Nutr* 1973;26:831-4.
163. Hamilakis Y. Food technologies/technologies of the body: the social context of wine and oil production and consumption in Bronze Age Crete. *World Archaeol* 1999;31:38-54.
164. Hamza S, Petit JM, Masson D, Jooste V, Binquet C, Sgro C, Guiu B, et al. Pnpla3 rs738409 GG homozygote status is associated with increased risk of hepatocellular carcinoma in cirrhotic patients. *J Hepatol* 2012;56:S281-S82.
165. Hansell NK, Agrawal A, Whitfield JB, Morley KI, Gordon SD, Lind PA, Pergadia ML, et al. Linkage analysis of alcohol dependence symptoms in the community. *Alcohol Clin Exp Res* 2010;34:158-63.
166. Harman DJ, Ryder SD, James MW, Jelpke M, Ottey DS, Wilkes EA, Card TR, et al. Direct targeting of risk factors significantly increases the detection of liver cirrhosis in primary care: a cross-sectional diagnostic study utilising transient elastography. *BMJ Open* 2015;5:e007516.
167. Hart CL, Morrison DS, Batty GD, Mitchell RJ, Davey Smith G. Effect of body mass index and alcohol consumption on liver disease: analysis of data from two prospective cohort studies. *BMJ* 2010;340:c1240.
168. Hasin D, Li Q, Mccloud S, Endicott J. Agreement between DSM-III, DSM-III-R, DSM-IV and ICD-10 alcohol diagnoses in US community-sample heavy drinkers. *Addiction* 1996;91:1517-27.
169. Hassan MM, Kaseb A, Etzel CJ, El-Serag H, Spitz MR, Chang P, Hale KS, et al. Genetic variation in the PNPLA3 gene and hepatocellular carcinoma in USA: risk and prognosis prediction. *Mol Carcinog* 2013;52 Suppl 1:E139-47.
170. He S, McPhaul C, Li JZ, Garuti R, Kinch L, Grishin NV, Cohen JC, et al. A sequence variation (I148M) in PNPLA3 associated with nonalcoholic fatty liver disease disrupts triglyceride hydrolysis. *J Biol Chem* 2010;285:6706-15.
171. Health Committee. Written evidence from the Department of Health (GAS 01). In. [www.publications.parliament.uk](http://www.publications.parliament.uk); 2012.

172. Heath AC, Bucholz KK, Madden PA, Dinwiddie SH, Slutske WS, Bierut LJ, Statham DJ, et al. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. *Psychol Med* 1997;27:1381-96.
173. Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA, McEvoy BP, et al. A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. *Biol Psychiatry* 2011;70:513-8.
174. Hernaez R, McLean J, Lazo M, Brancati FL, Hirschhorn JN, Borecki IB, Harris TB, et al. Association between variants in or near PNPLA3, GCKR, and PPP1R3B with ultrasound-defined steatosis based on data from the third National Health and Nutrition Examination Survey. *Clin Gastroenterol Hepatol*. 2013;11:1183-90. e2.
175. Hernández-Nazará ZH, Ruiz-Madrigal B, Martínez-López E, Roman S, Panduro A. Association of the epsilon 2 allele of APOE gene to hypertriglyceridemia and to early-onset alcoholic cirrhosis. *Alcohol Clin Exp Res* 2008;32:559-66.
176. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58.
177. Hiroki A, Ohtake A, Yamamoto S, Satoh Y, Takayanagi M, Amaya Y, Takiguchi M, et al. Cloning and sequence analysis of a full length cDNA encoding human mitochondrial 3-oxoacyl-CoA thiolase. *BBA-Gene Struct Expr* 1993;1216:304-06.
178. Hirschberg HJ, Simons JW, Dekker N, Egmond MR. Cloning, expression, purification and characterization of patatin, a novel phospholipase A. *Eur J Biochem* 2001;268:5037-44.
179. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;4:45-61.
180. Hishikawa D, Shindou H, Kobayashi S, Nakanishi H, Taguchi R, Shimizu T. Discovery of a lysophospholipid acyltransferase family essential for membrane asymmetry and diversity. *Proc Natl Acad Sci U S A* 2008;105:2830-5.
181. Hobbs HH, Cohen JC, inventors; Sequence Variations in PNPLA3 Associated with Hepatic Steatosis. USA. 2010.
182. Hoek JB, Cahill A, Pastorino JG. Alcohol and mitochondria: a dysfunctional relationship. *Gastroenterology* 2002;122:2049-63.
183. Hoshida Y, Fuchs BC, Tanabe KK. Genomic risk of hepatitis C-related hepatocellular carcinoma. *J Hepatol* 2012;56:729-30.
184. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* 2011;1:457-70.
185. Hrubec Z, Omenn GS. Evidence of genetic predisposition to alcoholic cirrhosis and psychosis: twin concordances for alcoholism and its biological end points by zygosity among male veterans. *Alcohol Clin Exp Res* 1981;5:207-15.

186. Hu M, Wang F, Li X, Rogers CQ, Liang X, Finck BN, Mitra MS, et al. Regulation of hepatic lipin-1 by ethanol: role of AMP-activated protein kinase/sterol regulatory element-binding protein 1 signaling in mice. *Hepatology* 2012;55:437-46.
187. Huang Y, Cohen JC, Hobbs HH. Expression and characterization of a PNPLA3 protein isoform (I148M) associated with nonalcoholic fatty liver disease. *J Biol Chem* 2011;286:37085-93.
188. Huang Y, He S, Li JZ, Seo YK, Osborne TF, Cohen JC, Hobbs HH. A feed-forward loop amplifies nutritional regulation of PNPLA3. *Proc Natl Acad Sci U S A* 2010;107:7892-7.
189. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *J Comp Graph Stat* 1995:299-314.
190. Illumina®. Genome-Wide DNA Analysis BeadChips. In: Data Sheet: DNA Analysis; 2010.
191. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, Ramirez J, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet* 2011;7:e1002078.
192. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299-320.
193. Iorio KR, Tabakoff B, Hoffman PL. Glutamate-induced neurotoxicity is increased in cerebellar granule cells exposed chronically to ethanol. *Eur J Pharmacol* 1993;248:209-12.
194. Jablon S, Neel JV, Gershowitz H, Atkinson GF. The NAS-NRC twin panel: methods of construction of the panel, zygosity diagnosis, and proposed use. *Am J Hum Genet* 1967;19:133.
195. Järveläinen HA, Orpana A, Perola M, Savolainen VT, Karhunen PJ, Lindros KO. Promoter polymorphism of the CD14 endotoxin receptor gene as a risk factor for alcoholic liver disease. *Hepatology* 2001;33:1148-53.
196. Jaye M, Lynch KJ, Krawiec T, Marchadier D, Maugeais C, Doan K, South V, et al. A novel endothelial-derived lipase that modulates HDL metabolism. *Nat Genet* 1999;21:424-28.
197. Jellinek EM. The disease concept of alcoholism: New Haven, Hillhouse Press, 1960: 266.
198. Jenkins CM, Mancuso DJ, Yan W, Sims HF, Gibson B, Gross RW. Identification, cloning, expression, and purification of three novel human calcium-independent phospholipase A2 family members possessing triacylglycerol lipase and acylglycerol transacylase activities. *J Biol Chem* 2004;279:48968-75.

199. Jiang DK, Sun J, Cao G, Liu Y, Lin D, Gao YZ, Ren WH, et al. Genetic variants in STAT4 and HLA-DQ genes confer risk of hepatitis B virus-related hepatocellular carcinoma. *Nat Genet* 2013;45:72-5.
200. Jin F, Qu LS, Shen XZ. Association between C282Y and H63D mutations of the HFE gene with hepatocellular carcinoma in European populations: a meta-analysis. *J Exp Clin Cancer Res* 2010;29:18.
201. Johansson LE, Hoffstedt J, Parikh H, Carlsson E, Wabitsch M, Bondeson AG, Hedenbro J, et al. Variation in the adiponutrin gene influences its expression and associates with obesity. *Diabetes* 2006;55:826-33.
202. Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, Bierut LJ, Page GP. Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet* 2013;132:509-22.
203. Johnson MK. The delayed neurotoxic effect of some organophosphorus compounds. Identification of the phosphorylation site as an esterase. *Biochem J*. 1969;114:711-17.
204. Johnson SB, Gordon E, McClain C, Low G, Holman RT. Abnormal polyunsaturated fatty acid patterns of serum lipids in alcoholism and cirrhosis: arachidonic acid deficiency in cirrhosis. *Proc Natl Acad Sci U S A* 1985;82:1815-8.
205. Jones AW. Evidence-based survey of the elimination rates of ethanol from blood with applications in forensic casework. *Forensic Sci Int* 2010;200:1-20.
206. Jones AW, Andersson L. Influence of age, gender, and blood-alcohol concentration on the disappearance rate of alcohol from blood in drinking drivers. *J Forensic Sci* 1996;41:922-6.
207. Jun DW, Han JH, Jang EC, Kim SH, Kim SH, Jo YJ, Park YS, et al. Polymorphisms of microsomal triglyceride transfer protein gene and phosphatidylethanolamine N-methyltransferase gene in alcoholic and nonalcoholic fatty liver disease in Koreans. *Eur J Gastroenterol Hepatol* 2009;21:667-72.
208. June HL, Foster KL, McKay PF, Seyoum R, Woods JE, Harvey SC, Eiler WJ, et al. The reinforcing properties of alcohol are mediated by GABA(A1) receptors in the ventral pallidum. *Neuropsychopharmacology* 2003;28:2124-37.
209. Kahali B, Liu YL, Daly AK, Day CP, Anstee QM, Speliotes EK. TM6SF2: catch-22 in the fight against nonalcoholic fatty liver disease and cardiovascular disease? *Gastroenterology* 2015;148:679-84.
210. Kaij L. Alcoholism in twins: Studies on the etiology and sequels of abuse of alcohol: Almqvist & Wiksell, 1960.
211. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, Nakamura Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* 2010;42:210-5.



212. Kamatani Y, Wattanapokayakit S, Ochi H, Kawaguchi T, Takahashi A, Hosono N, Kubo M, et al. A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet* 2009;41:591-5.
213. Kamath PS, Kim WR. The model for end-stage liver disease (MELD). *Hepatology* 2007;45:797-805.
214. Kamper-Jørgensen M, Grønbaek M, Tolstrup J, Becker U. Alcohol and cirrhosis: dose–response or threshold effect? *J Hepatol* 2004;41:25-30.
215. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *JASA* 1958;53:457-81.
216. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, et al. The UCSC Genome Browser Database. *Nucleic Acids Res* 2003;31:51-4.
217. Kawaguchi T, Sumida Y, Umemura A, Matsuo K, Takahashi M, Takamura T, Yasui K, et al. Genetic polymorphisms of the human *PNPLA3* gene are strongly associated with severity of non-alcoholic fatty liver disease in Japanese. *PLoS One* 2012;7:e38322.
218. Kendler KS, Neale MC, Heath AC, Kessler RC, Eaves LJ. A twin-family study of alcoholism in women. *Am J Psychiatry* 1994;151:707-15.
219. Kennedy OJ, Roderick P, Buchanan R, Fallowfield JA, Hayes PC, Parkes J. Systematic review with meta-analysis: coffee consumption and the risk of cirrhosis. *Aliment Pharmacol Ther* 2016;43:562-74.
220. Keogh B. Personalised Medicine Strategy. In: England N, editor.; 2015.
221. Kershaw EE, Hamm JK, Verhagen LA, Peroni O, Katic M, Flier JS. Adipose triglyceride lipase: function, regulation by insulin, and comparison with adiponutrin. *Diabetes* 2006;55:148-57.
222. Kielhorn FW. The history of alcoholism: Bruhl-Cramer's concepts and observations. *Addiction* 1996;91:121-8.
223. Kienesberger PC, Oberer M, Lass A, Zechner R. Mammalian patatin domain containing proteins: a family with diverse lipolytic activities involved in multiple biological functions. *J Lipid Res* 2009;50 Suppl:S63-8.
224. Kim YJ, Go MJ, Hu C, Hong CB, Kim YK, Lee JY, Hwang JY, et al. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat Genet* 2011;43:990-5.
225. Kitamoto T, Kitamoto A, Yoneda M, Hyogo H, Ochi H, Nakamura T, Teranishi H, et al. Genome-wide scan revealed that polymorphisms in the *PNPLA3*, *SAMM50*, and *PARVB* genes are associated with development and progression of nonalcoholic fatty liver disease in Japan. *Hum Genet* 2013;132:783-92.
226. Klatsky AL, Armstrong MA. Alcohol, smoking, coffee, and cirrhosis. *Am J Epidemiol* 1992;136:1248-57.

227. Kondo F. Histological features of early hepatocellular carcinomas and their developmental process: for daily practical clinical application. *Hepatol Int*. 2009;3:283-93.
228. Koskenvuo M, Langinvainio H, Kaprio J, Lonnqvist J, Tienari P. Psychiatric hospitalization in twins. *Acta Genet Med Gemellol (Roma)* 1984;33:321-32.
229. Kozlitina J, Smagris E, Stender S, Nordestgaard BG, Zhou HH, Tybjaerg-Hansen A, Vogt TF, et al. Exome-wide association study identifies a *TM6SF2* variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2014;46:352-56.
230. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci* 2009;24:561-73.
231. Krebs HA, Freedland RA, Hems R, Stubbs M. Inhibition of hepatic gluconeogenesis by ethanol. *Biochem J* 1969;112:117-24.
232. Kumar V, Kato N, Urabe Y, Takahashi A, Muroyama R, Hosono N, Otsuka M, et al. Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. *Nat Genet* 2011;43:455-8.
233. Kumari M, Schoiswohl G, Chitraju C, Paar M, Cornaciu I, Rangrez AY, Wongsirirot N, et al. Adiponutrin functions as a nutritionally regulated lysophosphatidic acid acyltransferase. *Cell Metab* 2012;15:691-702.
234. Kumashiro N, Yoshimura T, Cantley JL, Majumdar SK, Guebre-Egziabher F, Kursawe R, Vatner DF, et al. Role of patatin-like phospholipase domain-containing 3 on lipid-induced hepatic steatosis and insulin resistance in rats. *Hepatology* 2013;57:1763-72.
235. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157:105-32.
236. Lake AC, Sun Y, Li JL, Kim JE, Johnson JW, Li D, Revett T, et al. Expression, regulation, and triglyceride hydrolase activity of Adiponutrin family members. *J Lipid Res* 2005;46:2477-87.
237. Lands WE. Metabolism of glycerolipides; a comparison of lecithin and triglyceride synthesis. *J Biol Chem* 1958;231:883-8.
238. Lazo M, Selvin E, Clark JM. Brief communication: clinical implications of short-term variability in liver function test results. *Ann Intern Med* 2008;148:348-52.
239. Leevy CM. Cirrhosis in alcoholics. *Med Clin North Am* 1968;52:1445-55.
240. Leibach WK. Cirrhosis in the alcoholic and its relation to the volume of alcohol abuse. *Ann N Y Acad Sci* 1975;252:85-105.
241. Leibach WK. Epidemiology of alcoholic liver disease. *Prog Liver Dis* 1976;5:494-515.

242. Leon DA, McCambridge J. Liver cirrhosis mortality rates in Britain from 1950 to 2002: an analysis of routine data. *Lancet* 2006;367:52-6.
243. Levine HG. The discovery of addiction. Changing conceptions of habitual drunkenness in America. *J Stud Alcohol* 1978;39:143-74.
244. Levy RE, Catana AM, Durbin-Johnson B, Halsted CH, Medici V. Ethnic differences in presentation and severity of alcoholic liver disease. *Alcohol Clin Exp Res* 2015;39:566-74.
245. Li D, Zhao H, Gelernter J. Strong association of the alcohol dehydrogenase 1B gene (*ADH1B*) with alcohol dependence and alcohol-induced medical diseases. *Biol Psychiatry* 2011;70:504-12.
246. Li D, Zhao H, Gelernter J. Strong protective effect of the aldehyde dehydrogenase gene (*ALDH2*) 504lys (\*2) allele against alcoholism and alcohol-induced medical diseases in Asians. *Hum Genet* 2012;131:725-37.
247. Li JZ, Huang Y, Karaman R, Ivanova PT, Brown HA, Roddy T, Castro-Perez J, et al. Chronic overexpression of PNPLA3 I148M in mouse liver causes hepatic steatosis. *J Clin Invest* 2012;122:4130-44.
248. Li S, Qian J, Yang Y, Zhao W, Dai J, Bei JX, Foo JN, et al. GWAS identifies novel susceptibility loci on 6p21.32 and 21q21.3 for hepatocellular carcinoma in chronic hepatitis B virus carriers. *PLoS Genet* 2012;8:e1002791.
249. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci U S A* 1987;84:6611-5.
250. Lieber CS. Perspectives: do alcohol calories count? *Am J Clin Nutr* 1991;54:976-82.
251. Lieber CS, DeCarli LM. Hepatic microsomal ethanol-oxidizing system. In vitro characteristics and adaptive properties in vivo. *J Biol Chem* 1970;245:2505-12.
252. Lieber CS, Decarli LM. Quantitative relationship between amount of dietary fat and severity of alcoholic fatty liver. *Am J Clin Nutr* 1970;23:474-&.
253. Lieber CS, Jones DP, Decarli LM. Effects of prolonged ethanol intake: production of fatty liver despite adequate Diets. *J Clin Invest* 1965;44:1009-21.
254. Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, de Moor MH, Smit AB, et al. A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations. *Twin Res Hum Genet* 2010;13:10-29.
255. Littleton J. Neurochemical mechanisms underlying alcohol withdrawal. *Alcohol Health Res World* 1998;22:13-24.
256. Liu J, Wang LN. Baclofen for alcohol withdrawal. *Cochrane Database of Systematic Reviews* 2015. *Cochrane Database Syst Rev* 2012;2:CD008502.

257. Liu J, Zeng FF, Liu ZM, Zhang CX, Ling WH, Chen YM. Effects of blood triglycerides on cardiovascular and all-cause mortality: a systematic review and meta-analysis of 61 prospective studies. *Lipids Health Dis* 2013;12:159.
258. Liu JZ, Almarri MA, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, Ducker SJ, et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* 2012;44:1137-41.
259. Liu LY, Fox CS, North TE, Goessling W. Functional validation of GWAS gene candidates for abnormal liver function during zebrafish liver development. *Dis Model Mech* 2013;6:1271-8.
260. Liu X, Invernizzi P, Lu Y, Kosoy R, Lu Y, Bianchi I, Podda M, et al. Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nat Genet* 2010;42:658-60.
261. Liu YL, Reeves HL, Burt AD, Tiniakos D, McPherson S, Leathart JB, Allison ME, et al. *TM6SF2* rs58542926 influences hepatic fibrosis progression in patients with non-alcoholic fatty liver disease. *Nat Commun* 2014;5:4309.
262. Long JC, Knowler WC, Hanson RL, Robin RW, Urbanek M, Moore E, Bennett PH, et al. Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. *Am J Med Genet* 1998;81:216-21.
263. Long JZ, Cravatt BF. The metabolic serine hydrolases and their functions in mammalian physiology and disease. *Chem Rev* 2011;111:6022-63.
264. Louvet A, Mathurin P. Alcoholic liver disease: mechanisms of injury and targeted treatment. *Nat Rev Gastroenterol Hepatol* 2015;12:231-42.
265. Lovinger DM, White G, Weight FF. Ethanol inhibits NMDA-activated ion current in hippocampal neurons. *Science* 1989;243:1721-4.
266. Ludwig J, Viggiano TR, McGill DB, Oh BJ. Nonalcoholic steatohepatitis: Mayo Clinic experiences with a hitherto unnamed disease. *Mayo Clin Proc* 1980;55:434-8.
267. Lukas SE, Mendelson JH, Benedikt RA. Instrumental analysis of ethanol-induced intoxication in human males. *Psychopharmacology (Berl)* 1986;89:8-13.
268. Maddrey WC, Boitnott JK, Bedine MS, Weber FL, Jr., Mezey E, White RI, Jr. Corticosteroid therapy of alcoholic hepatitis. *Gastroenterology* 1978;75:193-9.
269. Mahdessian H, Taxiarchis A, Popov S, Silveira A, Franco-Cereceda A, Hamsten A, Eriksson P, et al. *TM6SF2* is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. *Proc Natl Acad Sci U S A* 2014;111:8913-18.
270. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, et al. RNA-guided human genome engineering via Cas9. *Science* 2013;339:823-6.
271. Mallory FB. Cirrhosis of the liver. *Bull. Johns Hopkins Hosp.* 1911;22:69-74.

272. Mancebo A, González-Diéguez ML, Cadahía V, Varela M, Pérez R, Navascués CA, Sotorríos NG, et al. Annual incidence of hepatocellular carcinoma among patients with alcoholic cirrhosis and identification of risk groups. *Clin Gastroenterol Hepatol*. 2013;11:95-101.
273. Mancina RM, Dongiovanni P, Petta S, Pingitore P, Meroni M, Rametta R, Boren J, et al. The MBOAT7-TMC4 Variant rs641738 Increases Risk of Nonalcoholic Fatty Liver Disease in Individuals of European Descent. *Gastroenterology* 2016.
274. Mann K, Hermann D, Heinz A. One hundred years of alcoholism: the Twentieth Century. *Alcohol Alcohol* 2000;35:10-5.
275. Mann K, Lehert P, Morgan MY. The efficacy of acamprosate in the maintenance of abstinence in alcohol-dependent individuals: results of a meta-analysis. *Alcohol Clin Exp Res* 2004;28:51-63.
276. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-53.
277. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;11:499-511.
278. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906-13.
279. Marcos M, Pastor I, Gonzalez-Sarmiento R, Laso FJ. Interleukin-10 gene polymorphism is associated with alcoholism but not with alcoholic liver disease. *Alcohol Alcohol* 2008;43:523-8.
280. Marcos M, Pastor I, Gonzalez-Sarmiento R, Laso FJ. A functional polymorphism of the NFKB1 gene increases the risk for alcoholic liver cirrhosis in patients with alcohol dependence. *Alcohol Clin Exp Res* 2009;33:1857-62.
281. Marcos M, Gomez-Munuera M, Pastor I, Gonzalez-Sarmiento R, Laso FJ. Tumor necrosis factor polymorphisms and alcoholic liver disease: a HuGE review and meta-analysis. *Am J Epidemiol* 2009;170:948-56.
282. Marcos M, Pastor I, Chamorro AJ, Ciria-Abad S, Gonzalez-Sarmiento R, Laso FJ. Meta-analysis: glutathione-S-transferase allelic variants are associated with alcoholic liver disease. *Aliment Pharmacol Ther* 2011;34:1159-72.
283. Marshall AW, Kingstone D, Boss M, Morgan MY. Ethanol elimination in males and females: relationship to menstrual cycle and body composition. *Hepatology* 1983;3:701-6.
284. Mason JN, Eshleman AJ, Belknap JK, Crabbe JC, Loftis JM, Macey TA, Janowsky A. NMDA receptor subunit mRNA and protein expression in ethanol-withdrawal seizure-prone and -resistant mice. *Alcohol Clin Exp Res* 2001;25:651-60.

285. Mathurin P, Hadengue A, Bataller R, Addolorato G, Burra P, Burt A, Caballeria J, et al. EASL Clinical Practical Guidelines: Management of Alcoholic Liver Disease. *J Hepatol* 2012;57:399-420.
286. Mbarek H, Ochi H, Urabe Y, Kumar V, Kubo M, Hosono N, Takahashi A, et al. A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. *Hum Mol Genet* 2011;20:3884-92.
287. McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, Nunez A, et al. Fermented beverages of pre- and proto-historic China. *Proc Natl Acad Sci U S A* 2004;101:17593-8.
288. McManus S, Meltzer H, Brugha T, Bebbington P, Jenkins R. Adult psychiatric morbidity in England, 2007: results of a household survey. 2009.
289. Mells GF, Floyd JA, Morley KI, Cordell HJ, Franklin CS, Shin SY, Heneghan MA, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat Genet* 2011;43:329-32.
290. Melum E, Franke A, Schramm C, Weismuller TJ, Gotthardt DN, Offner FA, Juran BD, et al. Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nat Genet* 2011;43:17-9.
291. Mezey E. Ethanol metabolism and ethanol-drug interactions. *Biochem Pharmacol* 1976;25:869-75.
292. Michel RH, McGovern PE, Badler VR. The first wine & beer. Chemical detection of ancient fermented beverages. *Anal Chem* 1993;65:408A-13A.
293. Mihic SJ, Ye Q, Wick MJ, Koltchine VV, Krasowski MD, Finn SE, Mascia MP, et al. Sites of alcohol and volatile anaesthetic action on GABAA and glycine receptors. *Nature* 1997;389:385-89.
294. Mitchell JM, O'Neil JP, Janabi M, Marks SM, Jagust WJ, Fields HL. Alcohol consumption induces endogenous opioid release in the human orbitofrontal cortex and nucleus accumbens. *Sci Transl Med* 2012;4:116ra6.
295. Morgan MY, Sherlock S. Sex-related differences among 100 patients with alcoholic liver disease. *BMJ* 1977;1:939-41.
296. Morgan MY, Ritson B, Alcohol MCo, Staff MCoA. *Alcohol and Health: A Guide for Health-Care Professionals*: BPR Publishers, 2010.
297. Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* 1955;7:277-318.
298. Mueller S, Millonig G, Seitz HK. Alcoholic liver disease and hepatitis C: a frequently underestimated combination. *World J Gastroenterol* 2009;15:3462-71.
299. Mühleisen TW, Leber M, Schulze TG, Strohmaier J, Degenhardt F, Treutlein J, Mattheisen M, et al. Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat Commun* 2014;5:3339.

300. Murugesan S, Goldberg EB, Dou E, Brown WJ. Identification of diverse lipid droplet targeting motifs in the PNPLA family of triglyceride lipases. *PLoS One* 2013;8:e64950.
301. Nakamura M, Nishida N, Kawashima M, Aiba Y, Tanaka A, Yasunami M, Nakamura H, et al. Genome-wide association study identifies TNFSF15 and POU2AF1 as susceptibility loci for primary biliary cirrhosis in the Japanese population. *Am J Hum Genet* 2012;91:721-8.
302. National Collaborating Centre for Mental Health. Alcohol-use disorders: diagnosis, assessment and management of harmful drinking and alcohol dependence (CG115): RCPsych Publications, 2011.
303. National Health Service. Alcohol poisoning In; 2015.
304. Naveau S, Giraud V, Borotto E, Aubert A, Capron F, Chaput JC. Excess weight risk factor for alcoholic liver disease. *Hepatology* 1997;25:108-11.
305. NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2013;41:D8.
306. Nettleship JE, Assenberg R, Diprose JM, Rahman-Huq N, Owens RJ. Recent advances in the production of proteins in insect and mammalian cells for structural biology. *J Struct Biol* 2010;172:55-65.
307. Nettleship JE, Watson PJ, Rahman-Huq N, Fairall L, Posner MG, Upadhyay A, Reddivari Y, et al.: Transient expression in HEK 293 cells: an alternative to *E. coli* for the production of secreted and intracellular mammalian proteins. In: *Insoluble Proteins*: Springer, 2015; 209-22.
308. NICE. Assessment and Management of Cirrhosis. In: *NICE in development [GID-CGWAVE0683]*; 2015.
309. Nischalke HD, Berger C, Luda C, Berg T, Muller T, Grunhage F, Lammert F, et al. The PNPLA3 rs738409 148M/M genotype is a risk factor for liver cancer in alcoholic cirrhosis but shows no or weak association in hepatitis C cirrhosis. *PLoS One* 2011;6:e27087.
310. Nishida N, Sawai H, Matsuura K, Sugiyama M, Ahn SH, Park JY, Hige S, et al. Genome-wide association study confirming association of HLA-DP with protection against chronic hepatitis B and viral clearance in Japanese and Korean. *PLoS One* 2012;7:e39175.
311. Novy R, Drott D, Yaeger K, Mierendorf R. Overcoming the codon bias of *E. coli* for enhanced protein expression. *InNovations* 2001;12:1-3.
312. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* 2014;10:e1004234.

313. O'Shea RS, Dasarathy S, McCullough AJ, Practice Guideline Committee of the American Association for the Study of Liver Diseases Practice Parameters Committee of the American College of Gastroenterology. Alcoholic liver disease. *Hepatology* 2010;51:307-28.
314. OECD. Tackling Harmful Alcohol Use: Economics and public health policy. In: Sassi F, editor.: OECD Publishing; 2015.
315. Office for National Statistics. Alcohol-related deaths in the United Kingdom, registered in 2012. In; 2014.
316. Oliver J, Agundez JA, Morales S, Fernandez-Arquero M, Fernandez-Gutierrez B, de la Concha EG, Diaz-Rubio M, et al. Polymorphisms in the transforming growth factor-beta gene (TGF-beta) and the risk of advanced alcoholic liver disease. *Liver Int* 2005;25:935-9.
317. Oneta CM, Simanowski UA, Martinez M, Allali-Hassani A, Pares X, Homann N, Conradt C, et al. First pass metabolism of ethanol is strikingly influenced by the speed of gastric emptying. *Gut* 1998;43:612-19.
318. ONS. Population Estimates by Ethnic Group Experimental. In. <http://www.ons.gov.uk/ons/rel/peeg/population-estimates-by-ethnic-group--experimental/-current-estimates/population-estimates-by-ethnic-group-mid-2009--experimental--.zip>: ONS; 2009.
319. Orchard C. Adult Drinking Habits in Great Britain, 2013. In: ONS, editor.; 2015.
320. Pam A, Kemker SS, Ross CA, Golden R. The "equal environments assumption" in MZ-DZ twin comparisons: an untenable premise of psychiatric genetics? *Acta Genet Med Gemellol (Roma)* 1996;45:349-60.
321. Panés J, Caballería J, Guitart R, Parés A, Soler X, Rodamilans M, Navasa M, et al. Determinants of ethanol and acetaldehyde metabolism in chronic alcoholics. *Alcohol Clin Exp Res* 1993;17:48-53.
322. Papadopoulos JS, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics* 2007;23:1073-9.
323. Pare´ G, Ridker PM, Rose L, Barbalic M, Dupuis J, Dehghan A, Bis JC, et al. Genome-wide association analysis of soluble *ICAM-1* concentration reveals novel associations at the *NFKB1K*, *PNPLA3*, *RELA*, and *SH2B3* loci. *PLoS Genet* 2011;7:e1001374.
324. Parés A, Caballeria J, Bruguera M, Torres M, Rodes J. Histological course of alcoholic hepatitis. Influence of abstinence, sex and extent of hepatic damage. *J Hepatol* 1986;2:33-42.
325. Park BL, Kim JW, Cheong HS, Kim LH, Lee BC, Seo CH, Kang TC, et al. Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: from GWAS to replication. *Hum Genet* 2013;132:657-68.



326. Partanen J, Bruun K, Markkanen T. Inheritance of drinking behavior: A study on intelligence, personality, and use of alcohol of adult twins: Finnish Foundation for Alcohol Studies, 1966.
327. Pasteur L. Mémoire sur la fermentation appliquée lactique. Mémoire sur la fermentation alcoolique. In: Mallet-Bachelier; 1857.
328. Patin E, Kutalik Z, Guergnon J, Bibert S, Nalpas B, Jouanguy E, Munteanu M, et al. Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. *Gastroenterology* 2012;143:1244-52 e1-12.
329. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;9:203-14.
330. Pavlov CS, Casazza G, Nikolova D, Tsochatzis E, Burroughs AK, Ivashkin VT, Gluud C. Transient elastography for diagnosis of stages of hepatic fibrosis and cirrhosis in people with alcoholic liver disease. *Cochrane Database Syst Rev* 2015;1:CD010542.
331. Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. Analyses and comparison of accuracy of different genotype imputation methods. *PLoS One* 2008;3:e3551.
332. Perez Y, Maffei M, Igea A, Amata I, Gairi M, Nebreda AR, Bernado P, et al. Lipid binding by the Unique and SH3 domains of c-Src suggests a new regulatory mechanism. *Sci Rep* 2013;3:1295.
333. Pickens RW, Svikis DS, McGue M, Lykken DT, Heston LL, Clayton PJ. Heterogeneity in the inheritance of alcoholism. A study of male and female twins. *Arch Gen Psychiatry* 1991;48:19-28.
334. Pingitore P, Pirazzi C, Mancina RM, Motta BM, Indiveri C, Pujia A, Montalcini T, et al. Recombinant PNPLA3 protein shows triglyceride hydrolase activity and its I148M mutation results in loss of function. *Biochim Biophys Acta* 2014;1841:574-80.
335. Pirazzi C, Adiels M, Burza MA, Mancina RM, Levin M, Stahlman M, Taskinen MR, et al. Patatin-like phospholipase domain-containing 3 (PNPLA3) I148M (rs738409) affects hepatic VLDL secretion in humans and in vitro. *J Hepatol* 2012;57:1276-82.
336. Pirazzi C, Valenti L, Motta BM, Pingitore P, Hedfalk K, Mancina RM, Burza MA, et al. PNPLA3 has retinyl-palmitate lipase activity in human hepatic stellate cells. *Hum Mol Genet* 2014;23:4077-85.
337. Pirola CJ, Sookoian S. The dual and opposite role of the TM6SF2-rs58542926 variant in protecting against cardiovascular disease and conferring risk for nonalcoholic fatty liver: A meta-analysis. *Hepatology* 2015;62:1742-56.
338. Png E, Thalamuthu A, Ong RT, Snippe H, Boland GJ, Seielstad M. A genome-wide association study of hepatitis B vaccine response in an Indonesian population

- reveals multiple independent risk variants in the HLA region. *Hum Mol Genet* 2011;20:3893-8.
339. Poikolainen K. Risk factors for alcohol dependence: a case-control study. *Alcohol Alcohol* 2000;35:190-6.
340. Potts JR, Verma S. Alcoholic hepatitis: diagnosis and management in 2012. *Expert Rev Gastroenterol Hepatol* 2012;6:695-710.
341. Power C, Rasko JE. Whither Prometheus' liver? Greek myth and the science of regeneration. *Ann Intern Med* 2008;149:421-6.
342. Prescott CA, Kendler KS. Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. *Am J Psychiatry* 1999;156:34-40.
343. Prescott CA, Caldwell CB, Carey G, Vogler GP, Trumbetta SL, Gottesman, II. The Washington University Twin Study of alcoholism. *Am J Med Genet B Neuropsychiatr Genet* 2005;134B:48-55.
344. Prescott CA, Sullivan PF, Kuo PH, Webb BT, Vittum J, Patterson DG, Thiselton DL, et al. Genomewide linkage study in the Irish affected sib pair study of alcohol dependence: evidence for a susceptibility region for symptoms of alcohol dependence on chromosome 4. *Mol Psychiatry* 2006;11:603-11.
345. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904-9.
346. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, Ge DL, et al. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* 2008;83:132-35.
347. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;26:2336-7.
348. Pugh RN, Murray-Lyon IM, Dawson JL, Pietroni MC, Williams R. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg* 1973;60:646-9.
349. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-75.
350. Purohit V, Gao B, Song BJ. Molecular mechanisms of alcoholic fatty liver. *Alcohol Clin Exp Res* 2009;33:191-205.
351. Qu LS, Jin F, Guo YM, Liu TT, Xue RY, Huang XW, Xu M, et al. Nine susceptibility loci for hepatitis B virus-related hepatocellular carcinoma identified by a pilot two-stage genome-wide association study. *Oncol Lett* 2016;11:624-32.

352. R Core Team. R: A language and environment for statistical computing. In: ISBN 3-900051-07-0; 2012.
353. Racusen D, Foote M. A Major Soluble Glycoprotein of Potato-Tubers. *J Food Biochem* 1980;4:43-52.
354. Rao R. Endotoxemia and gut barrier dysfunction in alcoholic liver disease. *Hepatology* 2009;50:638-44.
355. Rauch A, Kutalik Z, Descombes P, Cai T, Di Iulio J, Mueller T, Bochud M, et al. Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. *Gastroenterology* 2010;138:1338-45, 45 e1-7.
356. Rausch V, Peccerella T, Seitz HK, Stickel F, Yagmur E, Herzig S, Mueller S. Histological hepatocyte damage precedes steatosis in ALD patients with genetic variant I148 M in PNPLA3. *Z Gastroenterol* 2013;51:K118.
357. Raynard B, Balian A, Fallik D, Capron F, Bedossa P, Chaput JC, Naveau S. Risk factors of fibrosis in alcohol-induced liver disease. *Hepatology* 2002;35:635-8.
358. Reed T, Page WF, Viken RJ, Christian JC. Genetic predisposition to organ-specific endpoints of alcoholism. *Alcohol Clin Exp Res* 1996;20:1528-33.
359. Rehm J, Roerecke M. Patterns of drinking and liver cirrhosis - what do we know and where do we go? *J Hepatol* 2015;62:1000-1.
360. Rehm J, Samokhvalov AV, Shield KD. Global burden of alcoholic liver diseases. *J Hepatol* 2013;59:160-8.
361. Rehm J, Mathers C, Popova S, Thavorncharoensap M, Teerawattananon Y, Patra J. Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. *Lancet* 2009;373:2223-33.
362. Rehm J, Taylor B, Mohapatra S, Irving H, Baliunas D, Patra J, Roerecke M. Alcohol as a risk factor for liver cirrhosis: A systematic review and meta-analysis. *Drug Alcohol Rev* 2010;29:437-45.
363. Reich T, Cloninger CR, Van Eerdewegh P, Rice JP, Mullaney J. Secular trends in the familial transmission of alcoholism. *Alcohol Clin Exp Res* 1988;12:458-64.
364. Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, Foroud T, et al. Genome-wide search for genes affecting the risk for alcohol dependence. *Am J Med Genet* 1998;81:207-15.
365. Ricciardelli LA, Connor JP, Williams RJ, Young RM. Gender stereotypes and drinking cognitions as indicators of moderate and high risk drinking among young women and men. *Drug Alcohol Depend* 2001;61:129-36.
366. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, Bergen SE, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013;45:1150-9.
367. Roche. LightCycler® 480 Software release 1.5.0. In; 2004.

368. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health* 1998;88:15-19.
369. Roerecke M, Rehm J. Irregular heavy drinking occasions and risk of ischemic heart disease: a systematic review and meta-analysis. *Am J Epidemiol* 2010;171:633-44.
370. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, Boerwinkle E, et al. Genetic variation in *PNPLA3* confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2008;40:1461-5.
371. Roshyara NR, Scholz M. fcGENE: a versatile tool for processing and transforming SNP datasets. *PLoS One* 2014;9:e97589.
372. Rossit AR, Cabral IR, Hackel C, da Silva R, Froes ND, Abdel-Rahman SZ. Polymorphisms in the DNA repair gene *XRCC1* and susceptibility to alcoholic liver cirrhosis in older Southeastern Brazilians. *Cancer Lett* 2002;180:173-82.
373. Rotily M, Durbec JP, Berthezene P, Sarles H. Diet and alcohol in liver cirrhosis: a case-control study. *Eur J Clin Nutr* 1990;44:595-603.
374. Rottenberg H, Waring A, Rubin E. Tolerance and cross-tolerance in chronic alcoholics: reduced membrane binding of ethanol and other drugs. *Science* 1981;213:583-5.
375. Ruhanen H, Perttinen J, Holtta-Vuori M, Zhou Y, Yki-Jarvinen H, Ikonen E, Kakela R, et al. *PNPLA3* mediates hepatocyte triacylglycerol remodeling. *J Lipid Res* 2014;55:739-46.
376. Rydel TJ, Williams JM, Krieger E, Moshiri F, Stallings WC, Brown SM, Pershing JC, et al. The crystal structure, mutagenesis, and activity studies reveal that patatin is a lipid acyl hydrolase with a Ser-Asp catalytic dyad. *Biochemistry* 2003;42:6696-708.
377. Salameh H, Raff E, Erwin A, Seth D, Nischalke HD, Falletti E, Burza MA, et al. *PNPLA3* Gene Polymorphism Is Associated With Predisposition to and Severity of Alcoholic Liver Disease. *Am J Gastroenterol* 2015;110:846-56.
378. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;74:5463-7.
379. Saunders JB, Walters JR, Davies AP, Paton A. A 20-year prospective study of cirrhosis. *Br Med J (Clin Res Ed)* 1981;282:263-6.
380. Savolainen VT, Pjarinen J, Perola M, Penttinen A, Karhunen PJ. Glutathione-S-transferase *GST M1* "null" genotype and the risk of alcoholic liver disease. *Alcohol Clin Exp Res* 1996;20:1340-5.
381. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014;511:421-27.

382. Schlegel S, Löfblom J, Lee C, Hjelm A, Klepsch M, Strous M, Drew D, et al. Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21 (DE3). *J Mol Biol* 2012;423:648-59.
383. Schrödinger L. The PyMOL molecular graphics system. In. Version 1.7.4 ed; 2002.
384. Schumann G, Coin LJ, Lourdasamy A, Charoen P, Berger KH, Stacey D, Desrivieres S, et al. Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proc Natl Acad Sci U S A* 2011;108:7119-24.
385. Sebastiani P, Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, et al. Genetic signatures of exceptional longevity in humans. *Science* 2010;2010:1190532.
386. Sedgwick P. Bias in observational study designs: case-control studies. *BMJ* 2015;350:h560.
387. Sela M, White FH, Jr., Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. *Science* 1957;125:691-2.
388. Seth D, Daly AK, Haber PS, Day CP. Patatin-like phospholipase domain containing 3: a case in point linking genetic susceptibility for alcoholic and nonalcoholic liver disease. *Hepatology* 2010;51:1463-5.
389. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014;15:335-46.
390. Sheron N, Moore M, Ansett S, Parsons C, Bateman A. Developing a 'traffic light' test with potential for rational early diagnosis of liver fibrosis and cirrhosis in the community. *Br J Gen Pract* 2012;62.
391. Sheron N, Moore M, O'Brien W, Harris S, Roderick P. Feasibility of detection and intervention for alcohol-related liver disease in the community: the Alcohol and Liver Disease Detection study (ALDDeS). *Br J Gen Pract* 2013;63:e698-705.
392. Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* 2006;1:2856-60.
393. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014;46:543-50.
394. Shukla SD, Pruett SB, Szabo G, Arteel GE. Binge ethanol and liver: new molecular developments. *Alcohol Clin Exp Res* 2013;37:550-7.
395. Sigel E, Steinmann M. Structure, function, and modulation of GABA<sub>A</sub> receptors. *J Biol Chem* 2012;287:40224-31.

396. Sigvardsson S, Bohman M, Cloninger CR. Replication of the Stockholm Adoption Study of alcoholism. Confirmatory cross-fostering analysis. *Arch Gen Psychiatry* 1996;53:681-7.
397. Singal A, Volk ML, Waljee A, Salgia R, Higgins P, Rogers MA, Marrero JA. Meta-analysis: surveillance with ultrasound for early-stage hepatocellular carcinoma in patients with cirrhosis. *Aliment Pharmacol Ther* 2009;30:37-47.
398. Singal AG, Manjunath H, Yopp AC, Beg MS, Marrero JA, Gopal P, Waljee AK. The effect of PNPLA3 on fibrosis progression and development of hepatocellular carcinoma: a meta-analysis. *Am J Gastroenterol* 2014;109:325-34.
399. Singh S, Murad MH, Chandar AK, Bongiorno CM, Singal AK, Atkinson SR, Thursz MR, et al. Comparative Effectiveness of Pharmacological Interventions for Severe Alcoholic Hepatitis: A Systematic Review and Network Meta-analysis. *Gastroenterology* 2015;149:958-70 e12.
400. Smagris E, BasuRay S, Li J, Huang Y, Lai KM, Gromada J, Cohen JC, et al. Pnpla3<sup>1148M</sup> knockin mice accumulate PNPLA3 on lipid droplets and develop hepatic steatosis. *Hepatology* 2015;61:108-18.
401. Snowberger N, Chinnakotla S, Lepe RM, Peattie J, Goldstein R, Klintmalm GB, Davis GL. Alpha fetoprotein, ultrasound, computerized tomography and magnetic resonance imaging for detection of hepatocellular carcinoma in patients with advanced cirrhosis. *Aliment Pharmacol Ther* 2007;26:1187-94.
402. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 2013;14:483-95.
403. Sørensen HT, Friis S, Olsen JH, Thulstrup AM, Møller M, Linet M, Trichopoulos D, et al. Risk of liver and other types of cancer in patients with cirrhosis: a nationwide cohort study in Denmark. *Hepatology* 1998;28:921-25.
404. Sørensen TA, Bentsen KD, Eghøj K, Orholm M, Højbye G, Offersen P. Prospective evaluation of alcohol abuse and alcoholic liver injury in men as predictors of development of cirrhosis. *Lancet* 1984;324:241-44.
405. Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, Gudnason V, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* 2011;7:e1001324.
406. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;5:e1000477.
407. Splettstoesser T. Workflow for solving the structure of a molecule by X-ray crystallography. In; 2015.

408. Sprengell CJ. The Aphorisms of Hippocrates and the Sentences of Celsus, with Explanations ... To which are Added Aphorisms Upon the Small-pox, Measles, and Other Distempers, 1735.
409. Stagos D, Chen Y, Brocker C, Donald E, Jackson BC, Orlicky DJ, Thompson DC, et al. Aldehyde dehydrogenase 1B1: molecular cloning and characterization of a novel mitochondrial acetaldehyde-metabolizing enzyme. *Drug Metab Dispos* 2010;38:1679-87.
410. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 2009;10:681-90.
411. Stickel F, Hampe J. Genetic determinants of alcoholic liver disease. *Gut* 2012;61:150-9.
412. Stickel F, Osterreicher CH, Halangk J, Berg T, Homann N, Hellerbrand C, Patsenker E, et al. No role of matrixmetalloproteinase-3 genetic promoter polymorphism 1171 as a risk factor for cirrhosis in alcoholic liver disease. *Alcohol Clin Exp Res* 2008;32:959-65.
413. Stickel F, Buch S, Lau K, Meyer zu Schwabedissen H, Berg T, Ridinger M, Rietschel M, et al. Genetic variation in the PNPLA3 gene is associated with alcoholic liver injury in caucasians. *Hepatology* 2011;53:86-95.
414. Stika HP. Traces of a possible Celtic brewery in Eberdingen-Hochdorf, Kreis Ludwigsburg, southwest Germany. *Veg Hist Archaeobot* 1996;5:81-88.
415. Strickland JA, Orr GL, Walsh TA. Inhibition of Diabrotica Larval Growth by Patatin, the Lipid Acyl Hydrolase from Potato Tubers. *Plant Physiol* 1995;109:667-74.
416. Striegel A, Yau WW, Kirkland JJ, Bly DD. Modern size-exclusion liquid chromatography: practice of gel permeation and gel filtration chromatography: John Wiley & Sons, 2009.
417. Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, Bassendine M, et al. IL28B is associated with response to chronic hepatitis C interferon- $\alpha$  and ribavirin therapy. *Nat Genet* 2009;41:1100-04.
418. Takamatsu M, Yamauchi M, Maezawa Y, Ohata M, Saitoh S, Toda G. Correlation of a polymorphism in the interleukin-1 receptor antagonist gene with hepatic fibrosis in Japanese alcoholics. *Alcohol Clin Exp Res* 1998;22:141S-44S.
419. Takamatsu M, Yamauchi M, Maezawa Y, Saito S, Maeyama S, Uchikoshi T. Genetic polymorphisms of interleukin-1 $\beta$  in association with the development of alcoholic liver disease in Japanese patients. *Am J Gastroenterol* 2000;95:1305-11.
420. Tam J, Liu J, Mukhopadhyay B, Cinar R, Godlewski G, Kunos G. Endocannabinoids in liver disease. *Hepatology* 2011;53:346-55.
421. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, et al. Genome-wide association of IL28B with response to pegylated

- interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nat Genet* 2009;41:1105-9.
422. Taylor B, Irving HM, Kanteres F, Room R, Borges G, Cherpitel C, Greenfield T, et al. The more you drink, the harder you fall: A systematic review and meta-analysis of how acute alcohol consumption and injury or collision risk increase together. *Drug Alcohol Depend* 2010;110:108-16.
423. Teli MR, Day CP, Burt AD, Bennett MK, James OF. Determinants of progression to cirrhosis or fibrosis in pure alcoholic fatty liver. *Lancet* 1995;346:987-90.
424. The National Center for Biotechnology Information. Basic Local Alignment Search Tool®. In; 2015.
425. The Science and Technology Committee. Alcohol guidelines. In: London: The Stationery Office Limited 2012.
426. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model: Springer-Verlag New York, 2000.
427. Thursz MR, Richardson P, Allison M, Austin A, Bowers M, Day CP, Downs N, et al. Prednisolone or Pentoxifylline for Alcoholic Hepatitis. *N Engl J Med* 2015;372:1619-28.
428. Tian C, Stokowski RP, Kershenovich D, Ballinger DG, Hinds DA. Variant in *PNPLA3* is associated with alcoholic liver disease. *Nat Genet* 2010;42:21-3.
429. Topaloglu AK, Lomniczi A, Kretzschmar D, Dissen GA, Kotan LD, McArdle CA, Koc AF, et al. Loss-of-function mutations in *PNPLA6* encoding neuropathy target esterase underlie pubertal failure and neurological deficits in Gordon Holmes syndrome. *J Clin Endocrinol Metab* 2014;99:E2067-75.
430. Trépo E, Guyot E, Ganne-Carrie N, Degre D, Gustot T, Franchimont D, Sutton A, et al. *PNPLA3* (rs738409 C > G) is a common risk variant associated with hepatocellular carcinoma in alcoholic cirrhosis. *Hepatology* 2012;55:1307-08.
431. Trépo E, Gustot T, Degre D, Lemmers A, Verset L, Demetter P, Ouziel R, et al. Common polymorphism in the *PNPLA3*/adiponutrin gene confers higher risk of cirrhosis and liver damage in alcoholic liver disease. *J Hepatol* 2011;55:906-12.
432. Trépo E, Nahon P, Bontempi G, Valenti L, Falletti E, Nischalke HD, Hamza S, et al. Association between the *PNPLA3* (rs738409 C>G) variant and hepatocellular carcinoma: Evidence from a meta-analysis of individual participant data. *Hepatology* 2014;59:2170-77.
433. Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P, Maier W, et al. Genome-wide association study of alcohol dependence. *Arch Gen Psychiatry* 2009;66:773-84.
434. Trotter T, Porter R. An Essay, Medical, Philosophical, and Chemical on Drunkenness and its Effects on the Human Body: Taylor & Francis, 1988.



435. True WR, Heath AC, Bucholz K, Slutske W, Romeis JC, Scherrer JF, Lin N, et al. Models of treatment seeking for alcoholism: the role of genes and environment. *Alcohol Clin Exp Res* 1996;20:1577-81.
436. Turner SD. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *bioRxiv* 2014:005165.
437. Tuyns AJ, Pequignot G. Greater risk of ascitic cirrhosis in females in relation to alcohol consumption. *Int J Epidemiol* 1984;13:53-7.
438. Tyburski EM, Sokolowski A, Samochowiec J, Samochowiec A. New diagnostic criteria for alcohol use disorders and novel treatment approaches—2014 update. *Arch Med Sci.* 2014;10:1191.
439. Uesugi T, Froh M, Arteel GE, Bradford BU, Thurman RG. Toll-like receptor 4 is involved in the mechanism of early alcohol-induced liver injury in mice. *Hepatology* 2001;34:101-08.
440. University of Oxford. Oxford Protein Production Facility,. In; 2015.
441. Urabe Y, Ochi H, Kato N, Kumar V, Takahashi A, Muroyama R, Hosono N, et al. A genome-wide association study of HCV-induced liver cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region. *J Hepatol* 2013;58:875-82.
442. Valenti L, De Feo T, Fracanzani AL, Fatta E, Salvagnini M, Arico S, Rossi G, et al. Cytotoxic T-lymphocyte antigen-4 A49G polymorphism is associated with susceptibility to and severity of alcoholic liver disease in Italian patients. *Alcohol Alcohol* 2004;39:276-80.
443. Valenti L, Motta BM, Soardo G, Iavarone M, Donati B, Sangiovanni A, Carnelutti A, et al. PNPLA3 I148M polymorphism, clinical presentation, and survival in patients with hepatocellular carcinoma. *PLoS One* 2013;8:e75982.
444. Verrill C, Markham H, Templeton A, Carr NJ, Sheron N. Alcohol-related cirrhosis - early abstinence is a key factor in prognosis, even in the most severe cases. *Addiction* 2009;104:768-74.
445. Walter SD, Stitt LW. Evaluating the survival of cancer cases detected by screening. *Stat Med* 1987;6:885-900.
446. Walters GD. The heritability of alcohol abuse and dependence: a meta-analysis of behavior genetic research. *Am J Drug Alcohol Abuse* 2002;28:557-84.
447. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* 2010;11:843-54.
448. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 2004;20:2138-9.

449. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;25:1189-91.
450. Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, Aulchenko YS, et al. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 2010;30:2264-76.
451. Watson PE, Watson ID, Batt RD. Total body water volumes for adult males and females estimated from simple anthropometric measurements. *Am J Clin Nutr* 1980;33:27-39.
452. Whitfield JB, Rahman K, Haber PS, Day CP, Masson S, Daly AK, Cordell HJ, et al. Brief Report: Genetics of Alcoholic Cirrhosis-GenomALC Multinational Study. *Alcohol Clin Exp Res* 2015;39:836-42.
453. Wiemann SU, Satyanarayana A, Tshuridu M, Tillmann HL, Zender L, Klempnauer J, Flemming P, et al. Hepatocyte telomere shortening and senescence are general markers of human liver cirrhosis. *FASEB J* 2002;16:935-42.
454. Wiens TK, Walker LJ. The chronic disease concept of addiction: Helpful or harmful? *Addict Res Theory* 2014:1-13.
455. Wijeyesakere SJ, Richardson RJ, Stuckey JA. Crystal structure of patatin-17 in complex with aged and non-aged organophosphorus compounds. *PLoS One* 2014;9:e108245.
456. Wilhelmsen KC, Swan GE, Cheng LS, Lessov-Schlaggar CN, Amos CI, Feiler HS, Hudmon KS, et al. Support for previously identified alcoholism susceptibility Loci in a cohort selected for smoking behavior. *Alcohol Clin Exp Res* 2005;29:2108-15.
457. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 2008;40:161-69.
458. Williams R, Aspinall R, Bellis M, Camps-Walsh G, Cramp M, Dhawan A, Ferguson J, et al. Addressing liver disease in the UK: a blueprint for attaining excellence in health care and reducing premature mortality from lifestyle issues of excess consumption of alcohol, obesity, and viral hepatitis. *Lancet* 2014;384:1953-97.
459. Wilson PA, Gardner SD, Lambie NM, Commans SA, Crowther DJ. Characterization of the human patatin-like phospholipase family. *J Lipid Res* 2006;47:1940-9.
460. Winberg ME, Motlagh MK, Stenkula KG, Holm C, Jones HA. Adiponutrin: A multimeric plasma protein. *Biochem Biophys Res Commun* 2014.
461. Winokur G, Reich T, Rimmer J, Pitts FNJ. Alcoholism. III. Diagnosis and familial psychiatric illness in 259 alcoholic probands. *Arch Gen Psychiatry* 1970;23:104-11.

462. Wong NA, Rae F, Simpson KJ, Murray GD, Harrison DJ. Genetic polymorphisms of cytochrome p4502E1 and susceptibility to alcoholic liver disease and hepatocellular carcinoma in a white population: a study and literature review, including meta-analysis. *Mol Pathol* 2000;53:88-93.
463. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. 1992.
464. World Health Organization. European health for all database (HFA-DB). In; 2005.
465. World Health Organization. Global Information System on Alcohol and Health (GISAH). In: World Health Organization, editor. Global status report on alcohol and health. Geneva; 2010.
466. World Health Organization. Global status report on alcohol and health 2014: World Health Organization, 2014.
467. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res* 2007;35:3375-82.
468. Wyszynski DF, Panhuysen CI, Ma Q, Yip AG, Wilcox M, Erlich P, Farrer LA. Genome-wide screen for heavy alcohol consumption. *BMC Genet* 2003;4 Suppl 1:S106.
469. Xie YD, Feng B, Gao Y, Wei L. Effect of abstinence from alcohol on survival of patients with alcoholic cirrhosis: A systematic review and meta-analysis. *Hepatol Res* 2014;44:436-49.
470. Xin YN, Zhao Y, Lin ZH, Jiang X, Xuan SY, Huang J. Molecular dynamics simulation of PNPLA3 I148M polymorphism reveals reduced substrate access to the catalytic cavity. *Proteins* 2013;81:406-14.
471. Xu Y, Peng B, Fu Y, Amos CI. Genome-wide algorithm for detecting CNV associations with diseases. *BMC Bioinformatics* 2011;12:331.
472. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res* 2015;43:W174-81.
473. Yang SQ, Lin HZ, Lane MD, Clemens M, Diehl AM. Obesity increases sensitivity to endotoxin liver injury: implications for the pathogenesis of steatohepatitis. *Proc Natl Acad Sci U S A* 1997;94:2557-62.
474. Yin SJ, Bosron WF, Magnes LJ, Li TK. Human liver alcohol dehydrogenase: purification and kinetic characterization of the  $\beta_2\beta_2$ ,  $\beta_2\beta_1$ ,  $\alpha\beta_2$ , and  $\beta_2\gamma_1$  "Oriental" isoenzymes. *Biochemistry* 1984;23:5847-53.
475. Yuan X, Waterworth D, Perry JR, Lim N, Song K, Chambers JC, Zhang W, et al. Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am J Hum Genet* 2008;83:520-8.

476. Zanier K, Nomine Y, Charbonnier S, Ruhlmann C, Schultz P, Schweizer J, Trave G. Formation of well-defined soluble aggregates upon fusion to MBP is a generic property of E6 proteins from various human papillomavirus species. *Protein Expr Purif* 2007;51:59-70.
477. Zhang HX, Zhai Y, Hu ZB, Wu C, Qian J, Jia WH, Ma FC, et al. Genome-wide association study identifies 1p36.22 as a new susceptibility locus for hepatocellular carcinoma in chronic hepatitis B virus carriers. *Nat Genet* 2010;42:755-39.
478. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998;280:1690-1.
479. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57:702-10.
480. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;33:2302-9.
481. Zimmerman MR. The paleopathology of the liver. *Ann Clin Lab Sci* 1990;20:301-6.
482. Zintzaras E, Stefanidis I, Santos M, Vidal F. Do alcohol-metabolizing enzyme gene polymorphisms increase the risk of alcoholism and alcoholic liver disease? *Hepatology* 2006;43:352-61.
483. Zuo L, Gelernter J, Zhang CK, Zhao H, Lu L, Kranzler HR, Malison RT, et al. Genome-wide association study of alcohol dependence implicates *KIAA0040* on chromosome 1q. *Neuropsychopharmacology* 2012;37:557-66.
484. Zuvich RL, Armstrong LL, Bielinski SJ, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genet Epidemiol* 2011;35:887-98.
485. Zwier MC, Chong LT. Reaching biological timescales with all-atom molecular dynamics simulations. *Curr Opin Pharmacol* 2010;10:745-52.

---

---

# APPENDICES

---

---

## DNA EXTRACTION

Three different techniques were used to obtain the genomic DNA samples that comprised the entire UCL cohort.

### Phenol-Chloroform

The majority of genomic DNA samples were extracted from lymphocytes extracted from whole blood using a phenol-chloroform based protocol. The first day, of this protocol involves separating lymphocytes from the other constituents of blood by. Separation is performed using rounds of rounds of centrifugation and washing with red-blood cell lysis buffer. The separated lymphocytes are then lysed using a lysis buffer with the addition of proteinase K and left overnight in a water bath at 50°C. The second day of this protocol involves removing non-DNA impurities, such as proteins, from the solution using phenol and chloroform:IAA phase separation. In this purification process proteins and other compounds dissolve into the organic layer (the phenol and chloroform) while the DNA remains in the solution of the aqueous layer. The aqueous layer can be separated from the organic layer and from this; the DNA is precipitated with concentrated sodium acetate and absolute ethanol. The precipitated DNA is washed in 70% ethanol to remove excess salts followed by rehydration in TE buffer.

#### *Day One*

1. Thaw blood in a 37°C water bath for 30-45 min.
2. Once the blood is thawed store on ice.
3. Transfer blood from Vacutainer tubes into falcon tubes.
4. Add 5-10 mL 1x Lysis Buffer, shake and then top up to 50 mL.
5. Spin solution at 3,000 rpm for 15 min and 4°C.
6. Pour off supernatant and re-suspend in 5-10 mL 1x Lysis Buffer to a final volume of 50 mL.
7. Spin solution at 3,000 rpm for 15 min and 4°C.
8. Pour off supernatant and re-suspend in 5 mL Proteinase K Buffer, vortex, then top to 10 mL.
9. Add 500 µL 10 % SDS.
10. Add 50 µL of 20 mg/mL Proteinase K solution
11. Mix then incubate in a shaking 55°C water bath overnight.

## Day Two

12. Weigh out 1g of PVPP into a fresh falcon tube and add 5ml TE.
13. Add 5 mL TE-equilibrated Phenol and 5 mL Chloroform:IAA (24:1).
14. Shake to emulsify.
15. Transfer lysate to falcon tube with PVPP.
16. Shake to emulsify.
17. Centrifuge at 3,000 rpm for 5-10 Min at room temperature.
18. Transfer the upper aqueous phase into new falcon tubes. Take care to avoid degraded proteins at the interface.
19. Back extract if necessary by adding TE Buffer and centrifuging again.
20. Add 1 to 1.5 mL 3M (0.1 X volume) sodium acetate then immediately add 100% Ethanol to a final volume of 35-40 mL.
21. Mix gently by inverting the tube once or twice.
22. Remove the White DNA using a fresh inoculation loop.
23. Briefly rinse the DNA in a 1.5 mL tube containing 70 % ethanol.
24. Allow the DNA to dry for five min.
25. Transfer the DNA to a 1.5 mL tube containing 500  $\mu$ L TE (use 250  $\mu$ L TE if the yield of DNA is low).
26. Leave at room temperature for 3-5 days to dissolve.

Supplementary Table 1 Standard media for phenol/chloroform based genomic DNA extraction

Media	Constituents
Tris-EDTA (TE) Buffer	10 mM Tris. 1 mM EDTA
Low Tris-EDTA buffer	10 mM Tris. 0.1 mM EDTA
10X Lysis buffer	100 mM NaCl, 100 mM EDTA,
Proteinase K buffer	50mM Tris.HCl (pH8.0), 50mM EDTA, 100mM NaCl
Proteinase K	Proteinase K enzyme (20 mg/mL)
1:24 Isoamylalcohol:chloroform solution	40 mL Isoamylalcohol: 960 mL chloroform
DNA precipitation solution	Absolute ethanol (Sigma-Aldrich, UK)
DNA precipitation solution	3 M Sodium Acetate
Phenol	Phenol (equilibrated with 10 mM Tris HCl, 1 mM EDTA ph 8.0)
Poly(vinylpolypyrrolidone) (PVPP)	Cross-linked form of PVPP
DNA cleaning solution	70% Ethanol

\*All of the reagents used to create these media were purchased from Sigma-Aldrich  
All buffers made in double-deionized H<sub>2</sub>O except where otherwise explicitly stated

## PureGene

The PureGene method was used to extract genomic DNA from whole blood samples. This method has been adapted from a protocol provided from a commercial Gentra Puregene kit. As with the other protocol, lymphocytes are separated from other constituents of the blood through stages of washing with red-blood cell lysis buffer and centrifugation. The lymphocytes are lysed using a lysis buffer containing proteinase K, releasing both the DNA and other cellular components into solution. These cell debris are removed from the solution through precipitation using a protein precipitation solution. The remaining eluent containing the DNA is mixed with DNA precipitation solution. The precipitated DNA is separated from the solution through centrifugation and is then washed in 70% Ethanol before rehydration in TE buffer.

1. Thaw blood in a 37°C water bath for 30 min.
2. Pour defrosted blood and 30 mL 1X RBC Lysis solution into a labelled 50 mL Falcon Tube.
3. Incubate samples for 5 minutes at RT. Invert several times during incubation.
4. Centrifuge for 5 minutes at 3000 RPM and 4°C.
5. Pour off supernatant into leaving the cell pellet.
6. Re-perform the earlier steps twice until the supernatant is visibly clear of haemoglobin.
7. Add 15 µL of Proteinase K solution to the cell pellet and any remaining RBC lysis solution. Vortex until homogeneous.
8. Add 10 mL of Cell lysis solution to each sample and vortex this for 10 seconds.
9. Leave in water bath at 55°C for a minimum of 30 minutes. The sample should turn a straw yellow colour.
10. Place sample in ice for 5 minutes. When cool add 3.33 mL of Protein Precipitation solution and vortex this. Leave the tube on ice, and vortex it intermittently until the mixture inside it turns translucent/opaque.
11. Spin solution at 3,000 rpm for 10 minutes at 4°C.
12. Carefully transfer supernatant, avoiding the precipitated protein pellet into another falcon tube containing 10 mL isopropanol.
13. Invert tube several times until the DNA strands precipitate.
14. Spin solution at 3,000 rpm for 5 minutes to pellet the DNA.
15. Pour off the supernatant keeping the DNA pellet. To this add 10 mL of 70% Ethanol, vortex and leave on shaker for 10 minutes.
16. Spin the DNA ethanol solution at 3,000 rpm for 5 min to pellet DNA.
17. Carefully pour off supernatant. Dry the DNA pellet for until all residual ethanol has evaporated.

18. Re-suspend DNA pellet in 500  $\mu\text{L}$  of TE buffer (use 250  $\mu\text{L}$  TE if the yield of DNA is low).
19. Incubate the samples in a water bath at 55°C for 1 hour.
20. Leave the DNA in a shaker at 37°C overnight.
21. Transfer samples into tubes.

Supplementary Table 2 Standard media for purgene based genomic DNA extraction

Media	Constituents
Red blood cell (RBC) Lysis Buffer	100mM NaCl, 10mM EDTA, 1.5M NH <sub>4</sub> Cl.
Cell Lysis Solution	10 mM Tris-HCl pH 8.0, 25 mM EDTA, 0.5% SDS
Protein Precipitation solution	5M Ammonium Acetate
DNA precipitation solution	Absolute isopropanol
DNA cleaning solution	70% Ethanol

\*All of the reagents used to create these media were purchased from Sigma-Aldrich. All buffers made in double-deionized H<sub>2</sub>O except where otherwise explicitly stated.

### From saliva

At sample recruitment sites without trained phlebotomists or where blood was difficult to obtain saliva DNA collection kits (Oragene® DNA (OG-500)) were used. The extraction protocol from these kits involves requires reagents and protocol provided by the manufacturer. This protocol involves the inactivation of nucleases through heat treatment overnight. The following day proteins and other turbid compounds are precipitated through salting out. DNA is precipitated through addition of ethanol and collected through centrifugation. DNA pellets were rehydrated in TE buffer.

### DNA Quantification

DNA quantification was performed using a Qubit® 2.0 Fluorometer. Using the double stranded DNA broad range assay (ThermoFisher scientific, Q32850), quantification was performed following manufacturer stated guidelines. All DNA concentrations were determined from the average of four readings taken from two independent replicate samples.

### DNA Storage

DNA samples were stored in several different formats for the purposes of routine genotyping and long-term storage (Supplementary Figure 1). Quantified DNA samples were normalised to a stock concentration with TE buffer and normalised to a maximum concentration of 50 ng/ $\mu\text{L}$ , 100 ng/ $\mu\text{L}$ , 200 ng/ $\mu\text{L}$  or 500 ng/ $\mu\text{L}$ . These long-term stock DNA samples were stored at -80°C. From the stock DNA normalised working stock DNA solutions were created at a concentration of 25 ng/ $\mu\text{L}$  for storage at 4°C. These samples were stored on 96-tube plates with sample ID mapped to a unique position on



each plate. Deep well plates were created maintaining the same 96-well layout as stored on a centrally backed up database. DNA samples were diluted to a concentration of 3.33 ng/ $\mu$ L in low TE-Buffer. These deep well plates were stored at -20°C. Using an epMotion 5070 Automated Pipetting system (Eppendorf, Stevenage, UK) 3.33 ng of genomic DNA from four 96-deep well plates were pipetted and dehydrated on to LightCycler® 480 Multiwell 384 plates (Roche, UK).



#### **Supplementary Figure 1 A schematic of DNA plate creation**

Stock DNA is transferred and normalised for routine use at 25 ng/ $\mu$ L. This DNA is then transferred to a deep-well plate at a working concentration from which it is automatically pipetted on to 384 well-plates.

### **KASPAR GENOTYPING**

All genotyping experiments performed in the UCL Molecular Psychiatry lab were performed using KASPAR genotyping (K-Bioscience, Hoddenson, UK).

The sequences of genotyping primers were determined using the software Primerpicker lite (version 0.27) and manually checked for efficacy using the in silico PCR tool on the UCSC genome browser<sup>216</sup>. For each variant that was genotyped, two allele specific forward primers and two unique reverse primers were purchased (SigmaAldrich, UK) at a reverse-phase cartridge purity level. For each set of primers, several PCR conditions (Supplementary Table 5) were tested for the optimal fluorescent detection of genotypes. This optimisation tested several different concentrations of magnesium chloride or the dimethyl sulfoxide in the assay mix (Supplementary Table 3). The two reverse primers were also tested for their efficacy with the most efficacious primer selected for genotyping.

Supplementary Table 3 Primer assay mix constituents

Primer	Volume ( $\mu\text{L}$ )
Forward Allele Specific primer 1	6
Forward Allele Specific primer 1	6
Reverse Primer	15
Water	73
Total	100

Supplementary Table 4 KASPAR genotyping optimisation conditions

	10 reactions					
	A	B	C	D	E	F
	1.8mM MgCl <sub>2</sub>	2.2mM MgCl <sub>2</sub>	2.5mM MgCl <sub>2</sub>	2.8mM MgCl <sub>2</sub>	5% DMSO	10% DMSO
DNA	10	10	10	10	10	10
2X RXN mix (+KTAQ)	20	20	20	20	20	20
Assay mix	1.1	1.1	1.1	1.1	1.1	1.1
5 mM MgCl <sub>2</sub>	0	3.2	5.6	8	0.0	0.0
Water	8.9	5.7	3.3	0.9	6.9	4.9
Total	40	40	40	40	40.0	40.0
DMSO	0	0	0	0	2.0	4.0

Abbreviations: RXN – reaction; KTAQ - ; DMSO – dimethyl sulfoxide

The PCR cycling conditions used in all assays are standard involving a hot-start activation and subsequent touch-down denaturation and annealing/extension stages (Supplementary Table 5). The KASPAR PCR reaction was performed either directly inside the LightCycler480 or in separate PCR machines machines (Eppendorf® Mastercycler® Pro Thermal Cyclers). The fluorescence detection stage was performed on the LightCycler480 (Roche, UK). The process of fluorescence detection is effected by the unique KASP chemistry (Supplementary Figure 2) from which genotype cluster plots can be created and hence genotypes determined.

For the routine genotyping of the thousands of samples present in the UCL cohort, the assay mix condition determined during optimisation was scaled up for the necessary number of samples and automatically plated on to LightCycler® 480 Multiwell 384 plates (Roche, UK).

Supplementary Table 5 LightCycler480™ KASPAR cycling conditions

KASP Cycling Conditions	
Hot start activation	94°C for 15 minutes
10 cycles of;	94°C for 20 seconds 65°C to 57°C for 60 seconds (decreasing by 0.8°C cycle)
26 cycles of:	94°C for 20 seconds 57°C for 60 seconds
Reading 1	1 <sup>st</sup> Reading at 37°C
3 cycles of:	94°C for 20 seconds 57°C for 60 seconds
Reading 2	2 <sup>nd</sup> Reading at 37°C
3 cycles of:	94°C for 20 seconds 57°C for 60 seconds
Reading 3	3 <sup>rd</sup> Reading at 37°C



**Supplementary Figure 2 The molecular biological mechanism of fluorescent genotype detection by KASPAR genotyping**

This technique couples allele specific PCR with the fluorescent dyes VIC and FAM and can be used to genotype SNPs as well as short insertion/deletion variants using a standard proprietary reaction mix and allele specific primers. The proprietary reaction mix contains two universal primers which are chemically bound to a fluorescent molecule which absorb light at unique wavelengths. The allele specific primers are custom designed for each genetic variant under investigation - these are designed to contain an allele specific base at their 3' end and a 5' tail sequence which is complementary to the nucleotide sequence present on the universal primers. During a PCR reaction, the each allele specific primers will only amplify the template DNA which contains the specific allele to which they bind. The two fluorescently labelled universal primers present in the reaction mix will anneal to the amplified templates and from the fluorescence ratios genotypes can be determined. Using a real-time PCR machine the absorbance ratio between the wavelengths at which VIC and FAM absorb the genotype of a DNA sample may be determined.

Supplementary Table 6 Top hit variants associated with alcohol-related cirrhosis in the German discovery cohort

Chromosome	Variant	Hg19 Position	Significance P	Odds Ratio	Lower CI (95%)	Upper CI (95%)
22	rs738408	44324730	5.26X10 <sup>-23</sup>	2.40	2.02	2.85
22	rs738409	44324727	5.26X10 <sup>-23</sup>	2.40	2.02	2.85
22	rs3747207	44324855	5.61X10 <sup>-23</sup>	2.40	2.02	2.85
22	rs2294915	44340904	5.03X10 <sup>-22</sup>	2.36	1.99	2.80
22	rs201016637	44325479	3.68X10 <sup>-18</sup>	2.28	1.90	2.74
22	rs1977081	44330128	1.15X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs4823173	44328730	1.31X10 <sup>-17</sup>	2.38	1.98	2.87
22	rs2294433	44329275	1.34X10 <sup>-17</sup>	2.38	1.98	2.87
22	rs12484801	44325565	1.38X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs12483959	44325996	1.38X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs12484809	44325631	1.39X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs12485100	44325516	1.39X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs16991158	44327179	1.44X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs36055245	44327192	1.44X10 <sup>-17</sup>	2.38	1.97	2.87
22	rs11090617	44326700	1.46X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs9625962	44326272	1.46X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs2076211	44329078	1.50X10 <sup>-17</sup>	2.37	1.97	2.86
22	rs12484700	44327273	1.52X10 <sup>-17</sup>	2.38	1.97	2.86
22	rs2294922	44379565	2.04X10 <sup>-17</sup>	2.10	1.77	2.50
22	rs1977080	44330031	2.10X10 <sup>-17</sup>	2.37	1.97	2.85
22	rs1997693	44331513	2.20X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs2896019	44333694	2.43X10 <sup>-17</sup>	2.37	1.97	2.85
22	rs12484466	44330213	2.81X10 <sup>-17</sup>	2.34	1.95	2.82
22	rs2072906	44333172	2.93X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs2281135	44332570	2.95X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs36038527	44332888	2.97X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs2896020	44333968	3.00X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs2401512	44333945	3.02X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs4823176	44334476	3.16X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs4823177	44334486	3.17X10 <sup>-17</sup>	2.36	1.96	2.84
22	rs1883348	44331815	3.22X10 <sup>-17</sup>	2.35	1.95	2.82
22	rs13056638	44331778	3.29X10 <sup>-17</sup>	2.35	1.95	2.82
22	rs1883349	44331943	3.33X10 <sup>-17</sup>	2.35	1.95	2.82
22	rs2281138	44332477	3.37X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs2281137	44332493	3.37X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs2076207	44333370	3.38X10 <sup>-17</sup>	2.35	1.96	2.83
22	rs2072907	44332653	3.38X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs2072905	44333479	3.39X10 <sup>-17</sup>	2.36	1.96	2.83
22	rs16991175	44335331	3.41X10 <sup>-17</sup>	2.36	1.96	2.83
22	rs35621602	44335406	3.44X10 <sup>-17</sup>	2.36	1.96	2.83
22	rs34352134	44335416	3.44X10 <sup>-17</sup>	2.36	1.96	2.83

22	rs34376930	44335453	3.45X10 <sup>-17</sup>	2.36	1.96	2.83
22	rs2073081	44335744	3.55X10 <sup>-17</sup>	2.36	1.96	2.83
22	rs4823178	44334529	3.58X10 <sup>-17</sup>	2.35	1.96	2.83
22	rs34879941	44332878	3.63X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs1010023	44336098	3.67X10 <sup>-17</sup>	2.36	1.96	2.83
22	rs1010022	44336310	3.85X10 <sup>-17</sup>	2.35	1.96	2.83
22	rs13056555	44339526	3.93X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs2281293	44334842	4.01X10 <sup>-17</sup>	2.35	1.95	2.83
22	rs926633	44337533	4.22X10 <sup>-17</sup>	2.35	1.96	2.83
22	rs36069781	44340086	5.14X10 <sup>-17</sup>	2.34	1.95	2.82
22	rs8142145	44336496	5.38X10 <sup>-17</sup>	2.34	1.95	2.82
22	rs73176497	44336957	5.69X10 <sup>-17</sup>	2.31	1.92	2.77
22	rs2294916	44340922	5.93X10 <sup>-17</sup>	2.34	1.95	2.82
22	rs4823179	44341193	6.22X10 <sup>-17</sup>	2.34	1.95	2.82
22	rs4823180	44341298	6.32X10 <sup>-17</sup>	2.34	1.95	2.82
22	rs13055900	44341666	6.56X10 <sup>-17</sup>	2.34	1.95	2.81
22	rs13055874	44341672	6.57X10 <sup>-17</sup>	2.34	1.95	2.81
22	rs4823181	44341606	6.60X10 <sup>-17</sup>	2.34	1.95	2.81
22	rs2008451	44342969	7.50X10 <sup>-17</sup>	2.34	1.94	2.81
22	rs1810508	44343151	7.61X10 <sup>-17</sup>	2.34	1.94	2.81
22	rs13054885	44345771	1.22X10 <sup>-16</sup>	2.32	1.93	2.79
22	rs12484795	44343626	2.90X10 <sup>-16</sup>	2.21	1.85	2.65
22	rs2092501	44347251	1.03X10 <sup>-14</sup>	2.21	1.84	2.67
22	rs34912062	44348446	1.04X10 <sup>-14</sup>	2.21	1.84	2.67
22	rs56373884	44356468	1.70X10 <sup>-14</sup>	2.20	1.83	2.66
22	rs1474745	44349236	1.78X10 <sup>-14</sup>	2.20	1.83	2.66
22	rs2294921	44361842	3.02X10 <sup>-14</sup>	2.19	1.82	2.64
22	rs3761472	44368122	3.92X10 <sup>-14</sup>	2.19	1.82	2.63
22	rs4823108	44381340	4.96X10 <sup>-13</sup>	2.11	1.74	2.57
22	rs12167845	44380767	5.02X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs71313378	44380170	5.24X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs9626079	44380009	5.30X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs2294923	44379740	5.52X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs2235777	44378809	5.58X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs2235776	44377999	5.77X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs4823183	44378672	5.90X10 <sup>-13</sup>	2.12	1.74	2.57
22	rs4823109	44381482	6.79X10 <sup>-13</sup>	2.09	1.72	2.54
22	rs61473277	44371406	7.88X10 <sup>-13</sup>	2.12	1.75	2.58
22	rs2281298	44391234	7.77X10 <sup>-11</sup>	1.94	1.61	2.34
22	rs2143571	44391686	9.17X10 <sup>-11</sup>	1.94	1.61	2.33
22	rs2073079	44385594	1.06X10 <sup>-10</sup>	1.94	1.61	2.33
22	rs3827385	44388817	1.14X10 <sup>-10</sup>	1.94	1.61	2.33
22	rs2073080	44394402	1.20X10 <sup>-10</sup>	1.94	1.61	2.33
22	rs2401514	44394019	1.27X10 <sup>-10</sup>	1.94	1.61	2.33
22	rs2294927	44382684	1.82X10 <sup>-10</sup>	1.65	1.41	1.95
22	rs2235778	44389514	1.88X10 <sup>-10</sup>	1.65	1.40	1.94
22	rs10656207	44387932	1.93X10 <sup>-10</sup>	1.65	1.40	1.94

22	rs1986095	44387108	1.94X10 <sup>-10</sup>	1.65	1.40	1.94
22	rs3788604	44388417	1.94X10 <sup>-10</sup>	1.65	1.40	1.94
22	rs6006602	44383400	1.99X10 <sup>-10</sup>	1.65	1.40	1.94
22	rs6006468	44383432	1.99X10 <sup>-10</sup>	1.65	1.40	1.94
22	rs6006469	44383617	2.18X10 <sup>-10</sup>	1.65	1.40	1.94
22	rs6006473	44393075	2.59X10 <sup>-10</sup>	1.65	1.40	1.93
22	rs2281292	44395389	2.91X10 <sup>-10</sup>	1.64	1.40	1.93
22	rs1007863	44395451	2.96X10 <sup>-10</sup>	1.64	1.40	1.93
22	rs12165526	44361713	5.68X10 <sup>-9</sup>	1.75	1.38	2.22
22	rs1883350	44328043	7.05X10 <sup>-9</sup>	1.62	1.37	1.91
22	rs67450864	44376335	1.12X10 <sup>-8</sup>	1.56	1.33	1.84
22	rs4823182	44377442	1.24X10 <sup>-8</sup>	1.56	1.33	1.84
22	rs66812091	44335670	2.18X10 <sup>-8</sup>	1.62	1.38	1.91
22	rs2294917	44341986	8.15X10 <sup>-8</sup>	0.59	0.49	0.71
20	rs11696461	40899315	2.24X10 <sup>-7</sup>	3.40	1.82	6.33
22	rs738407	44323955	4.25X10 <sup>-7</sup>	1.51	1.28	1.77
22	rs2294926	44382533	8.00X10 <sup>-7</sup>	1.49	1.26	1.76
22	rs2179642	44391588	1.01X10 <sup>-6</sup>	1.50	1.27	1.77
22	rs2281297	44390568	1.22X10 <sup>-6</sup>	1.49	1.26	1.76
22	rs6006474	44393241	1.32X10 <sup>-6</sup>	1.50	1.27	1.77
22	rs11912828	44348116	2.36X10 <sup>-6</sup>	0.61	0.50	0.76
22	rs2294919	44342325	2.81X10 <sup>-6</sup>	0.62	0.50	0.76
22	rs6006599	44382004	2.92X10 <sup>-6</sup>	1.51	1.27	1.78
14	rs74893904	28476657	3.19X10 <sup>-6</sup>	1.79	1.36	2.36
18	rs57083541	77755460	3.29X10 <sup>-6</sup>	2.46	1.50	4.04
14	rs67608825	28476537	3.30X10 <sup>-6</sup>	1.79	1.36	2.35
22	rs139052	44327012	3.68X10 <sup>-6</sup>	1.59	1.29	1.95
8	rs117106349	77431741	4.07X10 <sup>-6</sup>	2.38	1.53	3.72
16	rs300019	86661105	4.09X10 <sup>-6</sup>	2.10	1.51	2.93
22	rs13055235	44400149	4.14X10 <sup>-6</sup>	2.18	1.56	3.06
22	rs56219234	44357894	4.34X10 <sup>-6</sup>	1.50	1.27	1.77
4	rs17886348	123533820	5.01X10 <sup>-6</sup>	1.98	1.51	2.60
22	rs738491	44354111	5.48X10 <sup>-6</sup>	1.49	1.26	1.76
4	rs17879298	123533834	5.80X10 <sup>-6</sup>	2.00	1.52	2.63
4	rs62324196	123530977	5.80X10 <sup>-6</sup>	2.00	1.52	2.63
4	rs62324197	123532202	5.83X10 <sup>-6</sup>	2.00	1.52	2.63
16	rs56237022	86659426	5.87X10 <sup>-6</sup>	2.06	1.48	2.86
11	rs143612609	1701967	5.88X10 <sup>-6</sup>	1.47	1.22	1.77
4	rs45589038	123532319	5.89X10 <sup>-6</sup>	2.00	1.52	2.63
5	rs13182940	76306771	6.09X10 <sup>-6</sup>	1.34	1.14	1.58
11	rs11039418	1702170	6.10X10 <sup>-6</sup>	1.47	1.22	1.77
4	rs45532838	123535124	6.13X10 <sup>-6</sup>	1.99	1.51	2.62
5	rs41096	76310835	6.22X10 <sup>-6</sup>	1.34	1.14	1.58
4	rs62324208	123557430	6.44X10 <sup>-6</sup>	1.99	1.50	2.62
4	rs62324207	123557331	6.44X10 <sup>-6</sup>	1.99	1.51	2.62
4	rs7667072	123556921	6.59X10 <sup>-6</sup>	1.99	1.51	2.62

5	rs6898403	76303398	6.81X10 <sup>-6</sup>	1.34	1.14	1.58
4	rs2893008	123552814	6.85X10 <sup>-6</sup>	2.00	1.52	2.63
4	rs6833795	123553388	7.13X10 <sup>-6</sup>	1.99	1.51	2.62
4	rs149787056	123552022	7.30X10 <sup>-6</sup>	1.99	1.51	2.62
5	rs7728968	76306006	7.42X10 <sup>-6</sup>	1.34	1.13	1.58
5	rs7723775	76305856	7.43X10 <sup>-6</sup>	1.34	1.13	1.58
4	rs142775319	123550665	7.62X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs62324201	123545119	7.63X10 <sup>-6</sup>	1.98	1.51	2.61
4	rs62324206	123548328	7.64X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs62324205	123548192	7.64X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs4326027	123547272	7.65X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs78780360	123547001	7.65X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs17880969	123541473	7.65X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs62324203	123546463	7.65X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs45441794	123544324	7.65X10 <sup>-6</sup>	1.99	1.51	2.61
4	rs62324202	123545147	7.65X10 <sup>-6</sup>	1.99	1.51	2.61
11	rs58237691	1702238	7.69X10 <sup>-6</sup>	1.45	1.21	1.75
4	rs78541112	123526068	7.75X10 <sup>-6</sup>	1.98	1.51	2.61
4	rs62324194	123524666	7.76X10 <sup>-6</sup>	1.98	1.51	2.61
4	rs62324193	123524632	7.76X10 <sup>-6</sup>	1.98	1.51	2.61
4	rs62324191	123523832	7.77X10 <sup>-6</sup>	1.98	1.51	2.61
4	rs62324192	123524257	7.80X10 <sup>-6</sup>	1.98	1.51	2.61
5	rs32900	76310114	7.92X10 <sup>-6</sup>	1.33	1.13	1.57
4	rs62324170	123521851	8.03X10 <sup>-6</sup>	1.98	1.51	2.60
4	rs74985076	123520646	8.09X10 <sup>-6</sup>	1.98	1.50	2.60
16	rs12051337	86663934	8.11X10 <sup>-6</sup>	2.10	1.51	2.93
4	rs62324169	123519601	8.16X10 <sup>-6</sup>	1.98	1.50	2.60
22	rs59436064	44362688	8.20X10 <sup>-6</sup>	1.46	1.23	1.73
4	rs6842843	123512132	8.28X10 <sup>-6</sup>	1.97	1.50	2.60
4	rs74388271	123508611	8.63X10 <sup>-6</sup>	1.97	1.50	2.59
4	rs62324166	123504864	8.71X10 <sup>-6</sup>	1.97	1.50	2.59
4	rs62324167	123507715	8.82X10 <sup>-6</sup>	1.96	1.49	2.58
2	rs3731944	45616849	9.19X10 <sup>-6</sup>	0.47	0.35	0.63
11	rs11211884	106597231	9.87X10 <sup>-6</sup>	1.38	1.17	1.62

Abbreviations: CI – confidence interval, Hg19 – human genome 19

Supplementary Sequence 1 The primary amino acid sequence sequence of PNPLA3  
Accession: NP\_079501

001-MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGA  
051-LHCVGVLSGIPLEQTLQVLSDLVRKARSRNIGIFHPSFNLSKFLRQGLCK  
101-CLPANVHQLISGKIGISLTRVSDGENVLVSDFRSKDEVVDALVCSCFIPF  
151-YSGLIPPSFRGVRYVDGGVSDNVPFIDAKTTITVSPFYGEYDICPKVKST  
201-NFLHVDITKLSLRLCTGNLYLLSRAFVPPDLKVLGEICLRGYLDAFRFLE  
251-EKGICNRPQPGLKSSSEGMDPEVAMP SWANMSLDSSPESAALAVRLEGDE  
301-LLDHLRLSILPWDESILDTLSPRLATALSEEMKDKGGYMSKICNLLPIRI  
351-MSYVMLPCTLPVESAI AIVQRLVTWLPDMPDDVLWLQWVTSQVFTRVLMC  
401-LLPASRSQMPVSSQQASPCTPEQDWPCWTPCSPKGC PAETKAEATPRSIL  
451-RSSLNFFLG NKVPAGAEGLSTFPSFSLEKSL



Supplementary Sequence 2 The nucleotide sequence of the PNPLA3 cDNA

CGCTTGC GGGCGCCGGGCGGAGCTGCTGCGGATCAGGACCCGAGCCGATTCCC GATCCC GACCC  
AGATCC TAACCCGCGCCCCCGCCCCGCCGCCGATGTACGACGCAGAGCGCGGCTGGAGC  
TTGTCCTTCGCGGGCTGCGGCTTCCTGGGCTTCTACCACGTCGGGGCGACCCGCTGCCTGAGCG  
AGCACGCCCCGCACCTCCTCCGCGACGCGCGCATGTTGTTTCGGCGCTTCGGCCGGGGCGTTGCA  
CTGCGTCGGCGTCTCTCCGGTATCCCGCTGGAGCAGACTCTGCAGGTCCTCTCAGATCTTGTG  
CGGAAGGCCAGGAGTCGGAACATTGGCATCTTCCATCCATCCTTCAACTTAAGCAAGTTCTTCC  
GACAGGGTCTCGGCAAATGCCTCCCGGCAAATGTCCACCAGCTCATCTCCGGCAAATAGGCAT  
CTCTCTTACCAGAGTGTCTGATGGGGAAAACGTTCTGGTGTCTGACTTTCGGTCCAAAGACGAA  
GTCGTGGATGCCTTGGTATGTTCTGCTTCATGCCTTTCTACAGTGGCCTTATCCCTCCTTCT  
TCAGAGGCGTGC GATATGTGGATGGAGGAGTGAGTGACAACGTACCCTTCATTGATGCCAAAAC  
AACCATCACCGTGTCCCCCTTCTATGGGGAGTACGACATCTGCCCTAAAGTCAAGTCCACGAAC  
TTTCTTCATGTGGACATCACCAAGCTCAGTCTACGCCTCTGCACAGGGAACCTCTACCTTCTCT  
CGAGAGCTTTTGTCCCCCGGATCTCAAGGTGCTGGGAGAGATATGCCTTCGAGGATATTTGGA  
TGCATTCAGGTTCTTGAAGAGAAGGGCATCTGCAACAGGCCCCAGCCAGGCTGAAGTCATCC  
TCAGAAGGGATGGATCCTGAGGTCGCCATGCCAGCTGGGCAAACATGAGTCTGGATTCTTCCC  
CGGAGTCGGCTGCCTTGGCTGTGAGGCTGGAGGGAGATGAGCTGCTAGACCACCTGCGTCTCAG  
CATCTGCCCTGGGATGAGAGCATCCTGGACACCCTCTCGCCAGGCTCGCTACAGCACTGAGT  
GAAGAAATGAAAGACAAAGGTGGATACATGAGCAAGATTTGCAACTTGCTACCCATTAGGATAA  
TGTCTTATGTAATGCTGCCCTGTACCCTGCCTGTGGAATCTGCCATTGCGATTGTCCAGAGACT  
GGTGACATGGCTTCCAGATATGCCCGACGATGTCCTGTGGTTGCAGTGGGTGACCTCACAGGTG  
TTCACTCGAGTGCTGATGTGTCTGCTCCCCGCCTCCAGGTCCCAAATGCCAGTGAGCAGCCAAC  
AGGCCTCCCATGCACACCTGAGCAGGACTGGCCCTGCTGGACTCCCTGCTCCCCGAGGGCTG  
TCCAGCAGAGACCAAAGCAGAGGCCACCCCGCGTCCATCCTCAGGTCCAGCCTGAACTTCTTC  
TTGGGCAATAAAGTACCTGCTGGTGTGAGGGGCTCTCCACCTTTCCAGTTTTTCACTAGAGA  
AGAGTCTGTGAGTCACTT GAGGAGCGAGTCTAGCAGATTTCTTTCAGAGGTGCTAAAGTTTCCC  
ATCTTTGTGCAGCTACCTCCGCATTGCTGTGTAGTGACCCCTGCCTGTGACGTGGAGGATCCCA  
GCCTCTGAGCTGAGTTGGTTTTATGAAAAGCTAGGAAGCAACTTTTCGCCTGTGCAGCGGTCCA  
GCAC TTA ACTCTAATACATCAGCATGCGTTAATTCAGCTGGTTGGGAAATGACACCAGGAAGCC  
CAGTGCAGAGGGTCCCTTACTGACTGTTTTCGTGGCCCTATTAATGGTCAGACTGTTCCAGCATG  
AGGTTCTTAGAATGACAGGTGTTTGGATGGGTGGGGCCCTTGTGATGGGGGGTAGGCTGGCCCA  
TGTGTGATCTTGTGGGTGGAGGGAAGAGAATAGCATGATCCCACTTCCCATGCTGTGGGAAG  
GGGTGCAGTTCGTCCCCAAGAACGACACTGCCTGT CAGGTGGTCTGCAAAGATGATAACCTTGA  
CTACTAAAAACGTCTCCATGGCGGGGGTAACAAGATGATAATCTACTTAATTTTAGAACACCTT  
TTTACCTAACTAAAATAATGTTTAAAGAGTTTTGTATAAAAATGTAAGGAAGCGTTGTACCT  
GTTGAATTTTGTATTATGTGAATCAGTGAGATGTTAGTAGAATAAGCCTTAAAAAAAAAAAAAA  
AAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Supplementary Table 7 MASCOT sequence alignment results for the SDS-PAGE gel bands that underwent mass spectrometry

Band1 - Maltose Binding Protein	>MBP:1	MKIKTGARILALSALTTMMFSASALAKIEEGKLVIIWINGDKGYNGLAIEVGG
	51	KKFEKDTGKIKVTVEHPDKLEEEKFPQVAATGDGPDIIIFWAHDFRGGYAAQSG
	101	LLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSIIYNKDLLPN
	151	PKTWEIIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENG
	201	KYDIKDVGVNDAGAKAGLTFLVDLIKXKHMNADTDYSIAEAAFNKGETAM
	251	TINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKE
	301	LAKFEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENA
	351	QKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK
	>MBP:1	MKIKTGARILALSALTTMMFSASALAKIEEGKLVIIWINGDKGYNGLAIEVGG
	51	KKFEKDTGKIKVTVEHPDKLEEEKFPQVAATGDGPDIIIFWAHDFRGGYAAQSG
101	LLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSIIYNKDLLPN	
151	PKTWEIIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENG	
201	KYDIKDVGVNDAGAKAGLTFLVDLIKXKHMNADTDYSIAEAAFNKGETAM	
251	TINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKE	
301	LAKFEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENA	
351	QKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK	
Band2 - Maltose Binding Protein	>MBP:1	MKIKTGARILALSALTTMMFSASALAKIEEGKLVIIWINGDKGYNGLAIEVGG
	51	KKFEKDTGKIKVTVEHPDKLEEEKFPQVAATGDGPDIIIFWAHDFRGGYAAQSG
	101	LLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSIIYNKDLLPN
	151	PKTWEIIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENG
	201	KYDIKDVGVNDAGAKAGLTFLVDLIKXKHMNADTDYSIAEAAFNKGETAM
	251	TINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKE
	301	LAKFEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENA
	351	QKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK
	>MBP:1	MKIKTGARILALSALTTMMFSASALAKIEEGKLVIIWINGDKGYNGLAIEVGG
	51	KKFEKDTGKIKVTVEHPDKLEEEKFPQVAATGDGPDIIIFWAHDFRGGYAAQSG
101	LLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSIIYNKDLLPN	
151	PKTWEIIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENG	
201	KYDIKDVGVNDAGAKAGLTFLVDLIKXKHMNADTDYSIAEAAFNKGETAM	
251	TINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKE	
301	LAKFEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENA	
351	QKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK	
Band3 - Maltose Binding Protein + PNPLA3	>PNPLA3:1	MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGA
	51	LHCVGLSGIPLAQTLQVLSDLVRKARSRNIGIFHPSFNLSKFLRQGLCK
	101	CLPANVHQLISGKIGISLTRVSDGENVLVSDFRSKDEVVDALVCSCFIPF
	151	YSGLIPPSFRGVRYVDGGVSDNVPFIDAKTTITVSPFYGEYDICKVKST
	201	NFLHVDITKLSRLCTGNLYLLSRAVFPPDLKVLGEICLRGYLDAFRFLE
	251	EKGICNRPQPLKSSSEGMDEPAMPVSWANMSLDSSPESAAALAVRLEGDE
	301	LLDHLRLSILPWDESILDTPRLATALSEEMKDKGGYMSKICNLLPIRI
	351	MSYVMLPCTLPVESAIIVQRLVTLWLPDMPDDVLWLQWVTSQVFTVRLMC
	401	LLPASRSQMPVSSQASPTPEQDWPCWTPCSPKGCPAETKAEATPRISIL
	451	RSSLNFFLGNKVPAGAEGSTFPFSFLESLSL
Band4 - Maltose Binding Protein + PNPLA3	>MBP:1	MKIKTGARILALSALTTMMFSASALAKIEEGKLVIIWINGDKGYNGLAIEVGG
	51	KKFEKDTGKIKVTVEHPDKLEEEKFPQVAATGDGPDIIIFWAHDFRGGYAAQSG
	101	LLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSIIYNKDLLPN
	151	PKTWEIIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENG
	201	KYDIKDVGVNDAGAKAGLTFLVDLIKXKHMNADTDYSIAEAAFNKGETAM
	251	TINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKE
	301	LAKFEFLENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENA
	351	QKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDAQTRITK
	>PNPLA3:1	MYDAERGWSLSFAGCGFLGFYHVGATRCLSEHAPHLLRDARMLFGASAGA
	51	LHCVGLSGIPLAQTLQVLSDLVRKARSRNIGIFHPSFNLSKFLRQGLCK
101	CLPANVHQLISGKIGISLTRVSDGENVLVSDFRSKDEVVDALVCSCFIPF	
151	YSGLIPPSFRGVRYVDGGVSDNVPFIDAKTTITVSPFYGEYDICKVKST	
201	NFLHVDITKLSRLCTGNLYLLSRAVFPPDLKVLGEICLRGYLDAFRFLE	
251	EKGICNRPQPLKSSSEGMDEPAMPVSWANMSLDSSPESAAALAVRLEGDE	
301	LLDHLRLSILPWDESILDTPRLATALSEEMKDKGGYMSKICNLLPIRI	
351	MSYVMLPCTLPVESAIIVQRLVTLWLPDMPDDVLWLQWVTSQVFTVRLMC	
401	LLPASRSQMPVSSQASPTPEQDWPCWTPCSPKGCPAETKAEATPRISIL	
451	RSSLNFFLGNKVPAGAEGSTFPFSFLESLSL	