Genome to Clinic: Viral determinants of clinical outcomes in chronic Hepatitis C Virus infection





Zhiqing Wang

St Catherine's College

University of Oxford

A thesis presented for the degree of Doctor of Philosophy of Clinical Medicine

Trinity Term 2025

Abstract

Hepatitis C virus (HCV) chronically infects more than 50 million people worldwide and remains a leading cause of liver disease. Although direct-acting antivirals can achieve high cure rates, challenges persist due to viral genetic diversity, reinfection risk, and the absence of an effective vaccine. Understanding how viral polymorphisms shape clinical outcomes is therefore critical for identifying prognostic markers, advancing vaccine design and long-term disease management.

The work in chapter 2 examined both viral and non-viral factors associated with antibody responses within HCV genotype 3a (gt3a) infection, aiming to provide insights into HCV vaccine development. Three commonly observed amino acid substitutions and three N-glycosylation motifs (one overlapping with an identified amino acid) were showing significant associations with antibody responses. Additionally, the within-patient nucleotide diversity was also associated with antibody binding against E1/E2. Regarding the non-viral factors, male patients, non-cirrhotic patients, and patients carrying the *IFNL4* gene haplotype P70 had significant lower binding.

In chapters 3 and 4, viral genome-wide association studies and whole HCV genome sequences were used to assess the viral determinants of HCV-related liver disease and viral replication advantage in the context of HCV gt3a and gt1a

infections. Five novel amino acid substitutions in gt3a and three in gt1a demonstrated significant associations with cirrhosis prevalence; these were referred to as Cirrhosis associated Polymorphisms (CAPs) in the chapters. In addition, five polymorphisms in gt3a and six in gt1a were associated with higher viral load and were classified as Higher Viral Load associated polymorphisms (HVLPs). The interplays between the identified determinants and particular host genetic factors (*IFNL4* gene and HLA alleles) were examined to elucidate how host-virus interactions influence immune recognition, viral persistence, and disease outcome. The result indicated that the *IFNL4* genotype influenced cirrhosis prevalence only when the virus carried particular amino acid residues at two gt3a CAP sites. Moreover, three HVLPs conferred replication advantages only in subjects who did not carry the corresponding HLA alleles.

Together, this work demonstrates that viral genetic factors play key roles in chronic HCV infection, providing new insights into HCV pathogenesis and highlighting viral polymorphisms as potential targets for improved prognostic and vaccine strategies.

Declaration

I declare that the work presented in this thesis is my own and has not been submitted for any other degree in this or any other University or Institute of Learning.

Work completed by others: Samples and Data used in chapters 2,3 and 4 was collected as part of the BOSON clinical trial, STOP-HCV-1 clinical trial and HCV Research UK biobank. Nucleic acid extraction, NGS Library construction and sequencing were performed by the Genomics team at Wellcome Centre for Human Genetics including Anthony Brown, Amy Trebes and Paolo Piazza. The bioinformatics pipeline used to generate consensus HCV sequences and variant data was designed and written by David Bonsall, Camilla Ip, Haiting Chai and Azim Ansari. Host genome-wide genotyping was performed by the Genomics team at the Wellcome Centre for Human Genetics. Genotype calling and imputation and principal component analysis were done by Jocelyn Quistrebert and Azim Ansari. The antibody response data presented in chapter 2, derived from ELISA based IgG binding and retroviral pseudotype-based neutralisation assays, were generated by Isla Humphreys and Alex Tarr. All experimental work in chapter 2 were conducted by Alex Tarr and his team at the University of Nottingham. Structure-based mapping analysis in chapter 2 was conducted by Thomas Bowden's team at University of Oxford. Ongoing functional work for chapters 3 and 4 is being carried out by Jens Bukh's team at University of Copenhagen.

Funding: The funding for my DPhil came from CSC-COI MD/PhD High-level Medical Innovative Talent Scholarship and Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (CIFMS). This work was funded by a grant from the Medical Research Council (MRC) (MR/K01532X/1; to the STOP-HCV Consortium). This work was also supported by the HCV Research UK (funded by the Medical Research Foundation) for their assistance in handling and coordinating the release of samples for these analyses.

Acknowledgements

Writing this thesis marks the completion of a long and transformative journey, one that would not have been possible without the support, guidance, and encouragement of many people to whom I owe my deepest gratitude.

First and foremost, I would like to express my sincere thanks to my supervisors, Professor Azim Ansari and Professor Eleanor Barnes, without whose guidance this work would not have been started, let alone completed. I had virtually no formal background in bioinformatics when I began this PhD. Under their mentorship, I gradually developed the skills needed to code, perform statistical analyses, and eventually conduct independent research. Their patience, expertise, and vision have been invaluable in shaping my growth as a scientist and in giving me the confidence to complete this journey.

Secondly, I would like to express my sincere gratitude to my DPhil examiners, Professor David Robertson and Professor Paul Klenerman, whose insightful comments and suggestions helped to improve this thesis.

In addition, I would also like to acknowledge my first doctoral studies in Professor Jiang's lab at Peking Union Medical College (PUMC), China, where my supervisor and colleagues played a pivotal role in fostering my scientific curiosity, independent

thinking, and practical skills in the lab. I will always cherish the memories of my time there and friendships I shared with my dear colleagues there. The experiences and mentorship I gained during that time laid the groundwork for my continued pursuit of research and strengthened my optimism about life.

I must thank Professor Alex Tarr, whose guidance and contributions were essential to the work presented in chapter 2, particularly in conducting the wet-lab experiments. His extensive knowledge of immune responses to viral infections was instrumental in driving the project forward, and without his expertise and support, the publication arising from this chapter would not have been possible.

This thesis would also not have been possible without the invaluable contributions of the BOSON Study team and the STOP-HCV investigators, who collected and provided the samples and data used throughout my work. I am sincerely thankful for access to such an exceptional resource, which has not only made my research possible but continues to enable many important studies in the field.

Finally, I am profoundly grateful to my loving parents, my partner Yukun Shi, and my dearest friend Xiaoyun Li, whose support sustained me throughout my years of study abroad. Choosing to pursue a second PhD was far from an easy decision, requiring not only a major shift in mindset, but also adaptation to a new environment. There were many moments of loneliness and self-doubt, but your deep and warm conversations gave me strength and reassurance to keep moving forward. Your unwavering belief in me has been the true foundation on which this work has been built.

Publications arising

Paper published from this work

Zhiqing Wang, Isla Humphreys, Jocelyn Quistrebert, Haiting Chai, Robert Stass, Josh Dhir, Alexandru Nisioi, Paul Radford, STOP-HCV consortium, Jonathan K Ball, William L Irving, Thomas A Bowden, Paul Klenerman, Eleanor Barnes, Jane A McKeating, Alexander W. Tarr, and M. Azim Ansari (2025). Interplay of Host and Viral Genetic Variations in Modulating Antibody Responses to Genotype 3a Hepatitis C Virus: Implications for Vaccine Design. *Cell Reports*. [Published]

Link to Thesis: This publication corresponds to the work presented in Chapter 2.

Author contributions:
M.A.A., A.W.T., and Z.W. designed the research and wrote the manuscript. Z.W. conducted the analysis to identify factors influencing antibody responses. I.H., J.D., A.N., and A.W.T. performed ELISA and neutralization assays. M.A.A. and A.W.T. supervised the project. Z.W., I.H., J.Q., H.C., R.S., J.D., A.N., P.R., J.K.B., W.L.I., T.A.B., P.K., E.B., J.A.M., A.W.T., and M.A.A. discussed the results, commented on, and provided critical reviews of the manuscript.

Papers in preparation

1. Zhiqing Wang, Van Long Pham, Haiting Chai, Ulrik Fahnøe, Robert Stass, Jocelyn Quistrebert, Yan Chen, George Airey, Santseharay Ramirez, STOP-HCV Consortia, Graham Cooke, John McLauchlan, William L. Irving, Paul Klenerman, Eleanor Barnes, Thomas Bowden, Jens Bukh, and M. Azim Ansari (2025). Virus

amino acid polymorphisms associated with liver disease in patients chronically infected with genotype 3a and 1a hepatitis C virus.

Link to Thesis: This manuscript is based on the work in Chapter 3, which examines viral determinants associated with HCV-related liver disease.

2. Zhiqing Wang, Haiting Chai, Van Long Pham, Ulrik Fahnøe, Robert Stass, Jocelyn Quistrebert, Yan Chen, George Airey, Santseharay Ramirez, Jennifer Benselin, Sarah McDonald, STOP-HCV Consortia, Graham Cooke, Graham Foster, John McLauchlan, William L. Irving, Paul Klenerman, Eleanor Barnes, Thomas Bowden, Jens Bukh, and M. Azim Ansari. (2025). Viral load in patients with chronic hepatitis C: association with virus amino acid determinants.

Link to Thesis: This manuscript is based on the work in Chapter 4, focusing on viral genetic factors influencing viral load and replication advantages.

Author contributions (both papers):
M.A.A. and Z.W. designed the research and wrote the manuscripts. M.A.A. and J.B. supervised the projects. Z.W. conducted the analysis to identify viral determinants associated with HCV-linked liver disease and virus replication. H.C. generated consensus HCV sequences and variant data. V.L.P. conducted functional work. All authors contributed to interpreting the results, commenting on the manuscripts, and providing critical revisions.

Contents

1 Introduction and Background 20

1.1 Hepatitis C Virus 21

1.1.1 Characteristics and Structure 21

1.1.2 Life cycle 22

1.1.3 Genetic diversity and viral evolution 26

1.2 Immune responses in HCV chronic infection 30

1.2.1 Innate immune response 30

1.2.2 Adaptive immune response 32

1.2.3 Implications for vaccine design 36

1.3 Liver disease induced by HCV chronic infection 39

1.4 Viral genome-wide association studies (V-GWAS) 42

1.5 Aims of thesis 44

2 Factors from host and virus in modulating antibody responses of genotype 3a Hepatitis C Virus chronic infection 47

2.1 Summary 48

2.2 Background 49

2.3 Methods 51

2.3.1 Patients and samples 52

2.3.2 E1/E2 antigen and ELISA 52

2.3.3 Neutralisation assays 53

2.3.4 Mutants 54

2.3.5 Sequencing 54

2.3.6 Quantification and statistical analysis 55

2.4 Results 58

2.4.1 Cohort description and measurement of HCV specific antibodies 58
2.4.2 Impact of reference antigen (UKN3A13.6) diversity on antibody responses . 59
2.4.3 Correlation between binding and neutralization 60
2.4.4 Non-genetic factors associated with antibody responses 61
2.4.5 Host genetic factors associated with antibody responses 63
2.4.6 Virus genetic factors associated with antibody responses 66
2.4.7 HCV mutants for validating the function of the identified polymorphisms 74
2.4.8 Viral glycosylation sequons associated with antibody responses 77
2.4.9 The association of intra-patient viral nucleotide diversity and antibody responses 81
2.4.10 Mapping antibody-associated polymorphisms to surface-exposed E1/E2 regions 85
2.5 Discussion 90
3 Viral factors associated with liver disease in patients chronically infected with genotype 3a and 1a hepatitis C virus 98
3.1 Summary 99
3.2 Background 101
3.3 Methods 104
3.3.1 Patients and samples 104
3.3.2 Viral sequencing 106
3.3.3 Host genome-wide genotyping and imputation 106
3.3.4 Host and virus genetic principal component analysis 106
3.3.5 Statistical analysis 107
3.4 Results 109
3.4.1 Cohort description 110
3.4.2 Association between HCV clades and Cirrhosis 111
3.4.3 Viral Genome-Wide Association Study (V-GWAS) for Cirrhosis 115
3.4.4 Covariation of CAPs and their impact on cirrhosis 127
3.4.5 Effects of CAPs on hepatocellular carcinoma (HCC) 140
3.4.6 Effects of CAPs on viral load (VL) 143

3.4.7 Interplay between host *IFNL4* gene and CAPs 146
3.4.8 Interplay between host HLA alleles and CAPs 152
3.4.9 Viral Genome-Wide Association Study (V-GWAS) for HCC 159
3.5 Discussion 161

4 Viral determinants of viral load in patients chronically infected with genotype 3a and 1a hepatitis C virus 167

4.1 Summary 168
4.2 Background 170
4.3 Methods 172
4.3.1 Patients and samples 172
4.3.2 Viral sequencing 172
4.3.3 Host genome-wide genotyping and imputation 173
4.3.4 Host and virus genetic principal component analysis 173
4.3.5 Statistical analysis 173
4.4 Results 175
4.4.1 Viral Genome-Wide Association Study (V-GWAS) on viral load 175
4.4.2 Covariation of HVLPs and their impact on viral load 184
4.4.3 Effects of HVLPs on liver disease 199
4.4.4 Impact of HLA alleles on HVLPs 205
4.5 Discussion 214

5 Concluding Remarks 221

5.1 Summary of main findings 222
5.2 Future directions 224
5.3 Overall contributions and implications 225

Bibliography 227

Appendix A 240

Appendix B 242

List of Figures

Figure 1.1: HCV genome and polyprotein.....22

Figure 1.2: Schematic representation of the HCV life cycle. 24

Figure 1.3: Unrooted Maximum likelihood phylogenetic tree of HCV reference sequence. 27

Figure 1.4: Classical model of antiviral immune response kinetics 33

Figure 2.1: The selection of reference isolate.....59

Figure 2.2: Correlation between antibody response and E1/E2 amino acid divergence from the reference antigen (UKN3A13.6) 60

Figure 2.3: The correlation between binding responses and neutralisation sensitivity....61

Figure 2.4: The impact of non-genetic factors on antibody binding response 62

Figure 2.5: The impact of non-genetic factors on neutralization. 63

Figure 2.6: The impact of *IFNL4* gene haplotypes on antibody binding response.....64

Figure 2.7: The impact of *IFNL4* gene haplotypes on antibody neutralization response. 65

Figure 2.8: Association between HCV amino acid polymorphisms in E1/E2 and antibody binding. 68

Figure 2.9: Association between amino acids at site 653 and antibody response. 69

Figure 2.10: Association between HCV amino acid polymorphisms in E1/E2 and antibody neutralization.....71

Figure 2.11: Association between amino acids at site 501 and antibody response. 71

Figure 2.12: Association between amino acids at site 533 and antibody response. 72

Figure 2.13: Neutralization sensitivity distribution across amino acid variants at site 501 and site 533. 75

Figure 2.14: Infectivity of HCV mutants using E1/E2 Pseudo-particles. 76

Figure 2.15: Locations and amino acid sequences of glycosylation sequons in the E1/E2 proteins in HCV genotype 3a samples in the study sequences. 78
Figure 2.16: Association between variants in glycosylation site N234 and antibody response. 80
Figure 2.17: Association between variants in glycosylation site N476 and antibody response. 81
Figure 2.18: Correlates of intra-patient viral nucleotide diversity and binding. 83
Figure 2.19: Correlates of intra-patient viral nucleotide diversity and neutralization. 84
Figure 2.20: Structural and sequence context of antibody-associated amino acid polymorphisms in HCV E1/E2. 87
Figure 2.21: Structural localisation and epitope context of antibody-associated polymorphisms in HCV E1/E2. 88
Figure 2.22: Location of sites associated with antibody responses in a structure of the E1/E2 heterocomplex. 91
Figure 3.1: Two gt1a clades associated with low cirrhosis prevalence from the HCV gt1a phylogeny.....114
Figure 3.2: Percentage of variance explained by HCV viral PCs. 117
Figure 3.3: Association between HCV gt3a amino acids and cirrhosis status. 119
Figure 3.4: Association between each amino acid residue and cirrhosis status at each of the five sites identified within HCV gt3a infection. 122
Figure 3.5: Association between HCV gt1a amino acids and cirrhosis status. 124
Figure 3.6: Association between each amino acid residue and cirrhosis status at each of the three sites identified within HCV gt1a infection.....126
Figure 3.7: Covariation between pairs of gt3a CAPs. 129
Figure 3.8: Conditional analysis of cirrhosis status after covariate adjustment. 130
Figure 3.9: Cumulative impact of gt3a CAPs on cirrhosis status. 133
Figure 3.10: Covariation between pairs of gt1a CAPs. 136

Figure 3.11: Association test between cirrhosis status and site 2321/938, with and without different adjustments.....137
Figure 3.12: Covariation of gt1a CAPs and their impact on cirrhosis status.....138
Figure 3.13: Association between gt3a CAPs and hepatocellular carcinoma (HCC) status.....141
Figure 3.14: Association between gt1a CAPs and hepatocellular carcinoma (HCC) status.....143
Figure 3.15: Association between gt3a CAPs and log_{10} viral load.....144
Figure 3.16: Association between gt1a CAPs and log_{10} viral load.....146
Figure 3.17: Association between *IFNL4* genotype and gt3a CAPs.....148
Figure 3.18: Association between *IFNL4* genotype and gt1a CAPs.....151
Figure 3.19: Conditional association analysis of HLA alleles with site1444.....155
Figure 3.20: Association between cirrhosis status and residues at site1444 with or without adjusting for HLA A*01:01 or DRB1*04:04.....156
Figure 3.21: Association between HCV amino acids and HCC status.....160
Figure 4.1: Association between HCV gt3a amino acids and log_{10} viral load.....178
Figure 4.2: Association between amino acid residues and viral load at five polymorphic sites in HCV gt3a infection.....179
Figure 4.3: Association between HCV gt1a amino acids and cirrhosis status.....182
Figure 4.4: Association between each amino acid residue and viral load at each of the six sites identified within HCV gt1a infection.....183
Figure 4.5: Covariation between pairs of gt3a HVLPs.....186
Figure 4.6: Conditional analysis of viral load after covariate adjustment.....187
Figure 4.7: Association between log_{10} viral load and multiple gt3a HVLPs.....190
Figure 4.8: Covariation between pairs of gt1a HVLPs.....193
Figure 4.9: Conditional analysis of viral load after covariate adjustment.....194
Figure 4.10: Association between log_{10} viral load and multiple gt1a HVLPs.....197
Figure 4.11: Association between cirrhosis status and gt3a HVLPs.....201

Figure 4.12: Association between HCC status and gt3a HVLPs.....202
Figure 4.13: Association between cirrhosis status and gt1a HVLPs.....204
Figure 4.14: Association between HCC status and gt1a HVLPs.....205
Figure 4.15: Association between log_{10} viral load and variation at NS5A 2216 with or without HLA C*14:02. 207
Figure 4.16: Interaction between HLA-C*14:02 and viral amino acid at site 2216. 209
Figure 4.17: Association between log_{10} viral load and variation at site 2133 and site 2218 with different adjustments.....211
Figure 4.18: Interaction between HLA alleles and gt1a HVLPs. 213
Figure B.1: Distribution of Cirrhosis status (presence vs. absence) across the HCV gt3a phylogeny.....243
Figure B.2: Distribution of Cirrhosis status (presence vs. absence) across the HCV gt1a phylogeny.....245
Figure B.3: Distribution of HCC status (presence vs. absence) across the HCV gt3a phylogeny.....247
Figure B.4: Distribution of HCC status (presence vs. absence) across the HCV gt1a phylogeny.....249

List of Tables

Table 2.1: Top five associations between HLA alleles and binding. 66

Table 2.2: Top five associations between HLA alleles and neutralization. 66

Table 2.3: Top five associations between E1/E2 amino acid variations and antibody binding. 67

Table 2.4: Top five associations between E1/E2 amino acid variations and antibody neutralization. 70

Table 2.5: The frequency of different amino acids at each site (site 653, 501, 533) using 507 gt3a isolates in the BOSON cohort. 73

Table 2.6: The mutant design strategy for E2 sites 501 and 533. 76

Table 2.7: The total count of potential N-linked glycosylation sites detected in each patient. 78

Table 2.8: Association between tested glycosylation motifs at 6 sites and antibody response. 80

Table 3.1: Viral polymorphisms associated with liver disease reported in prior studies. 103

Table 3.2: Demographic and clinical characteristics of the study cohort. 111

Table 3.3: Demographic and clinical characteristics of different clades. 115

Table 3.4: Top ten associations between gt3a HCV amino acid variations and cirrhosis status. 120

Table 3.5: Top ten associations between gt1a HCV amino acid variations and cirrhosis status. 125

Table 3.6: Associations between gt3a CAP combinations and cirrhosis status.....134
Table 3.7: Associations between gt1a CAP combinations and cirrhosis status.....139
Table 3.8: Interaction analysis of NS5A 2057 and *IFNL4* genotype. 149
Table 3.9: Interaction analysis of NS5A 2414 and *IFNL4* genotype. 150
Table 3.10: Top associations (FDR < 5%) between HLA alleles and the residue present at NS3 position 1444. 154
Table 3.11: Interaction analysis of NS3 1444 and HLA A*01:01.....157
Table 3.12: Interaction analysis of NS3 1444 and HLA DRB1*04:04. 158
Table 3.13: The distribution of amino acid at position 1444 across the presence or absence of HLA alleles.....166
Table 4.1: Top ten associations between gt3a HCV amino acid variations and viral load. 178
Table 4.2: Top ten associations between gt1a HCV amino acid variants and cirrhosis status. 182
Table 4.3: Associations between gt3a HVLP combinations and log_{10} viral load.....191
Table 4.4: Associations between gt1a HVLP combinations and log_{10} viral load.....199
Table 4.5: The top five associations between HLA alleles and the residue present at position NS5A 2216.....207
Table 4.6: The top five associations between HLA alleles and the residue present at position NS5A 2133.....210
Table 4.7: The top five associations between HLA alleles and the residue present at position NS5A 2218.....210
Table 4.8: Amino acid distribution at HLA-associated viral sites.....218
Table A.1: Engineered mutants corresponding to CAPs and HVLPs in HCV genotype 3a. 240
Table A.2: Engineered mutants corresponding to CAPs and HVLPs in HCV genotype 1a 241

Abbreviations

HCV	Hepatitis C Virus
E1	Envelope 1
E2	Envelope 2
Non-structural proteins	NS proteins
gt	genotype
HVR	hypervariable regions
DAA	Direct Acting Antiviral
NGS	Next generation sequencing
IFN	Interferon
DCs	Dendritic cells
APCs	Antigen presenting cells
IFNL4	Interferon Lambda 4
nAbs	neutralizing antibodies
TCF-1	T cell factor 1
PD-1	programmed cell death-1
ECM	extracellular matrix
TGF-β	transforming growth factor beta
PDGF	platelet derived growth factor

HSCs ..... hepatic stellate cells

DAMPs ..... damage-associated patterns

HCC..... Hepatocellular Carcinoma

SVR ..... Sustained Virologic Response

HLA ..... Human Leukocyte Antigen

FDR ..... false-discovery rate

VL ..... viral load

CI ..... confidence interval

ISG ..... interferon-stimulated gene

HCVpp..... Hepatitis C Virus pseudoparticles

V-GWAS..... Viral Genome-Wide Association Study

LogOR..... log odds ratio

PC ..... Principal component

CAPs ..... Cirrhosis Associated Polymorphisms

HVLPs ..... Higher Viral load associated Polymorphisms

Chapter 1

Introduction and Background

1.1 Hepatitis C Virus

1.1.1 Characteristics and Structure

Hepatitis C Virus (HCV), identified in 1989 by Michael Houghton's team at Chiron, belongs to the Flaviviridae family, which also includes clinically important viruses such as Dengue and Zika. It is a bloodborne virus primarily transmitted through transfusion of unscreened blood and blood products, unsafe healthcare procedures, sharing devices for injection drug use, and contaminated equipment in tattoo parlors. Mother-to-infant and sexual transmission routes also occur, although they account for only a minor proportion in HCV prevalence [1]. Globally, an estimated 50 million people have chronic hepatitis C virus infection, with about 1.0 million new infections occurring per year, making it a critical public-health concern [2].

HCV virions are 45 - 65 nm in diameter and are enclosed in a lipid bilayer envelope that anchors two envelope glycoproteins, envelope 1 (E1) and envelope 2 (E2). The genome is a positive single-strand RNA of approximately 9.6kb, with an open reading frame (ORF) that encodes a polyprotein of roughly 3,000 amino acids [3]. The amino-terminal part of the ORF encodes the structural proteins (core (C), E1 and E2, Figure 1.1), whereas the remaining portion is subsequently translated and processed into seven non-structural proteins (viroporin p7, nonstructural proteins 2 (NS2), NS3, NS4A, NS4B, NS5A, and NS5B, Figure 1.1) [4]. Low-density lipoproteins (LDLs) and very-low-density lipoproteins (VLDLs) are associated with the formation of HCV virions, generating so called lipo-viro-particles (LVPs). Beyond the HCV core protein and RNA genome, LVPs incorporate apolipoprotein

B (APOB) and other exchangeable apolipoproteins (e.g., APOE, APOC), appearing as large spherical complexes with high infectivity. Non-enveloped viral capsids with limited infectivity may also be detected in the circulation of HCV-infected individuals [5].

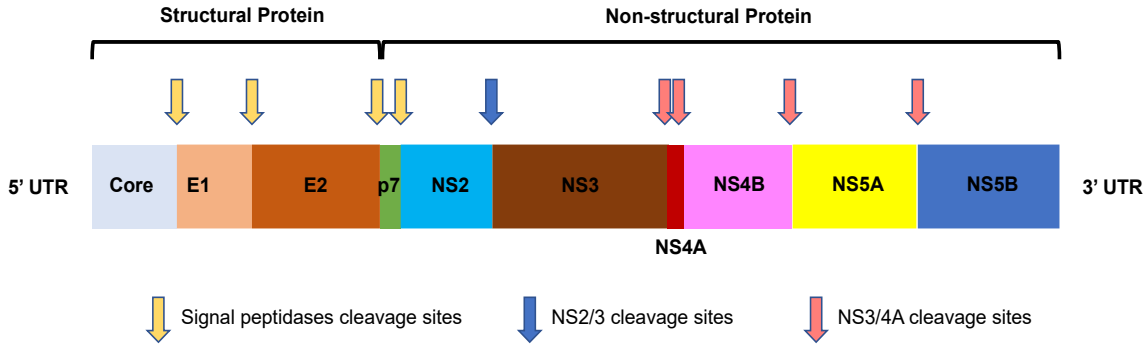


Figure 1.1: HCV genome and polyprotein.

The HCV genome consists of a single open reading frame (ORF) flanked by untranslated regions (UTRs) at both the 5′ and 3′ ends. Translation of the ORF is initiated by an internal ribosome entry site (IRES), resulting in the synthesis of a large polyprotein. This polyprotein is subsequently cleaved into 10 distinct viral proteins. Cellular signal peptidases mediate the cleavage of the core protein from E1, as well as the processing of E1, E2, and p7 (indicated by yellow arrows). The NS2–NS3 protease undergoes self-cleavage (blue arrow), while the NS3 protease–located within the N-terminal third of NS3–along with its membrane-associated cofactor NS4A, processes the remaining non-structural proteins: NS3, NS4A, NS4B, NS5A, and NS5B (pink arrows).

1.1.2 Life cycle

The cellular life cycle of HCV broadly includes seven steps: attachment, entry, uncoating, translation, replication, assembly and release (Figure 1.2) [6]. A

comprehensive understanding of each HCV protein's role in the viral life cycle is critical for advancing antiviral strategies, elucidating pathogenesis, and developing vaccines. Two envelope glycoproteins, E1 and E2, mediate viral entry into permissive host cells, working in concert with apolipoproteins at the surface of LVPs and multiple cell-surface molecules to facilitate initial viral attachment. Low-affinity binding to cells may involve glycosaminoglycans and LDL receptor in the initial attachment stage. Subsequently, E1 and E2 proteins interact with several entry factors–including cluster of differentiation 81 (CD81) and scavenger receptor class B member 1 (SRB1) – followed by claudin 1, occludin and potentially other molecules such as claudin 6, claudin 9, epidermal growth factor receptor (EGFR), and ephrin receptor type A2. These multi-receptor complexes contribute to organ and species tropism and are associated with tight junctions, enabling direct cell-to-cell transmission of the virus [7-9]. After attachment, HCV entry occurs through clathrin-dependent endocytosis, subsequently undergoes fusion between viral and endosomal membranes, releasing the nucleocapsid into the cytoplasm. During this process, the E1 envelope glycoprotein is believed to act as the fusogen that mediates membrane fusion.

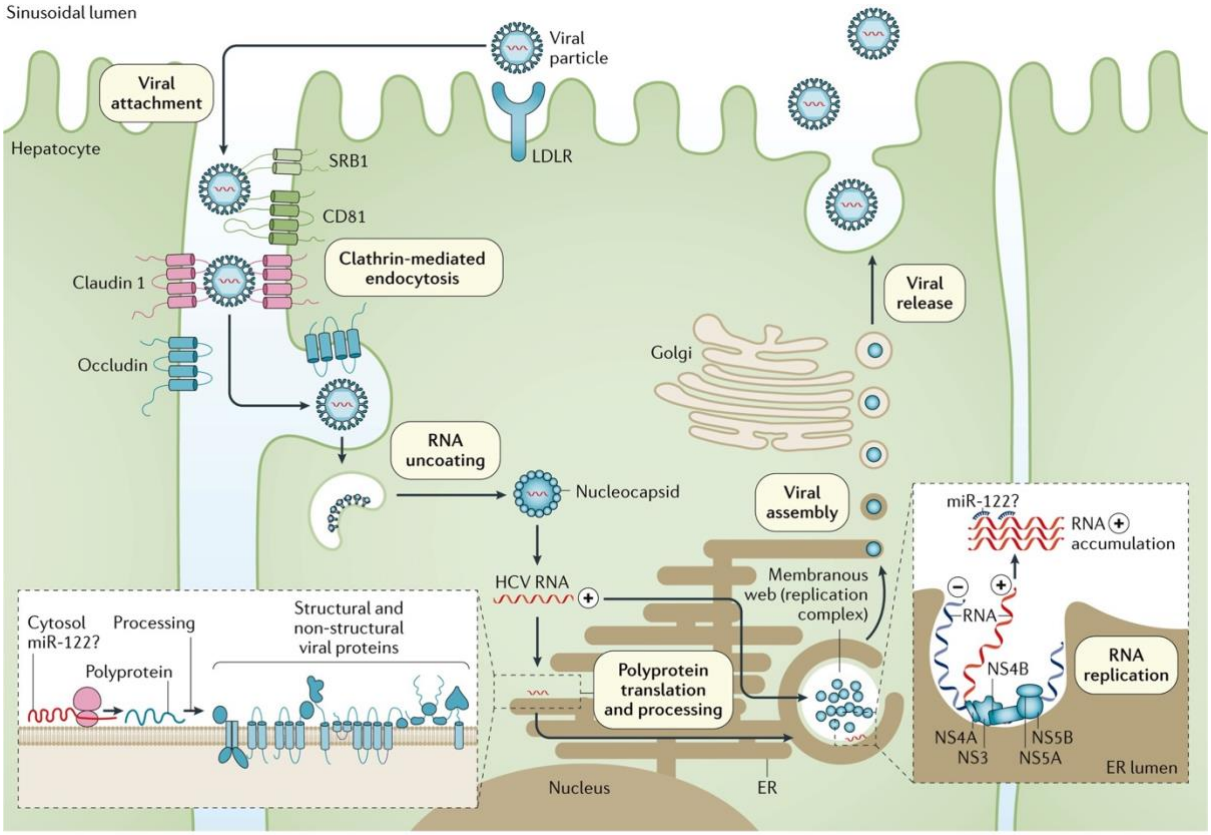


Figure 1.2: Schematic representation of the HCV life cycle.

The life cycle of HCV involves viral attachment, clathrin-mediated endocytosis, polyprotein translation and processing, RNA replication and, finally, viral assembly and release. The figure was reproduced from *Manns, M.P., et al.* (2017) [10]**.**

Following uncoating, the positive, single-stranded RNA released into the cytosol serves as an mRNA template for synthesizing HCV polyprotein. HCV 5' untranslated region (5'UTR) initiates translation of the viral ORF because it harbours an internal ribosome entry site (IRES) [11]. The large precursor polyprotein is synthesised at the endoplasmic reticulum membrane and subsequently cleaved into three structural proteins and seven non-structural proteins, as mentioned previously. Processing of HCV structural proteins requires two host peptidases (signal peptidase and signal peptide peptidase), while cleavage of non-structural proteins depends on two viral peptidases, NS2 and

NS3/4A. These well-characterized proteases have established enzymatic activities and represent major targets for antiviral drug development. After processing, the mature viral proteins remain associated with intracellular membranes [12, 13].

HCV RNA replication occurs through a complex, multi-step mechanism. As the RNA-dependent RNA polymerase of HCV, NS5B catalyses RNA synthesis using the viral genome as a template and starts the process without a primer. NS3 is an enzyme with multiple functions, including *N*-terminal protease activity, *C*-terminal helicase and NTPase activities. The helicase domain of NS3 regulates viral replication by separating nascent and template RNA strands, unwinding RNA secondary structures, and displacing RNA-binding proteins [14, 15]. Unlike NS3 and NS5B, which possess enzymatic activities, NS5A is a multifunctional protein involved in both HCV replication and assembly despite limited structure information. Its dimeric conformation forms a basic channel that contributes to RNA binding, with Domain I and II being essential for replication complex function. The phosphorylation status of NS5A regulates the balance between viral RNA replication and subsequent life cycle stages. However, this dynamic phosphorylation complicates efforts to define clear structure-function relationships [16]. NS4B is an integral membrane protein involved in HCV-induced membrane rearrangements, culminating in the formation of the membranous web–a specialized replication complex that physically compartmentalizes and supports viral RNA synthesis [10]. Once assembled, the replication machinery amplifies the HCV RNA genome, which is subsequently either translated to produce new viral proteins or packaged into nascent virus particles [14].

Various host factors also assist in the HCV life cycle. Cyclophilin A, also known as peptidylprolyl isomerase A, is required for efficient HCV replication by interacting with NS5A and NS5B. A liver-specific miRNA miR-122 also plays important roles in the HCV life cycle. It serves as a multifunctional cofactor for replication by (1) acting as an RNA chaperone or 'riboswitch', binding to two conserved sites in the internal ribosome entry site (IRES), (2) protecting the 5' UTR from exonuclease degradation, and (3) promoting viral translation. During the establishment phase of infection, miR-122 contributes equally to genome stability and translational enhancement. However, its role shifts predominantly toward translational promotion during the maintenance phase. Additional host factors and pathways, such as phosphatidylinositol 4-kinase IIIα and cholesterol and fatty acid biosynthesis pathways, are also involved in HCV replication [17, 18].

1.1.3 Genetic diversity and viral evolution

HCV possesses extraordinary genetic diversity, reflected in its classification into eight major genotypes and more than 90 distinct subtypes (Figure 1.3) [19]. When comparing complete HCV genomes, strains from different genotypes typically differ at 30–35% of their nucleotide sites, whereas strains from different subtypes within the same genotype differ by at least 15% of nucleotide positions [20]. The most common HCV genotypes causing infection globally are 1 and 3 [21]. Genotype 1 is highly prevalent in Europe, America, and far eastern countries, whereas genotype 3 predominates in South Asia [22]. Together, genotypes 2,4,5 and 6 account for less than 10% of global infections [23]. Genotype 7 has been

identified in few people from central Africa [24] and genotype 8 was recently reported in four epidemiologically unlinked patients from the state of Punjab, India [25, 26].

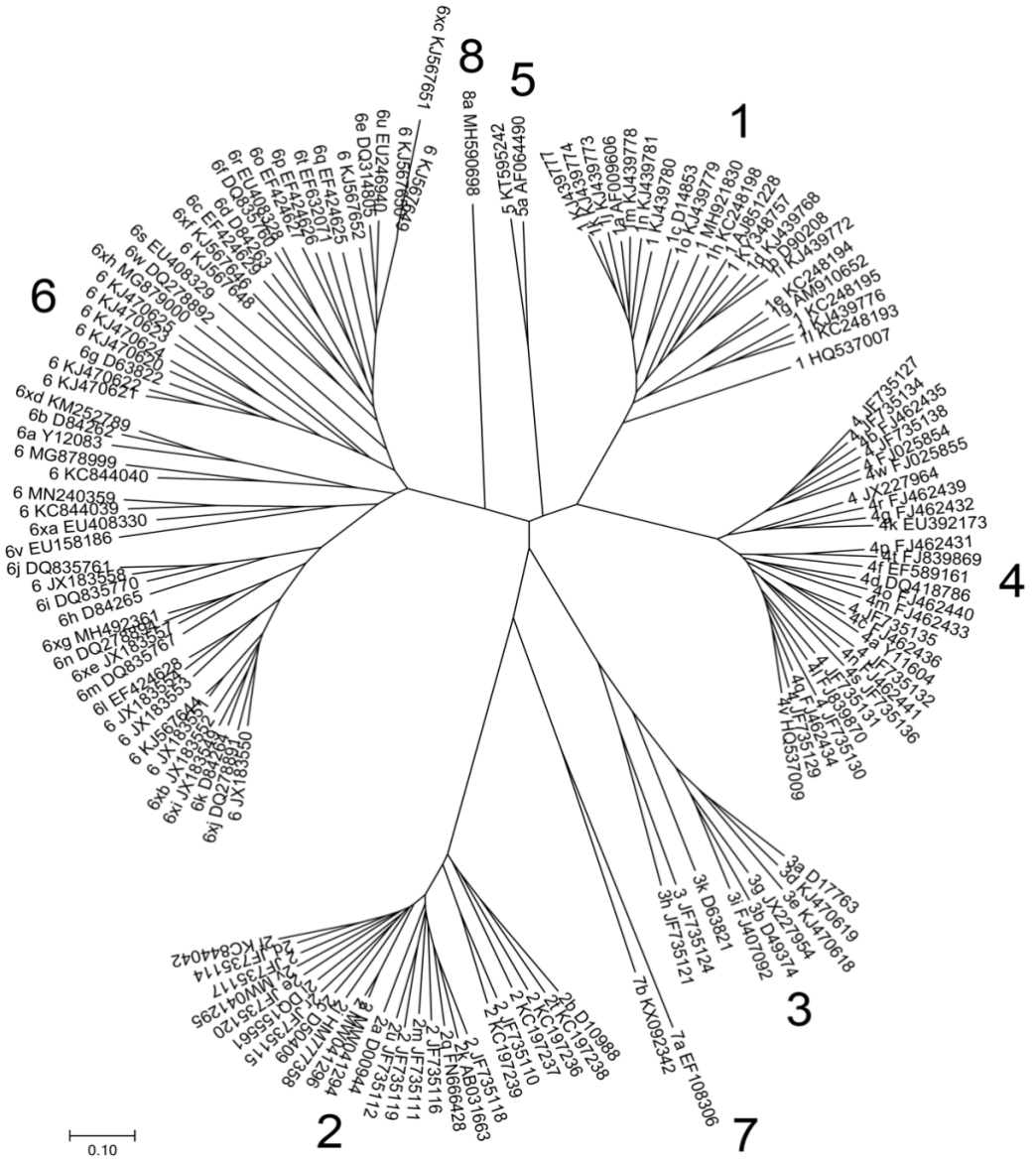


Figure 1.3: Unrooted Maximum likelihood phylogenetic tree of HCV reference sequence.

The figure was reproduced from International Committee on Taxonomy of Viruses (ICTV), 2022 [27].

This enormous diversity arises from the virus's highly dynamic evolutionary process. Like many RNA viruses, HCV exploits multiple mechanisms of genetic variation to ensure its survival. Mutation is the principal source of this variation. HCV replicates using an error-prone, non-proofreading RNA-dependent RNA polymerase (NS5B polymerase), which introduces nucleotide substitutions at an estimated rate of 1.4×10^{-3} to 1.9×10^{-3} per nucleotide site per year [28, 29]. Combined with the virus's extremely high replication rate, this results in the continual production of numerous distinct but closely related viral variants. These variants coexist within each infected individual as a heterogeneous quasispecies population [30-32], characterised by constant turnover, high replication rates and ongoing competition among viral genomes.

Some of the mutations during replication are synonymous; they have no impact on the amino acid sequence of the HCV polyprotein but may influence RNA secondary structure. By contrast, non-synonymous mutations modify encoded amino acids, generating protein variants. The genetic mutations are not evenly distributed across the HCV genome. Regions encoding essential viral functions (translation/replication-associated proteins) or exhibiting strong structural constraints (non-coding 5' and 3' UTRs) are highly conserved. The capsid-encoding region is also well conserved, maintaining approximately 80% sequence identity across viral isolates. E1 and E2 proteins harbour the most variable regions. In particular, the hypervariable regions (HVR1 and HVR2) within E2 can differ by more than 50% between isolates from different individuals [33, 34]. This pattern

reflects a balance between structural constraints and adaptive pressures acting on the virus.

A major driver of this adaptive evolution is host immunity. HLA-restricted T cell pressure selects for mutations that alter epitope presentation or recognition. Numerous HLA alleles have shown associations with viral sequence variation [35] and clinical outcomes of HCV infection [36]. For example, HLA-B27, HLA-B57 and HLA-DQB1*0301 have been reported to be linked to viral clearance, reflecting their ability to present conserved viral epitopes that are difficult for the virus to mutate without incurring a substantial fitness cost. In contrast, escape from other HLA-restricted responses can occur more readily and contributes to viral persistence. Such escape mutations often reduce viral fitness and may therefore be accompanied by compensatory substitutions elsewhere in the genome. In parallel, neutralising antibodies impose selective pressure on exposed regions of E1 and E2, particularly within hypervariable regions, favouring variants that evade antibody binding and facilitate immune evasion [37, 38].

Beyond adaptive immunity, host genetic factors further shape viral diversity. Polymorphisms in genes involved in interferon signalling, natural killer (NK) cell activation, and immune regulation–most notably within the IFNL3/IFNL4 locus–have been strongly associated with differences in viral clearance, persistence, and sequence diversity [39, 40]. Together, these host-driven immune pressures promote continual viral adaptation, contributing to HCV's remarkable genetic heterogeneity and its capacity to establish chronic infection. This extensive viral diversity, in combination with host genetic and immunological factors, has made

effective control of viral spread particularly challenging. Although antiviral therapies have substantially improved treatment outcomes, viral eradication at the population level and the development of a broadly protective preventive vaccine have remained elusive. Understanding these evolutionary processes is therefore essential for interpreting the host–virus interactions and for informing strategies aimed at durable viral control, as explored in subsequent sections and chapters.

1.2 Immune responses in HCV chronic infection

Spontaneous HCV clearance occurs in about 25% of individuals, generally in the first 6 months of infection[41]. HCV clearance during that phase is associated with rapid induction of innate immune responses, particularly interferon-stimulated genes, followed by a delayed but robust adaptive immunity. However, most patients fail to eliminate the virus and progress to chronic infection despite the presence of ongoing innate and adaptive immune response.

1.2.1 Innate immune response

The innate immune response plays an important role in HCV infection. It limits the dissemination of virus by inducing infected hepatocyte apoptosis and by stimulating antigen-specific adaptative immunity. During acute HCV infection, the release of HCV RNA into hepatocyte cytoplasm activates toll-like receptor (TLR3) and retinoic acid-inducible gene-I (RIG-I)-like receptor (RLR) on infected hepatocytes, triggering the generation of interferon type I ((IFN-α and IFN-β)) and

III (IFN-λ). These interferons activate natural killer (NK) cells and inhibit viral replication [42]. Type I and III interferons bind to their respective receptors, leading to phosphorylation of STAT1 and STAT2 and the formation of IFN-stimulated gene factor 3 (ISGF3). ISGF3 is a transcription factor that is essential for generating IFN-stimulated antiviral genes in the nucleus [43, 44]. Notably, hepatocytes are highly susceptible to HCV if type I and III IFN signalling is impaired, as the receptor for type III IFN (IFNLR) is expressed on epithelial cells, hepatocytes, and dendritic cells (DCs) [43, 45]. During HCV infection, IFNs and ISGs are typically upregulated, generating an inflammatory response against viral threats. At the same time, negative regulators such as ubiquitin-specific peptidase 18 (USP18) modulate this pathway. USP18, originally identified as a deISGylase that removes the ubiquitin-like ISG15 protein from substrate proteins, suppresses interferon signalling and thereby contributes to viral persistence [45-47].

NK cells play a paramount role in the cellular innate immune response to acute HCV infection. They act via cytolysis of infected hepatocytes and cytokine production to directly inhibit HCV replication, resulting in an initial reduction of circulating HCV RNA levels [48, 49]. IFN-γ and TNFα produced by NK cells further inhibit viral replication. Activated NK cells also induce perforin- and granzyme B-mediated apoptosis of HCV-infected host cells, although this cytolytic action causes collateral damage to surrounding hepatocytes and tissues. Beyond direct antiviral effects, NK cells help shape the adaptive immune response. Initial control of HCV viral load mediated by NK cells includes the stimulation of adaptive immunity by activating T cells. Virus-specific $CD4^+$ T, $CD8^+$ T and B cells are

induced by antigen presenting cells (APCs), specifically DCs, during adaptive immune response activation. DCs bind to the Nkp30 receptor on NK cells and release IL-12 and IL-15, which activate NK cells; in turn, activated NK cells secrete IFN-γ and TNFα, reciprocally promoting the maturation and antigen presentation [42]. The subsets of DCs such as $CD11c^+$ myeloid DC (mDC1) and $CD141^+$ myeloid DC (mDC2) release IL-12, promoting differentiation of $CD4^+$ and $CD8^+$ T cells and thereby strengthening adaptive immunity [42, 50].

1.2.2 Adaptive immune response

Humoral response

Adaptive immune responses, including humoral antibodies and T cell immunity, play a central role in controlling HCV infection [51]. Virus-specific antibodies typically appear approximately 8–12 weeks after early acute HCV infection (Figure 1.4). Most patients acutely infected with HCV produce antibodies against epitopes within both structural and non-structural proteins; however, most of them often show limited antiviral activity due to the high genetic variability of the virus and the presence of diverse viral population within a single patient. Only a small fraction of antibodies, known as neutralizing antibodies (nAbs) [52], provide protective activity through effectively inhibiting viral binding, entry or uncoating [53].

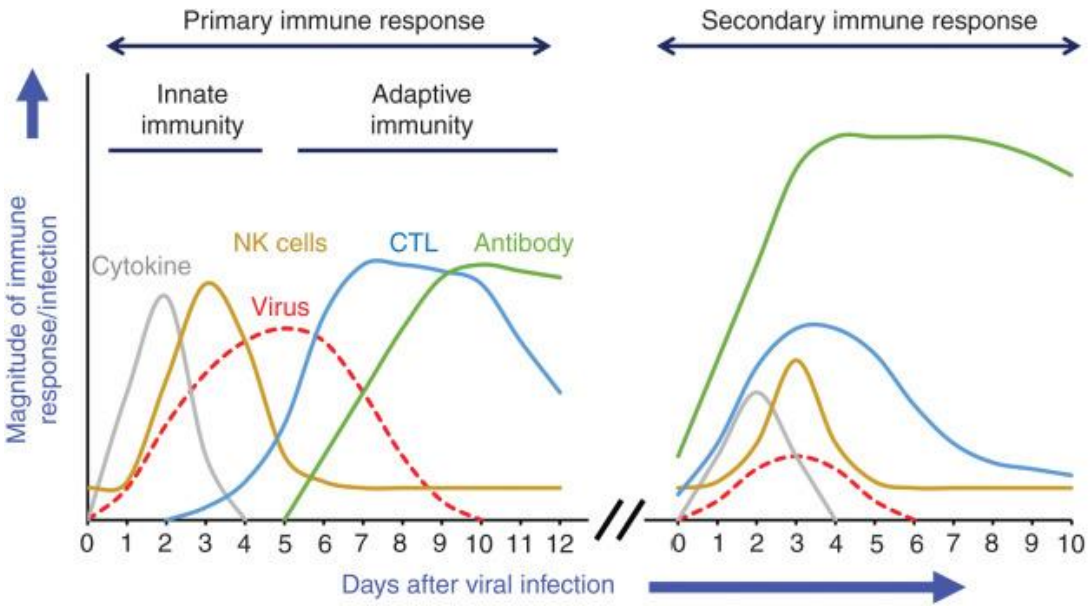


Figure 1.4: Classical model of antiviral immune response kinetics

During a primary infection, innate immunity and early virus-specific cytotoxic T lymphocyte (CTL) activity are crucial for limiting viral replication, while antiviral antibodies appear later in the course of infection. Circulating antibodies then decline gradually but can remain detectable for decades after primary infection or vaccination. In a secondary infection with the same virus, pre-existing neutralizing antibodies reduce the initial viral load, and memory B and T cells are rapidly reactivated to achieve swift and effective control. Antibody responses to HCV do not conform to this classical pattern. The figure was reproduced from *Law, M., et al.* (2021) [54]**.**

Generally, when the host is re-exposed to the same pathogens, pre-existing antibodies can immediately lower pathogen load, while the memory B-cell response is recalled rapidly to produce high levels of antibodies to combat the pathogens. However, nAbs generated during a previous HCV infection are usually insufficient to protect against reinfection with a different HCV strain or genotype due to the high mutation rate of HCV [55, 56]. HCV can mutate rapidly to escape

these nAbs, and only a small fraction of patients can produce truly cross-nAb responses (estimated <10%) [57-59]; even in these cases, serum antibody levels remain relatively low. All known nAbs target epitopes within the HCV envelope glycoproteins E1 and E2, whose hypervariable regions enable rapid development of resistance to broad nAbs [60, 61].

Taken together, it has therefore been widely thought that nAbs play only a limited role in controlling HCV replication, particularly during an acute infection phase, as spontaneous viral clearance can occur without producing sufficient levels of nAbs. Nevertheless, multiple studies have reported potential correlations between early nAbs production and spontaneous virus clearance [57, 62, 63]. Because HCV-specific antibodies decay rapidly following viral clearance, the maintenance of nAbs may require continual stimulation of antigen-specific B cells. Indeed, an increase in the frequency of HCV E2-specific B cells is observed as acute HCV infection progresses to the persistent phase [64].

T cell immunity

Most data are available on T cell responses, and there is broad consensus that HCV-specific T cells play a central role in the control of infection. Broad, multi-specific CD4+ and CD8+ T cell responses are linked to spontaneous HCV clearance, even in cases where viral resolution occurs after 6 months or oscillating viremia [65-67]. Both $CD4^+$ and $CD8^+$ T cells target distinct epitopes across the HCV polyprotein, with a particular focus on non-structural proteins that are critical for viral replication [68]. HCV-specific $CD8^+$ T cells are the primary effector cells. They express activation surface markers and inhibitory receptors such as

programmed cell death-1 (PD-1), whose expression declines with resolution of infection [69]. This downregulation of PD-1 promotes increased expression of CD127 (the alpha-chain of the IL-7 receptor), which is required for T cell homeostatic proliferation, and higher CD127 expression on HCV-specific $CD8^+$ T cells is associated with viral elimination [70-72]. In addition, the expression of the transcription factor T-bet in activated $CD8^+$ T cells regulates the production of IFN-γ and cytotoxic molecules [73], which inhibit HCV replication and mediate the cytolysis of infected hepatocytes. $CD4^+$ T cells provide essential help for priming CD8+ T cell responses and activating DCs through the production of IL-2 and IFN-γ, which also play crucial role in acute and chronic HCV infection [74, 75].

Progression from acute to chronic HCV infection is often attributed to insufficient or early loss of T cell responsiveness [76]. Two main mechanisms result in the failure of T cell immunity: T cell exhaustion and viral escape [77]. Generally, T cell exhaustion occurs under conditions of prolonged antigen exposure and persistent inhibitory signaling, leading to a profound loss of T cell function. This exhaustion is antigen-specific, affecting only T cells directed against the corresponding epitopes. Additional features of exhausted T cells include high co-expression of inhibitory receptors, metabolic changes, impaired survival, altered gene expression profiles and epigenetic modifications [78]. Indeed, these features have been observed in most HCV-specific $CD8^+$ T cells isolated from individuals with persistent infection, which include distinct subsets such as TCF-1^+CD127^+PD-1^+ memory-like cells and PD-1^{hi}Eomes^{hi}TOX^{hi}CD127-terminally exhausted subset [79]. Two transcription factors, TCF-1 (T cell factor 1) and TOX (thymocyte selection associated high mobility group box), play important roles in the biology of T cells exhaustion. The

overexpression of TOX improves the survival of progenitor exhausted $CD8^+$ T cells that is also characterized by co-expressing TCF-1, while TCF-1 helps maintain the exhausted T cell pool [80-82]. Viral escape mutations in the highly mutable HCV genome also emerge under CD8+ T cell pressure, typically within the first six months of infection. These mutations can disrupt antigen recognition through altering proteasomal processing, reducing peptide-MHC class I binding affinity, or impairing recognition by the T cell receptor. Collectively, these changes enable HCV to evade CD8+ T cell surveillance and promote viral persistence [83, 84].

The above mechanisms of CD8+ T cell failure are directly associated with CD4+ T cell response. Broad, long-lived HCV-specific CD4+ T cells are associated with HCV spontaneous clearance. However, in patients with chronic infection, HCV-specific CD4+ T cells are either short-lived or present at very low frequencies, display impaired proliferative and cytokine production, and express high levels of inhibitory receptors such as PD-1 and CTLA4 (cytotoxic T-lymphocyte associated protein 4) that inhibit CD8+ T cell function [85, 86]. The mechanisms underlying the failure to sustain CD4+ T cell responses in persisting infection remain unclear.

1.2.3 Implications for vaccine design

Though the discovery of directly acting antiviral drugs (DAAs) has led to a dramatic increase in HCV cure rates, it is unlikely that the global HCV burden will be eliminated in the near future. DAA treatment faces several challenges: (a) HCV can remain asymptomatic for years, leading to undiagnosed cases and ongoing liver damage during this period [87]; (b) DAAs do not prevent reinfections [88]; (c)

the emergence of DAA-resistant viruses; (d) HCV remains widespread in developing nations and among marginalized groups, where barriers to testing and care hinder effective management [89]. Therefore, there remains a critical need for a widely available and effective vaccine [23, 90]. Such a vaccine would need to robust and broadly induce both humoral and cell-mediated immunity. Traditional approaches, including the generation of live-attenuated or inactivated whole viruses preparations, have faced great challenges. These include limited in-vitro systems supporting full HCV replication and significant safety concerns regarding the translation of cell-culture derived HCV vaccines to humans [91]. Alternative strategies have been attempted, including (a) recombinant protein vaccines with adjuvant designed to elicit humoral responses and (b) DNA, synthetic peptides and nonreplicating viral vectors designed to stimulate T cell immunity [92-95].

One of the most advanced protein-based vaccines employs recombinant form of full-length E1E2 envelope (Env) glycoproteins from the genotype 1a strain HCV-1 (Chiron vaccine). This candidate has progressed to human phase I studies after demonstrating strong immunogenicity and good antibody response against homologous or heterologous HCV in animal models [96-98]. In a subsequent clinical trial (NCT00500747) of the Chiron vaccine, induction of nAbs and proliferative T-cell responses against E1E2 Env in healthy human volunteers was validated [99, 100]. However, cross-nAb responses were elicited in only a few subjects, and the majority of the nAb responses remained isolate-specific and directed toward variable regions of the E1E2 proteins [101].

Within the field of prophylactic T cell vaccines, two vaccine candidates have advanced to human preclinical and clinical trials. The first candidate was a prototype vaccine comprising HCV core protein formulated with the ISCOMATRIX adjuvant, which elicited T cell responses in only 25% of healthy participants and therefore was not progressed to testing in at-risk populations [102]. In contrast, the second T cell vaccine candidate used a heterologous prime-boost vaccination strategy based on a replicative-defective simian adenoviral vector (ChAd3) and a modified vaccinia Ankara (MVA) vector encoding non-structural proteins [95, 103]. Early proof of principle using this technology was demonstrated in a small chimpanzee study showing that adenoviral vectored vaccines encoding an HCV genotype 1b antigen could suppress acute viraemia [93]. A subsequent study in healthy volunteers also showed strong priming of CD4+ and CD8+ HCV-specific T-cell responses [95, 103]. However, these T cells were not cross-reactive across all genotypes and failed to induce protective effects against chronic HCV infection in high-risk people who inject drugs (PWIDs; NCT01436357) [104-106].

While disappointing, the key findings of this vaccine trial are not entirely unexpected, as they align with HCV's known ability to evade adaptive immunity—whether naturally induced or vaccine elicited—and establish persistent infection. These results underscore the need to better understand the factors that drive effective immune responses. Clearly, HCV genetic diversity as discussed previously represents a major hurdle to generate pan-genotypic HCV vaccines. Nevertheless, new approaches are seeking to overcome the limitations, including the use of modified E2 sequences with deletion of HVR [107] and conserved segment T cell immunogens expressed in viral vectors [108]. Advances in defining

neutralizing epitopes, together with the resolution of E1/E2 structures, will help optimize HCV glycoprotein immunogens by stabilizing their conformation and exposing key neutralization sites. The success of RNA vaccines against the SARS-CoV-2 pandemic provides insights into HCV vaccine development, indicating the feasibility of encoding envelope genes from all major HCV genotypes within RNA platforms and potentially enabling broader and more potent immune responses [91].

1.3 Liver disease induced by HCV chronic infection

Currently, liver disease is the eleventh-leading cause of death in the world. Within this burden, liver cirrhosis alone demonstrates striking regional disparities in mortality: it now stands as the tenth-leading cause of death in Africa (rising from thirteenth position in 2015), ninth-leading in both South-East Asia and Europe, and fifth-leading in the Eastern Mediterranean region [109]. For several decades, HCV chronic infection has remained a leading aetiology of liver cirrhosis and its associated complications that include portal hypertension, impaired synthetic function, liver failure, and HCC.

Building on this epidemiological perspective, it is essential to consider the cellular and molecular processes underlying HCV-associated liver disease. Generally, liver fibrogenesis is featured by the progressive accumulation of extracellular matrix (ECM), which distorts the physiological architecture of the liver [110]. Pathogenetically, viral infection, metabolic factors, or toxins can induce liver damage, resulting in a wound-healing process known as fibrosis leads to excessive

matrix deposition. The damaged hepatocytes and infiltration of immune cells secrete many profibrogenic growth factors and proinflammatory cytokines, including transforming growth factor beta (TGF-β) and platelet-derived growth factor (PDGF), activating the trans-differentiating of hepatic stellate cells (HSCs) into collagen-producing myofibroblasts [111, 112]. In short-term injury, these myofibroblasts normally resolve by apoptosis or inactivation to prevent persistent fibrosis [113]. However, persistent HSC activation disrupts the balance between ECM deposition and dissolution, which results in progressive liver fibrosis and ultimately advanced liver disease [114]. As activated HSCs are the predominant precursors of myofibroblasts in fibrotic liver (accounting for more than 90% of collagen-producing cells), several signalling pathways have been suggested to be highly associated with their activation, including TGF-β signalling, PDGF signalling, oxidative stress, the inflammasome (NLRP3)- Caspase1 pathway and WNT/β-catenin signalling [115].

Once cirrhosis has developed, the liver enters a precancerous state that substantially increases the risk of HCC. Within this altered microenvironment, repeated cycles of hepatocyte death and regeneration promote genomic instability, while oxidative stress, inflammatory cytokines, and angiogenic signals drive the emergence of dysplastic nodules [116, 117]. Over time, these nodules may accumulate additional genetic and epigenetic changes and progress to early HCC, which can then evolve into advanced, clinically detectable HCC. This stepwise process explains why cirrhosis is the single strongest risk factor for HCC and underscores the need for regular surveillance in cirrhotic patients [118].

Beyond these general mechanisms, chronic HCV infection adds additional layers of complexity. Hepatocyte death induced by chronic HCV infection leads to release of damage-associated patterns (DAMPs) that can directly activate HSCs [119-121]. More importantly, chronic inflammation driven by persistent antiviral immune responses remains the primary driver of myofibroblast activation and ECM deposition [122]. Given that, the host immune response to HCV infection enhances the secretion of multiple growth factors, inflammatory cytokines, and chemokines by Kupffer cells and lymphocytes, which further stimulate HSC activation and fibrogenesis [123, 124]. HCV core protein and non-structural proteins themselves can modulate fibrogenesis through multiple pathways [115]. Although DAAs remarkably reduce the risk of liver cirrhosis and hepatocellular carcinoma (HCC), likely by partially suppressing viral pro-fibrogenic and carcinogenic activities, patients with pre-existing advanced liver damage remain at residual HCC risk after sustained virologic response (SVR), possibly due to irreversible epigenetic or microenvironmental changes [125-127].

Not all individuals with chronic HCV infection progress to cirrhosis or HCC at the same rate. This marked inter-individual variability in disease outcome suggests that factors beyond cellular pathways and immune responses also play a critical role. The impact of host genetic factors on liver disease progression has become increasingly evident in recent years. In the context of HCV chronic infection, polymorphisms such as *PNPLA3* rs738409 (C>G), variants in *IFNL4* gene, and particular HLA alleles have shown strong associations with the risk of HCV-induced liver disease [128-130]. However, relatively few studies have examined viral determinants to date; only limited evidence suggests that specific amino acid

polymorphisms within the core, NS3 and NS5A regions of HCV gt1b showed association with cirrhosis and HCC risk [131, 132]. Given this gap in knowledge, systematic analyses using multidimensional cohorts integrating clinical outcomes with host and viral sequencing data are therefore required to comprehensively investigate these viral contributions.

1.4 Viral genome-wide association studies (V-GWAS)

Genome-wide association studies (GWAS) are a hypothesis-free approach designed to identify genetic variants associated with traits of interest by testing hundreds of thousands to millions of common variants across the entire genome [133, 134]. Originally developed in human genetics, GWAS have been conducted to elucidate the genetic architecture of complex traits, confirm heritability estimates, and identify novel biological pathways and therapeutic targets. In a typical framework, GWAS test associations between single-nucleotide polymorphisms (SNPs) and phenotypes – such as major human diseases – using linear or logistic regression. Effects are typically expressed as either betas for quantitative traits or odds ratios for case-control studies. Generally, only the main effects of individual SNPs are estimated, as identifying epistatic interactions between SNPs and SNP-environment interactions remains methodologically challenging due to the additional burden of multiple testing [135, 136].

Advances in high-throughput sequencing technologies have enabled the extension of GWAS methodologies beyond the human genome to microorganisms, including

bacteria, protozoa, and viruses [137]. Microbial GWAS shift the unit of analysis from host genetic variation to pathogen genetic diversity, with the aim of identifying microbial variants that influence clinically relevant outcomes such as drug resistance, virulence, transmissibility, and disease progression [138]. While conceptually similar to human GWAS, microbial GWAS require methodological adaptations to accommodate fundamental differences in genomic data sources, data scale, and quality control protocols. A successful application of microbial GWAS identified genetic determinants of pyomyositis in Staphylococcus aureus [139], revealing that the presence of Panton-Valentine leucocidin (PVL) locus increased the risk of the condition. In another study [140], variations in *Streptococcus pneumoniae* explained large amounts of the invasiveness potential but had no effect on severity of pneumococcal meningitis. Furthermore, microbial GWAS was used to determine vaccine candidate coverage across 2083 Group A *Streptococcus* (GAS) genomes [141].

Viral GWAS represent a distinct subset of microbial GWAS, characterised by compact genomes, rapid mutation rates, and substantial within-host genetic diversity [142, 143]. To date, GWAS approaches have been successfully applied to viral pathogens focusing directly on viral genomes. A notable example is provided by genome-to-genome analyses of HIV infection, in which significant associations were observed between host genetic variation and multiple HIV-1 amino acid variants influencing plasma viral load [144]. In another study, viral GWAS have been used to identify novel and known markers associated with HIV drug resistance [145].

In the context of HCV infection, a systematic genome-to-genome study of 542 individuals with chronic HCV was performed to identify human genes driving viral polymorphisms [146]. Additionally, viral GWAS was also used to identify novel HCV polymorphisms associated with sofosbuvir treatment outcome [147]. By moving beyond candidate-gene approaches, viral GWAS provides a powerful, interpretable, genome-wide methodology to systematically characterise the viral genetic factors contributing to the observed heterogeneity in HCV clinical outcomes–an approach utilized extensively in the results chapters of this thesis.

1.5 Aims of thesis

Collectively, the clinical outcomes of the HCV chronic infection remain a major global health concern. While DAAs can cure most cases, reinfection, the lack of prognostic strategies for progression diseases, and the absence of a licensed vaccine continue to pose challenges. Clinical outcomes in HCV infection are shaped by a complex interplay between viral replication, host immune responses, and the development of liver disease, but viral genetic factors driving these outcomes remain incompletely defined. For example, polymorphisms in the E1/E2 envelope proteins can influence viral entry efficiency, neutralization sensitivity, and immune evasion, whereas changes in the NS5A region are linked to treatment response and viral replication capacity. Such viral diversity, particularly the extreme sequence variability in the E1 and E2 coding regions, has been a major barrier to developing a broadly protective vaccine that elicits neutralizing antibodies against the full spectrum of circulating HCV isolates. Understanding these viral

determinants is therefore critical not only for vaccine development but also for predicting disease progression and guiding the next generation of antiviral strategies. Despite this importance, no large-scale, systematic study has comprehensively examined viral genetic variations in relation to immune responses, liver disease, and viral fitness across well-characterized patient cohorts.

To address this gap, this thesis integrates multi-cohort datasets with next generation sequencing to identify viral genetic determinants associated with different clinical outcomes. Each results chapter focuses on one of these key areas:

1. The first results chapter (Chapter 2) aims to provide insights into HCV vaccine development design by:
    1) Identifying host and viral genetic factors that modulate antibody binding and neutralization responses.
    2) Accessing their effects on antibody responses.

2. The second results chapter (Chapter 3) focus on clarifying the potential mechanisms contributing to the etiology of HCV-related liver disease from viral genetic perspective by:
    1) Identifying the amino acid polymorphic sites that influence the HCV-related liver disease.
    2) Investigating the interplay of host genetic factors and the identified viral polymorphisms.
    3) Clarifying how liver disease associated viral determinants directly, or via host modulation, influence disease progression.

3. The third results chapter (Chapter 4) aims to identify HCV replication associated viral polymorphisms:

1) Identifying amino acid polymorphic sites associated with differences in HCV RNA levels.
2) Investigating host - viral interactions.
3) Elucidating mechanisms of how viral polymorphisms directly, or via host modulation, influence viral replication.

Chapter 2[1]

Factors from host and virus in modulating antibody responses of genotype 3a Hepatitis C Virus chronic infection

[1] **The work presented in this chapter has been published in *Cell Reports*.**

2.1 Summary

Background and Aims: Hepatitis C virus (HCV) exhibits significant genetic diversity and is a cause of severe liver complications. The viral envelope glycoproteins E1 and E2, key targets for neutralizing antibodies, are highly variable. The factors from host and virus that influence antibody binding and neutralization remain poorly understood, posing a hurdle for HCV vaccine development. This chapter reports how host and viral genetic factors modulate antibody responses among patients infected with gt3a HCV, providing insights for designing broadly effective vaccines.

Methods: 60 patients used in this study were from BOSON clinical trial. Baseline plasma samples were collected to examine the polyclonal antibody responses. Entire consensus genome sequences for HCV were generated and only the E1/E2 sequences were analysed for the association with antibody responses by using linear regression. The relationship between host genetic factor (*IFNL4* genotypes and HLA molecules) and antibody responses was also investigated in this work.

Results: Non-genetic factors including sex, cirrhosis status and prior IFN treatment were found to be associated with antibody binding response. Host variants in *IFNL4* gene were associated with antibody binding. With respect to viral factors, polymorphisms at three specific E1/E2 amino acid positions and sequons at three N-glycosylation sites (including one site common to both analyses) were significantly correlated with antibody binding and/or neutralisation sensitivity. Furthermore, greater intra-patient diversity within the E2 hypervariable region 1 (HVR1) associated with stronger antibody binding.

Conclusions: This chapter demonstrated that genetic variation in both the virus and the host showed impact on the potency of the antibody response to HCV infection, and these findings have the potential to inform the development of more effective and personalized HCV vaccines.

2.2 Background

Despite the great success of direct acting antiviral therapies (DAA) [148], progress on developing an effective vaccine has been slow [149]. Virus-neutralizing antibodies target the HCV-encoded E1 and E2 glycoproteins that mediate viral entry [62, 150]. Though the molecular targets of these antibodies are well defined [149], the antibody correlates of protection remain unclear. The E2 protein is the major receptor binding protein that directly engages with the host cell membrane entry factors [151-155], while E1 protein is proposed to stabilize the structure of E2 [156], and contributes to receptor binding and fusion events that lead to delivery of the virus genome into a cell [155, 157]. The E1/E2 complex possesses a novel structure with E1 performing a ‘grasping’ mechanism to interact with the stem of E2 [156, 158]. This structure is consistent with the E2 protein being the receptor binding protein [159], which is known to interact with the molecules CD81 and scavenger receptor class B type 1 (SR-B1) on the surface of hepatocytes [152, 160].

Understanding the functional consequences of polymorphisms in the viral proteins informs the development of new antivirals and vaccines. Even minor differences in amino acid sequence can result in considerable functional changes. Genetic variability is not uniform across the virus genome, with extreme diversity exhibited in the E1 and E2 coding genes. This has provided a barrier to development of a vaccine that can elicit antibodies capable of neutralizing the wide range of *in vivo* HCV isolates [161-163]. This chapter primarily employed a genetic screening approach to uncover new virus amino acid variants in the HCV E1/E2 coding region linked to anti-E1/E2 antibody responses in patients with HCV genotype 3a infection, unveiling three commonly observed amino acid substitutions in the HCV E2 protein that associated with anti-E1/E2 antibody binding and neutralisation sensitivity.

Glycosylation is a major and crucial co- and post- translational modification of proteins that leads to glycoprotein formation by means of various glycopeptide linkages [164]. The β-glycosylamine linkage of N-acetylglucosamine (GlcNAc) to asparagine (N-glycosylation) represents the most widely distributed carbohydrate–peptide bond and is the most common type of glycosylation. The binding and cleavage of the sugar-amino acid linkage are critical to the biological activity of glycoproteins. Alterations in glycosylation are associated with diverse diseases. The HCV E1 and E2 proteins are extensively glycosylated with N-linked glycans, with up to 5 sites on E1 and up to 11 sites on E2, accounting for approximately one-third of the heterodimer molecular mass [165]. Despite the high sequence heterogeneity, many of the N-glycosylation sites in the E1 and E2 proteins are conserved among the various HCV genotypes and are important for evasion of antibody binding[166, 167]. In this chapter, to further reveal the role of N-

glycosylation in antibody response, all N-glycans observed in the E1/E2 region were tested for association with antibody binding and neutralization, demonstrating that distinct N-glycan motifs present at three specific N-glycosylation sites (one overlapping with an identified amino acid site) in E1/E2 were associated with antibody responses.

Host genetic variation is one of the factors that drives antibody responses. Genetic variations in the interferon lambda 4 (*IFNL*4) gene are associated with the spontaneous clearance and IFN-based treatment response of HCV infection [168, 169]. Two polymorphisms in *IFNL4* gene, rs11322783 [ΔG > TT] and rs117648444 [G > A], create four haplotypes [170], two that do not produce IFNλ4 protein (TT/G or TT/A: IFNλ4-Null) and two that express IFNλ4 protein variants (ΔG/G: IFNλ4-P70 and ΔG/A: IFNλ4-S70). Patients with the IFNλ4-Null and IFNλ4-S70 variants exhibit lower hepatic interferon-stimulated gene expression, which is associated with increased viral clearance and response to IFN-based therapy compared to those with the IFNλ4-P70 variant [171]. Additionally, genetic variation in the HLA region has been reported to be associated with spontaneous clearance of HCV infection [172]. In this work, the impact of host *IFNL4* gene variants and HLA alleles on anti-E1/E2 responses were also examined, identifying an association between the *IFNL4* haplotype related to production of IFNλ4-P70 and lower levels of antibody response.

2.3 Methods

2.3.1 Patients and samples

The study was conducted in collaboration with STOP-HCV, a consortium sponsored by the Medical Research Council, United Kingdom, which contributed to study design. This study used samples from 60 patients with HCV infection selected from BOSON clinical trial [173], an open-label, randomized phase 3 study evaluating the efficacy and safety of sofosbuvir plus ribavirin, with or without peginterferon-alfa, in 592 DAA treatment-naïve individuals with genotype 2 or 3 HCV infection. Of the 60 selected subjects, 54 were infected with HCV gt3a. The BOSON study protocol was approved by each institution's review board or ethics committee before study initiation (registration number: NCT01962441). ELISA based IgG binding and retroviral pseudotype-based neutralisation assays were implemented using samples recovered at baseline, without DAA treatment. All patients provided written informed consent before undertaking any study-related procedures. The study protocol was approved by each institution’s review board or ethics committee before study initiation.

2.3.2 E1/E2 antigen and ELISA

The polyclonal antibody response directed to the HCV E1 and E2 envelope glycoproteins were detected using well characterized antigen representing genotype 3a (UKN3A13.6 [174], also coded as UKNP3.2.1 in some studies when included in panels of patient-isolated clones, essentially as described previously [59]). Briefly, wells of Maxisorp 96-well plates were coated with *Galanthus nivalis* agglutinin (GNA) (Sigma; 1μg·mL^{-1}) and blocked with PBS containing 5% BSA.

Cell lysates of HEK293T cells transfected with constructs expressing the genes encoding these E1/E2 proteins were recovered 72 hours after transfection and diluted 1:5 in PBS containing 0.05% Tween (PBST) before adding to coated wells. After washing, patient plasma samples were added, diluted in PBST-5% BSA, between 1:100 and 1:10000 to titrate seroreactivity. A dilution of 1:300 was subsequently selected for further experimentation. Bound antibody was detected using anti-human Ig HRP and 3,5,3′,5′-tetramethylbenzidine (TMB) substrate was used to detect bound antibodies and the reaction was stopped after 15 minutes. The ELISA experiments were performed in duplicate and for analysis we used the mean for each sample.

2.3.3 Neutralisation assays

Murine Leukaemia Virus cores pseudotyped with the HCV E1/E2 genes representing isolate UKN3A13.6 were created as previously described [175], using a luciferase reporter construct [176]. Virus pseudotypes were incubated with sera (heat-inactivated prior to use at 56°C for 30 minutes) for one hour at 25°C before adding to target HuH7 cells for four hours at 37°C, in a 5% CO_2 incubator. Following infection, media was exchanged, and cells incubated for a further 72 hours. Following cell lysis, infection was assayed using a Promega luciferase assay substrate. Luciferase activity was measured, and data were normalized to an uninhibited control and a preparation of pseudotypes created without a viral glycoprotein. Percentage values were calculated by normalizing to an uninhibited infection (100%) and the signal generated in an assay performed with pseudotypes

lacking viral glycoprotein (0%). The neutralisation experiments were performed in triplicate and for analysis we used the mean for each sample.

2.3.4 Mutants

Single point mutants were generated using the wild-type UKN3A13.6 plasmid (accession AY894683.1) as template. Mutagenesis was performed using a Q5 Mutagenesis Kit (NEB) following the manufacturer's protocol (primers available on request). Mutants T501N and E533K were generated, and infectivity assessed using the pseudotype model described above.

2.3.5 Sequencing

Viral RNA was isolated from 500 µL plasma using the NucliSENS magnetic extraction system (bioMerieux). Libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England BioLabs) with a maximum of 10 ng total RNA template. 500 ng of pooled library was enriched using the xGen Lockdown protocol from Integrated DNA Technologies (IDT) (Rapid Protocol for DNA Probe Hybridization and Target Capture Using an Illumina TruSeq or Ion Torrent Library (v1.0)) with equimolar-pooled 120-nt DNA oligonucleotide probes (IDT) followed by a 12-cycle, modified, on-bead, post-enrichment PCR re-amplification step. The cleaned post-enrichment library was normalized with the aid of qPCR and sequenced with 151-base paired-end reads on a single run of the Illumina MiSeq using v2 chemistry. De-multiplexed sequence-read pairs were trimmed of low-quality bases using QUASR

(v7.0120) and of adaptor sequences using CutAdapt (v1.7.1). The remaining read pool was screened against a BLASTn database containing 165 HCV genomes, which covered its diversity both to choose an appropriate reference and to select those reads that formed a population for *de novo* assembly with Vicuna (v1.3). The assembly was finished with V-FAT v1.0.

2.3.6 Quantification and statistical analysis

Phylogenetics

Whole-genome viral nucleotide consensus sequences for each patient in the study (n = 57) and gt3a HCV reference sequences (n = 9) were aligned using MAFFT [177] with default settings. The alignment file containing only the E1 and E2 regions was uploaded to the NGPhylogeny server (https://ngphylogeny.fr/) and a maximum-likelihood tree was generated with default settings. R Statistical Software (Version 4.1.1) and "ape" and "phytools" packages [178, 179] were used to read the newick format tree and root the tree at midpoint and plot it. Linear regression in R (lm function) was used to test the impact of number of amino acid differences between the UKN3A13.6 reference and the viral consensus sequences for each patient.

Correlation between baseline binding and neutralisation

To test for the relationship between baseline binding and neutralisation, univariable linear regression model was performed to calculate the *p* value and R squared. Binding was the explanatory variable and neutralisation was the response variable

in the model. The 95% confidence interval for the best-fit line was calculated by "predict" function from "stats" package in R.

Association between non-viral factors and antibody response

A multivariable linear regression model was used to test for the association between clinical phenotypes and antibody response. Antibody response was used as the outcome variable and the patient's sex, cirrhosis status, prior IFN-based treatment status, age and log of baseline viral load were included as covariates in the model. The antibody response data were standardized using "scale" function from "base" package in R. The *p* values for the associations were from linear regression model. The "metafor" package in R was used to plot the *p* values and 95% confidence intervals of the association test as a forest plot. To test the association between three *IFNL4* gene haplotypes and antibody response, the univariable linear regression was performed. The antibody response was the explanatory variable and the *IFNL4* gene haplotypes were the response variable in the model. To test for HLA alleles association with antibody response, antibody response was used as the explanatory variable and HLA alleles as response variables to perform univariable linear regressions. The HLA allele data were imputed as previously described [146].

Association between viral amino acids polymorphisms in E1/E2 and antibody response

Linear regression was used to test for association between the presence and absence of each amino acid residue at variable sites E1 and E2 and antibody response, including patient's sex, cirrhosis status, prior IFN-based treatment status,

log of baseline viral load and *IFNL4* gene haplotypes as covariates. The threshold number for the tested residues was set to greater than ten isolates. The multiple tests were corrected at a false-discovery rate (FDR) of 5% and 20%, respectively. The associated viral amino acids sites calculated from the tests were performed linear regression for association with antibody response individually, including the same confounding factors above. All the *p* values were from the linear regression models.

Association between the N-linked glycosylation motifs and antibody response

A linear regression model was used to test for association between the presence or absence of each N-linked glycosylation motif at glycosylated sites at E1 and E2 proteins. The multivariate model included sex, cirrhosis status, prior IFN-based treatment status, log of baseline viral load and *IFNL4* gene haplotypes as covariates to control for possible confounding. The sites were only tested where there was some variability in the motifs with a threshold count of 10 or more. False-discovery rate (FDR) of 20% was used to correct for multiple testing where this study combined the p values generated from binding and neutralisation association tests in one FDR procedure.

Correlation between within patient viral nucleotide diversity and antibody response

Univariable linear regression model was used to test for association between within patient viral diversity at the nucleotide level and antibody response. The mean nucleotide diversity was calculated from next-generation sequencing reads

presenting as the measure of within patient viral nucleotide diversity. The mean nucleotide diversity across the whole genome, E1 combined with E2, E1, E2 and hypervariable region 1 were tested for the correlation with antibody response. All the *p* values were from the linear regression models.

2.4 Results

2.4.1 Cohort description and measurement of HCV specific antibodies

Samples from 60 individuals from the BOSON clinical trial (registration number: NCT01962441) were studied [173]. Utilising baseline plasma samples, the polyclonal antibody binding response directed to the full length E1 and E2 envelope glycoproteins was assessed by ELISA, using an E1/E2 expression construct produced in human HEK293T cells. As target antigen a well characterized clone (UKN3A13.6) was utilized [174]. This expression clone was previously identified as functional in a retrovirus pseudotype infection assay, confirming expression of correctly folded protein able to utilize cellular receptors for specific entry. The sequence of this antigen clustered among the samples used in this work in a maximum likelihood phylogenetic tree (Figure 2.1) based on E1 and E2-encoding nucleotide sequences. Additionally, the neutralisation sensitivity of the antibody responses in each individual was evaluated using retroviruses pseudotyped with an identical full length E1/E2 construct representing UKN3A13.6. The ELISA and neutralisation assays were reproducible with mean coefficient of variations of 8.5% and 14.4%, respectively.

Entire consensus genome sequences for HCV were generated for 57 of the infections. Of these, 54 individuals were infected with HCV genotype 3a (gt3a), and of the remaining three, two were infected with gt3b and one with gt2b. To minimize the impact of virus genetic heterogeneity, this chapter focused on samples from the 54 patients infected with gt3a viruses.

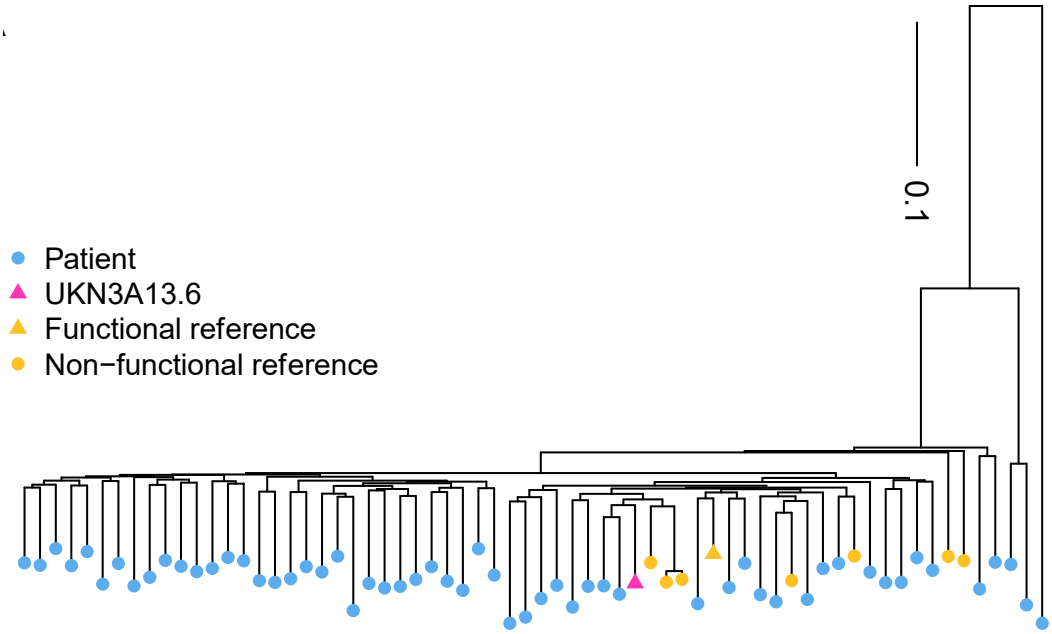


Figure 2.1: The selection of reference isolate

Maximum likelihood phylogenetic tree showing the relationship between the study isolates (blue) and reference isolates (yellow). Functional and non-functional reference isolates are denoted by triangles and circles respectively. The chosen reference isolate UKN3A13.6 is highlighted with a red triangle.

2.4.2 Impact of reference antigen (UKN3A13.6) diversity on antibody responses

After performing antibody binding and neutralisation assays, a linear regression was used to determine whether binding and neutralisation were associated with number of amino acid differences between the reference UKN3A13.6 antigen and the HCV consensus sequences of the study samples. No significant associations between binding and number of amino acid differences (*P* = 0.36, Figure 2.2A) or

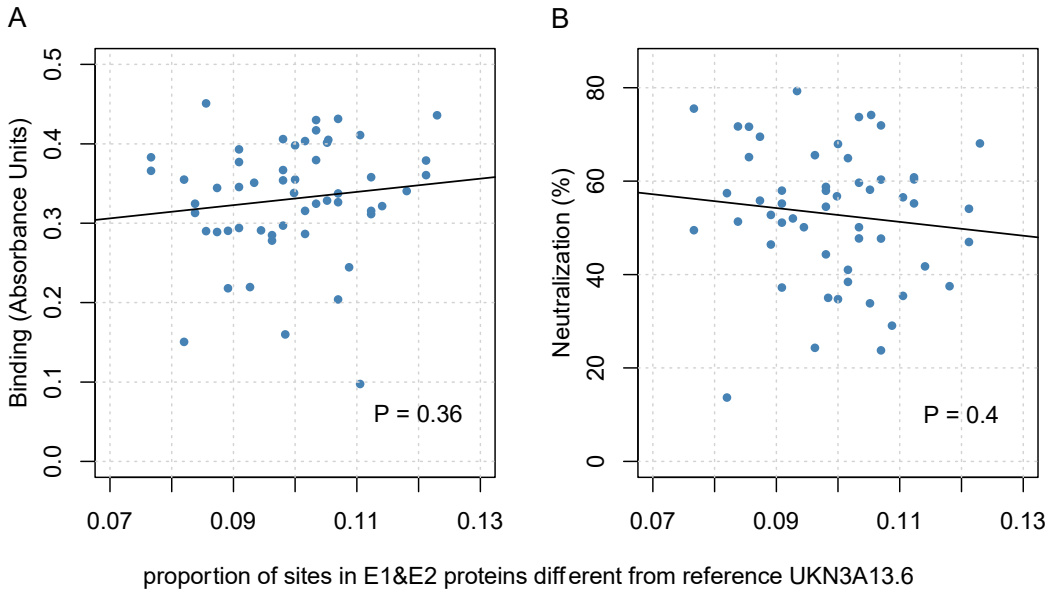


Figure 2.2: Correlation between antibody response and E1/E2 amino acid divergence from the reference antigen (UKN3A13.6)

neutralisation and number of amino acid differences (*P* = 0.4, Figure 2.2B) were found.

Linear regression p-values for the association are shown. The x-axis indicates the proportion of amino acid differences between study isolates and the reference UKN3A13.6 antigen. Each blue dot indicates a study sample and the solid black lines show the best fit linear regression lines and grey area indicate its 95% confidence interval. P values are estimated using a linear regression model. (A) is for binding and (B) is for neutralization.

2.4.3 Correlation between binding and neutralization

The relationship between binding and neutralization was also examined. A strong positive association was observed in a univariate linear regression model (P = 5.3 x 10-6, Figure 2.3), with binding explaining approximately 33% of the variation in neutralization.

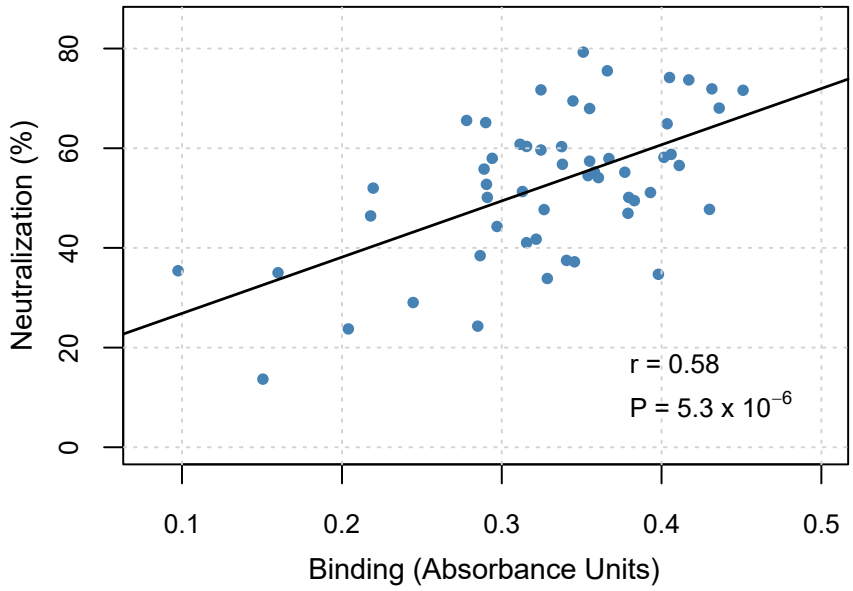


Figure 2.3: The correlation between binding responses and neutralisation sensitivity

The solid black line represents the best fit linear regression line where the p-value for the slope being non-zero is ****p = 5.3 x 10^{-6}.

2.4.4 Non-genetic factors associated with antibody responses

Binding

A multivariable linear regression analysis was used to investigate the association between antibody response and several factors, including sex, cirrhosis status, prior interferon treatment, age and $\log_{10}$ viral load into the model. Binding and

neutralisation values were standardized to have a mean of zero and standard deviation of one (Figure 2.4). Patient's sex had the largest impact on the antibody binding response where males had lower anti-E1/E2 binding levels than females (P = 0.011). Cirrhosis and previous interferon treatment status were also associated with binding, with patients without cirrhosis or with previous interferon treatment displaying lower antibody binding responses ($P_{cirrhosis}$ = 0.029, $P_{prior\text{-}treatment}$ = 0.04). Higher viral load was associated with higher binding, but this effect was not statistically significant ($P_{viral\text{-}load}$ = 0.11).

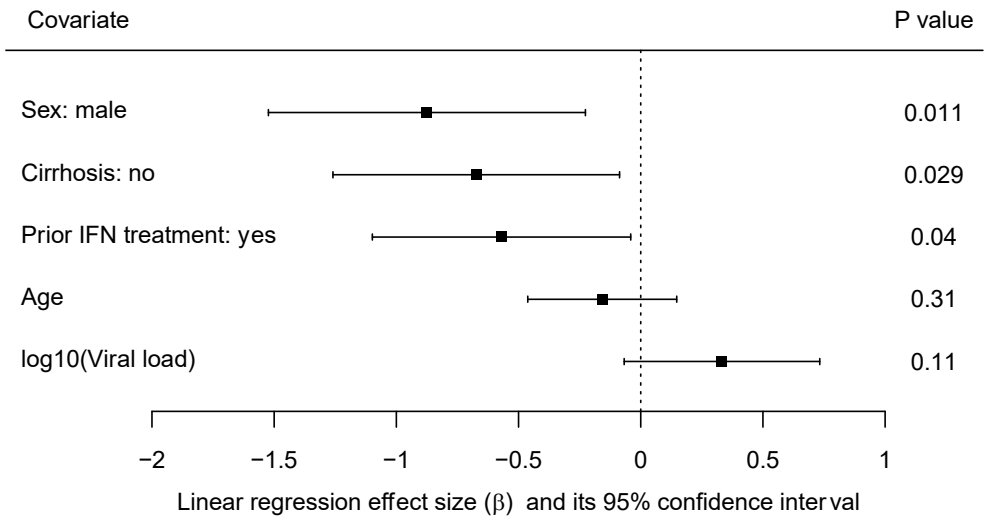


Figure 2.4: The impact of non-genetic factors on antibody binding response
Forest plot of the effect sizes and their confidence intervals for non-genetic factors tested against binding. The squares show the linear regression estimated effect sizes (in units of standard deviation) for each covariate and the lines show its 95% confidence interval. The p values for each covariate are shown on the right (n = 54).

Neutralization

The same factors mentioned above were also tested for the association with neutralisation using linear regression (Figure 2.5). Prior interferon treatment was marginally associated with lower levels of neutralisation sensitivity (P = 0.044). Male sex and absence of cirrhosis were associated with lower neutralisation sensitivity, but the effects were not statistically significant (P_{male} = 0.089, $P_{prior\text{-}reatment}$ = 0.1).

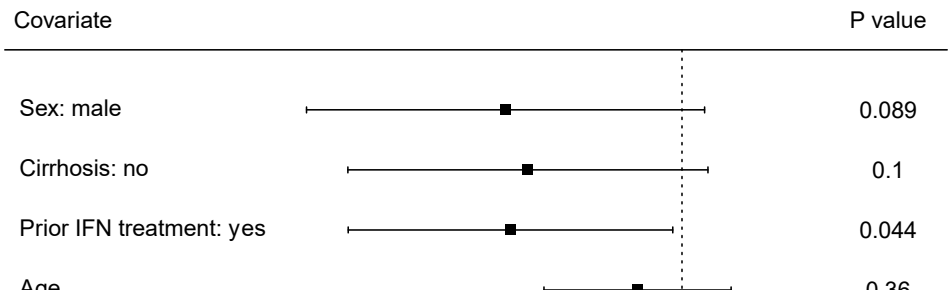


Figure 2.5: The impact of non-genetic factors on neutralization.

Forest plot of the effect sizes and their confidence intervals for non-viral factors tested against neutralization. The squares show the linear regression estimated effect sizes for each covariate and the lines show its 95% confidence interval. The p values (from linear regression) for each covariate is shown on the right (n = 54).

2.4.5 Host genetic factors associated with antibody responses

IFNL4 gene haplotypes and binding response

Since *IFNL4* gene haplotypes are highly associated with HCV spontaneous clearance and IFN-based treatment response, the association between the three haplotypes linked to the three protein variants (IFNλ4-Null, IFNλ4-S70 and IFNλ4-

P70) and antibody binding were examined (Figure 2.6). Individuals with one or two copies of IFNλ4-P70 were classified as IFNλ4-P70, while individuals with two copies of IFNλ4-S70 or one copy of IFNλ4-S70 and one copy of IFNλ4-Null were classified as IFNλ4-S70. Patients classified as IFNλ4-P70 were observed to have significantly lower antibody binding than patients with IFNλ4-Null (*P* = 0.012) in

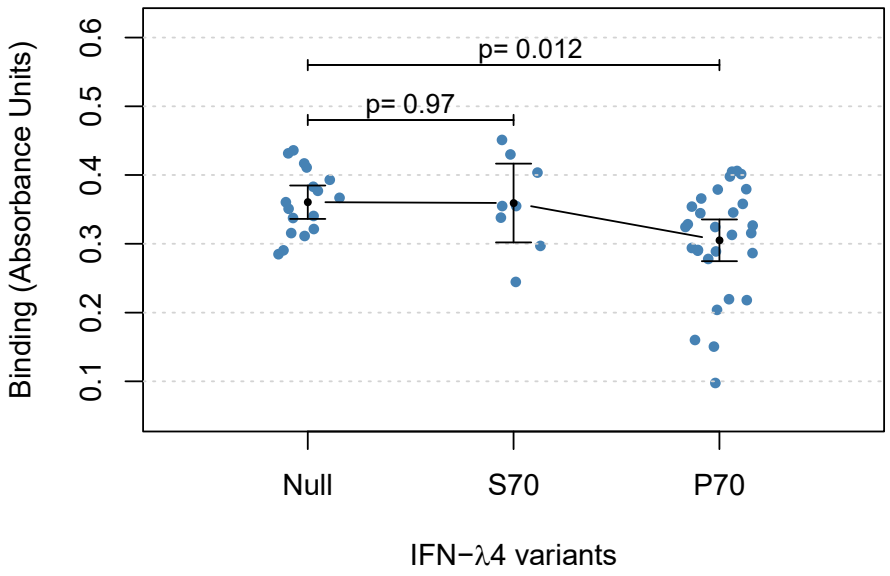


Figure 2.6: The impact of *IFNL4* gene haplotypes on antibody binding response.

regression analysis. However, antibody binding levels were similar in patients with IFNλ4-S70 and IFNλ4-Null variants (*P* = 0.97).

Binding stratified by the host *IFNL4* gene haplotypes linked to the three protein variants (IFNλ4-Null, IFNλ4-S70 and IFNλ4-P70). The black dots and lines indicate the mean and its 95% confidence interval for each group. The p values for the difference in mean relative to the IFNλ4-Null were calculated using a linear regression model (p_{S70} = 0.97, *p_{P70} = 0.012).

IFNL4 gene haplotypes and neutralization response

Similar analysis was performed in the association between *IFNL4* gene haplotypes and antibody neutralisation sensitivity. A modest reduction was observed in

antibody neutralisation sensitivity in IFNλ4-P70 individuals compared to the other two groups, although, the result was not statistically significant (Figure 2.7).

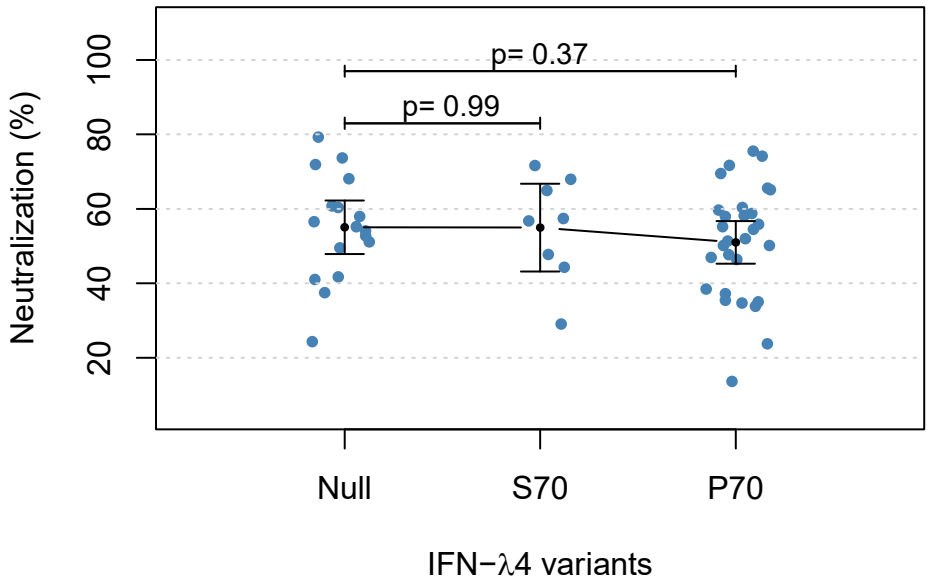


Figure 2.7: The impact of *IFNL4* gene haplotypes on antibody neutralization response.

Neutralization stratified by the host IFNλ4 protein haplotypes. The black dots and lines indicate the mean and 95% confidence interval (CI) for each group. P values were calculated using linear regression model.

HLA molecules and antibody responses

The association between 33 two-field HLA alleles and antibody response in HCV infection was conducted using univariate linear regression. To ensure statistical robustness, the tested HLA alleles were observed from at least 10 individuals. In terms of binding, no significant association was observed at a 20% false discovery rate (FDR). HLA-A*03:01 allele was nominally associated with reduction in antibody neutralisation sensitivity (P = 0.023), but after multiple testing correction the effect was not statistically significant (Table 2.1 and 2.2).

HLA allele	Allele count	P value	Effect size	Standard error	q value
DPA1*02:01	16	0.068	0.493	0.265	0.409
DQB1*03:01	15	0.176	-0.344	0.251	0.409
B*07:02	12	0.195	-0.344	0.262	0.409
C*06:02	10	0.205	-0.389	0.303	0.409
DRB3*01:01	20	0.216	-0.351	0.280	0.409

Table 2.1: Top five associations between HLA alleles and binding.
The p values, effect size and standard error are from linear regression tests and are shown in the table. Q value was estimated using qvalue function in R.

HLA allele	Allele count	P value	Effect size	Standard error	q value
A*03:01	15	0.022	-0.571	0.243	0.742
DRB3*02:02	26	0.128	-0.285	0.185	0.837
DRB3*03:01	12	0.138	0.392	0.261	0.837
A*01:01	14	0.144	0.375	0.253	0.837
C*07:01	15	0.146	0.369	0.250	0.837

Table 2.2: Top five associations between HLA alleles and neutralization.
The p values, effect size and standard error are from linear regression tests and are shown in the table. Q value was estimated using qvalue function in R.

2.4.6 Virus genetic factors associated with antibody responses

Binding response associated viral polymorphisms

To assess the binding associated viral determinants, associations between amino acid variants in E1 and E2 proteins and antibody response were examined using linear regression analysis. Sex, cirrhosis status, prior IFN-based treatment, $\log_{10}$ of baseline viral load and *IFNL4* haplotypes were included as covariates to

minimize possible confounding effects. Binding response data was treated as response variable and the presence or absence of amino acid residues in polymorphic sites of E1 and E2 glycoproteins were used as explanatory variables. To ensure statistical robustness, residues that were present in at least ten isolates were tested in the analysis, which resulted in testing 123 residues (at 77 sites) in each of the assays. Table 2.3 showed a list of the top five associations. At a false-discovery rate (FDR) of 5%, one viral polymorphism – site 653 in E2 (numbering relative to H77 polyprotein) – that was identified significantly associated with antibody binding (Figure 2.8). Increasing the FDR to 20% did not lead to any new findings. Specifically, residues at site 653 (most common sequence surrounding the site 653 RGERCDIEDRD) were significantly associated with antibody binding, where asparagine (N) at this site was associated with reduced binding relative to aspartic acid (D, Figure 2.9A, $P = 6.8 \times 10^{-5}$). However, this site was not associated with antibody neutralisation ($P = 0.19$, Figure 2.9B).

The tested site	The most associated amino acid	Amino acids in reducing frequency order at the tested site	P value	Effect size	Standard error	q value
653	**D**	**DN**	**6.80E-05**	**1.133**	**0.259**	**0.006**
407	A	APS	0.009	-0.698	0.254	0.372
466	R	KRNESG	0.025	0.571	0.246	0.372
500	S	SALRTKV	0.036	-0.536	0.248	0.372
540	E	EKTQNR	0.038	-0.553	0.259	0.372

Table 2.3: Top five associations between E1/E2 amino acid variations and antibody binding.

The table presents the first five significant results from linear regression models testing the associations between antibody binding and polymorphisms at 77 sites (in total 123 residues were tested at these 77 sites, but only the most associated

residue per site is shown in the table, numbered according to the H77 reference sequence). P values, effect sizes, and standard errors are reported, along with q-values representing FDR-corrected p-values to account for multiple testing. Associations at 5% FDR are highlighted in bold.

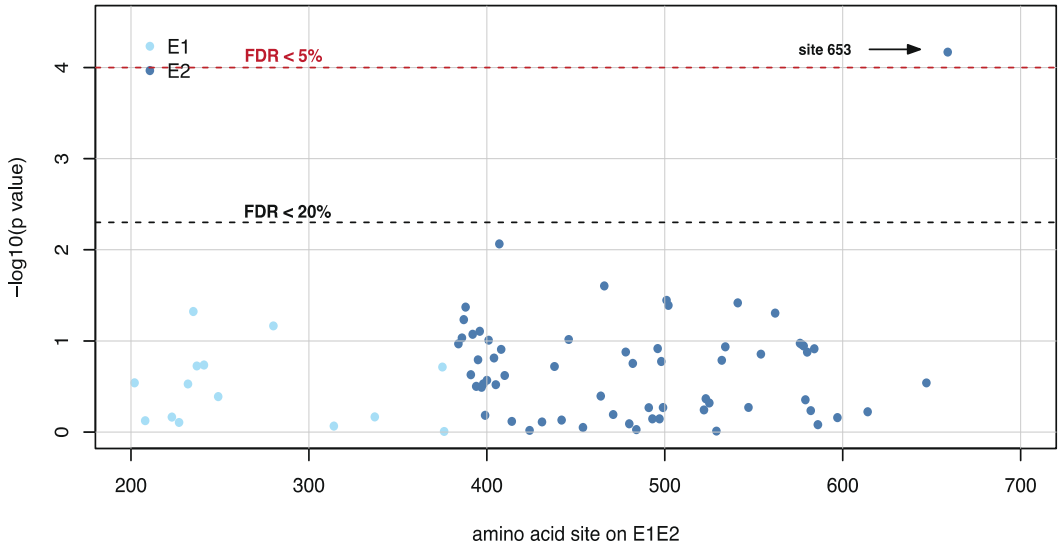


Figure 2.8: Association between HCV amino acid polymorphisms in E1/E2 and antibody binding.

Site 653 (relative to the H77 polyprotein) is significantly associated with antibody binding response at 5% FDR. The red and black dashed lines represent FDR thresholds of 5% and 20%, respectively.

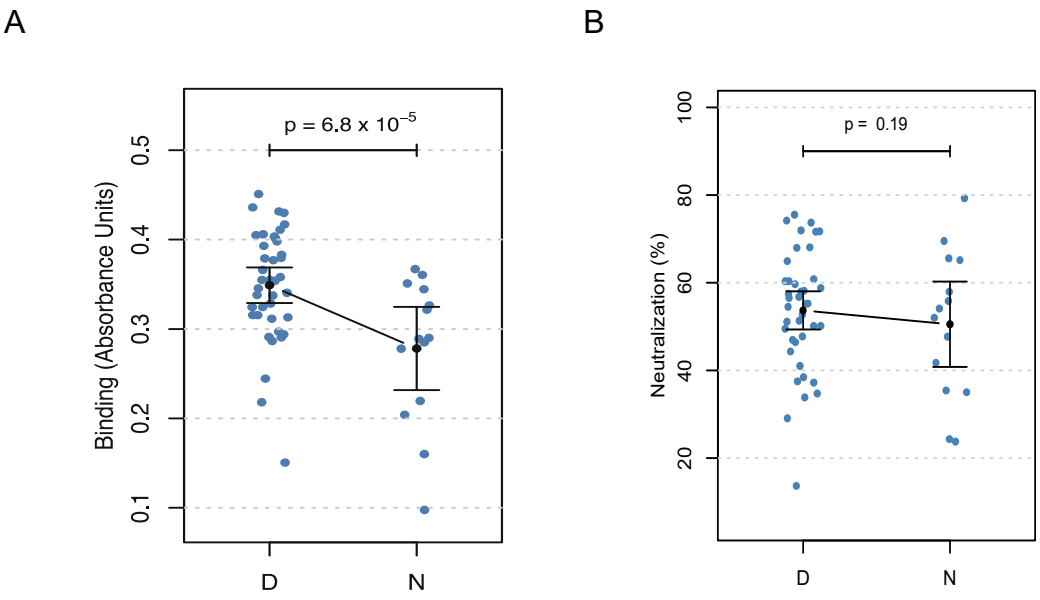


Figure 2.9: Association between amino acids at site 653 and antibody response.

Association for binding (A) and neutralization (B). In patients whose virus carried aspartic acid (D) at this site, the antibody binding response was higher relative to patients whose virus carried asparagine (N) at this site. ****p = 6.8 X 10^{-5}, calculated using linear regression. However, no association was observed between site 653 and neutralization (P = 0.19). The black dots and lines indicate the mean and 95% confidence interval (CI) for each group.

Neutralization response associated viral polymorphisms

The neutralization sensitivity associated viral polymorphisms were calculated in the same way. No sites reached significance at 5% FDR; however, at 20% FDR, two sites were observed to have association with neutralisation sensitivity (sites 501 and 533 in E2) (Table 2.4, Figure 2.10). Amino acid variations at sites 501 and 533 in E2 protein were associated with antibody neutralisation sensitivity, where asparagine (N) at site 501 was associated with higher levels of neutralisation sensitivity relative to other residues (P = 0.001, Figure 2.11A, most common

sequence surrounding the site 501 IVPALNVCGPV). At site 533, possessing a glutamic acid (E) residue was associated with higher neutralisation sensitivity (*P* = 0.0034) compared to other residues at this site (Figure 2.12A, most common sequence surrounding the site 533 TWGENETDVFL). The impact of these residues on binding was consistent with their impact on neutralisation sensitivity, however the effects were non-significant (P_{501} = 0.067, P_{533} = 0.053, Figures 2.11B and 2.12B). These three sites are highly variable in wildtype gt3a isolates. To see the frequency of the amino acids at these three associated sites in the full BOSON dataset (n = 507), the frequency of all the observed amino acids was calculated as shown in the Table 2.5.

The tested site	The most associated amino acid	Amino acids in reducing frequency order at the tested site	P value	Effect size	Standard error	q value
501	**N**	**SNTDGEHQ**	**0.001**	**0.963**	**0.274**	**0.108**
533	**E**	**EKAQVDT**	**0.003**	**0.909**	**0.295**	**0.187**
561	V	VLTI	0.024	0.616	0.264	0.403
235	T	TKAHIMRV	0.025	-0.649	0.279	0.403
471	P	PST	0.025	0.691	0.298	0.403

Table 2.4: Top five associations between E1/E2 amino acid variations and antibody neutralization.

The table presents the first five significant results from linear regression models testing the association between antibody neutralization and polymorphisms at 77 sites (in total 123 residues were tested at these 77 sites, but only the most associated residue per site is shown in the table, numbered according to the H77 reference sequence). P values, effect sizes, and standard errors are reported, along with q-values representing FDR-corrected p-values to account for multiple testing. Associations at 20% FDR are highlighted in bold.

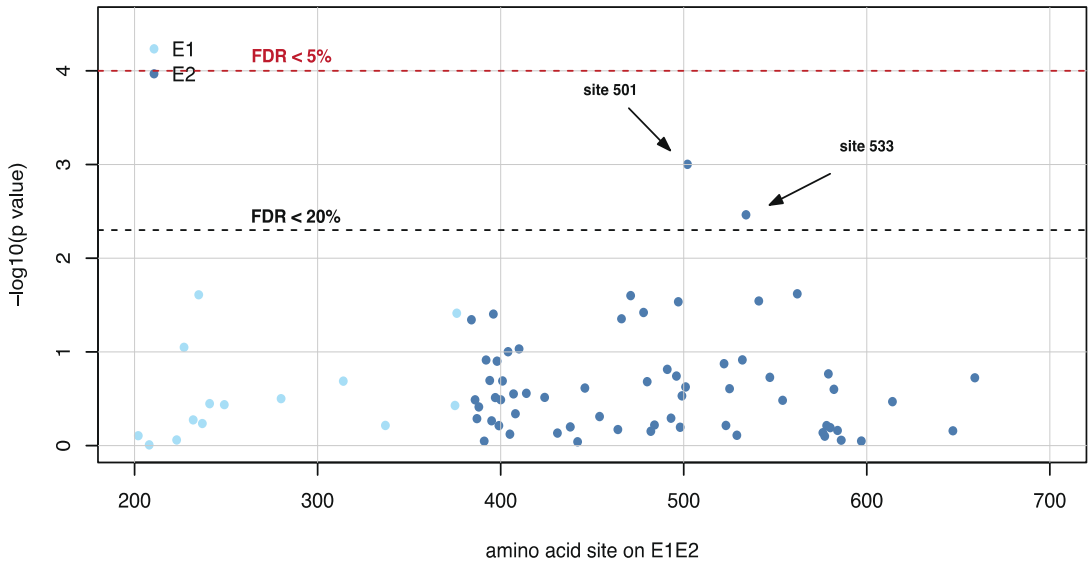


Figure 2.10: Association between HCV amino acid polymorphisms in E1/E2 and antibody neutralization.

Site 501 and 533 (relative to the H77) are significantly associated with antibody neutralization sensitivity at 20% FDR. The red and black dashed lines represent FDR thresholds of 5% and 20%, respectively.

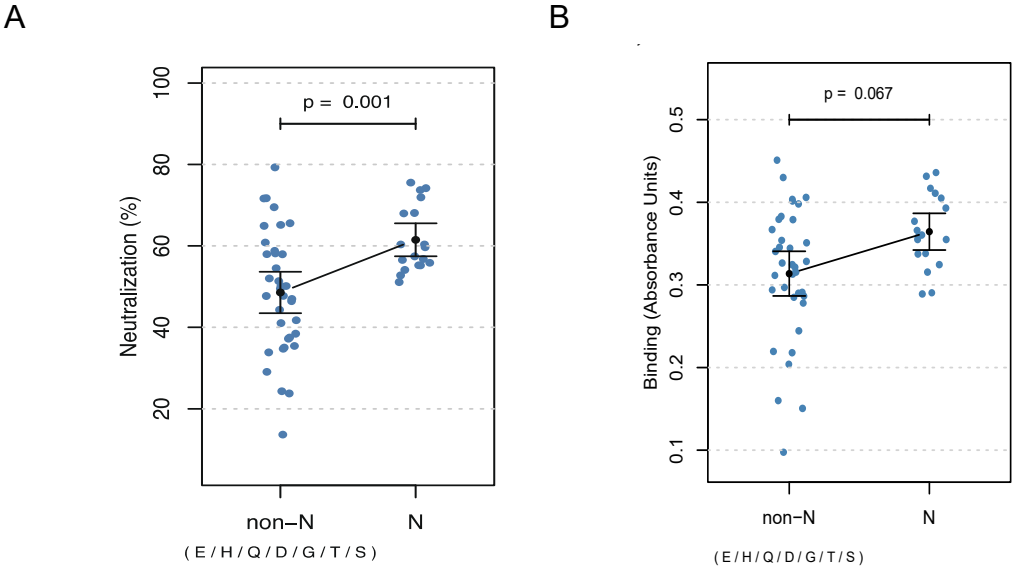


Figure 2.11: Association between amino acids at site 501 and antibody response.

Association for neutralization (A) and binding (B). Asparagine (N) at site 501 was the most associated residue. In patients whose virus carried asparagine (N) at this site, the antibody neutralisation sensitivity was higher relative to patients whose virus carried other amino acids at this site. **p = 0.001, calculated using linear regression. The amino acids listed in the non-N group are ordered in increasing frequency. However, this site was not statistically associated with binding (P = 0.067). The black dots and lines indicate the mean and 95% confidence interval (CI) for each group.

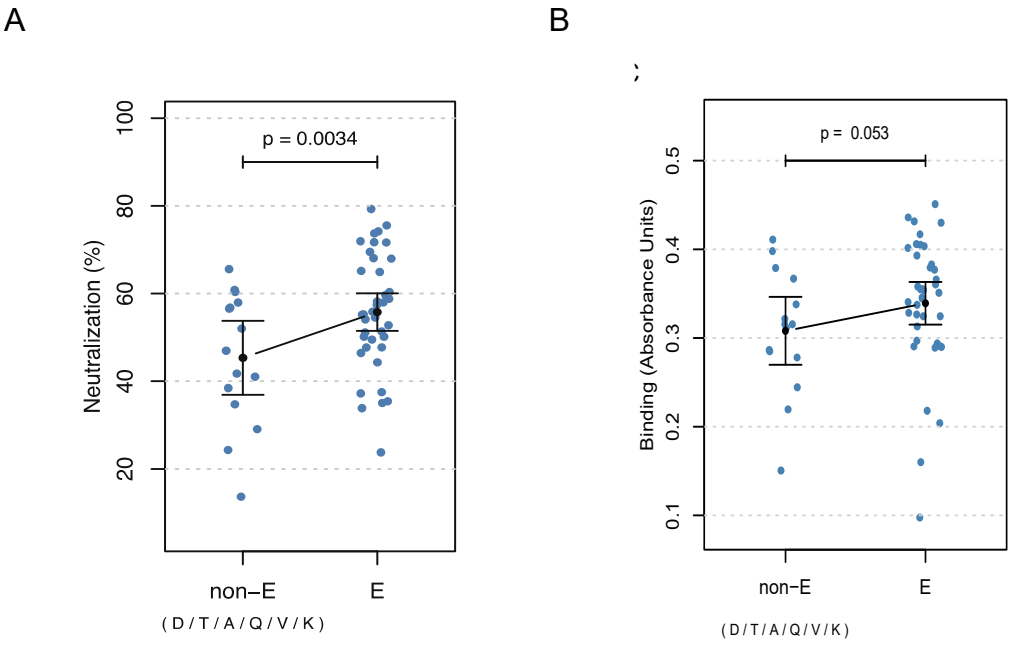


Figure 2.12: Association between amino acids at site 533 and antibody response.

Association for neutralization (A) and binding (B). Glutamic acid (E) at site 533 was the most associated residue. In patients whose virus carried glutamic acid (E) at this site, the antibody neutralisation sensitivity was higher relative to patients whose virus carried other amino acids at this site. **p = 0.0034, calculated using linear regression. The amino acids listed in the non-E group are ordered in increasing frequency. This site was marginally associated with binding (P = 0.053). The black dots and lines indicate the mean and 95% confidence interval (CI) for each group.

Amino acid at site 653	Count	Percent
D	406	80.08%
N	90	17.75%
E	10	1.97%

Amino acid at site 501	Count	Percent
S	181	35.70%
N	159	31.36%
T	59	11.64%
D	30	5.92%
G	19	3.75%
K	19	3.75%
E	17	3.35%
R	16	3.16%
Q	4	0.79%
H	1	0.20%

Amino acid at site 533	Count	Percent
E	402	79.29%
K	52	10.26%
A	14	2.76%
Q	13	2.56%
D	11	2.17%
V	8	1.58%
T	2	0.39%
N	1	0.20%
P	1	0.20%

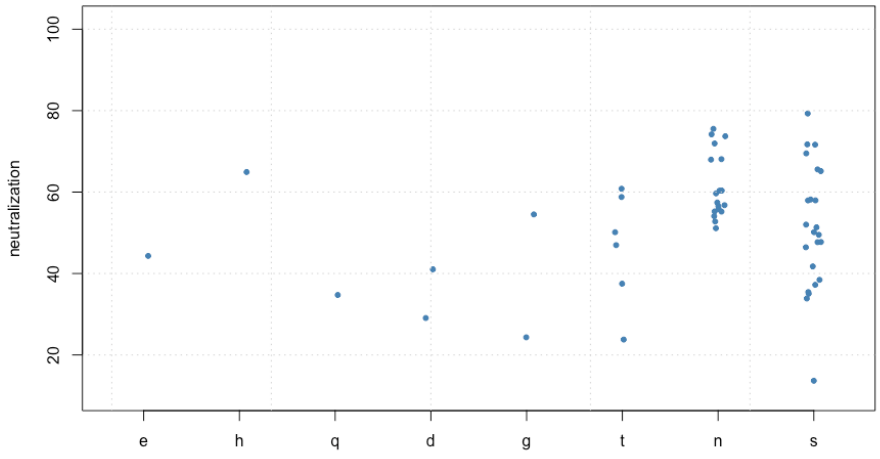
Table 2.5: The frequency of different amino acids at each site (site 653, 501, 533) using 507 gt3a isolates in the BOSON cohort.

The top, medium and bottom tables indicate the frequency of different amino acids at site 653, site 501 and at site 533 respectively in the BOSON dataset. The counts and percent for each amino acid in 507 gt3a samples from BOSON cohort are shown in the tables.

2.4.7 HCV mutants for validating the function of the identified polymorphisms

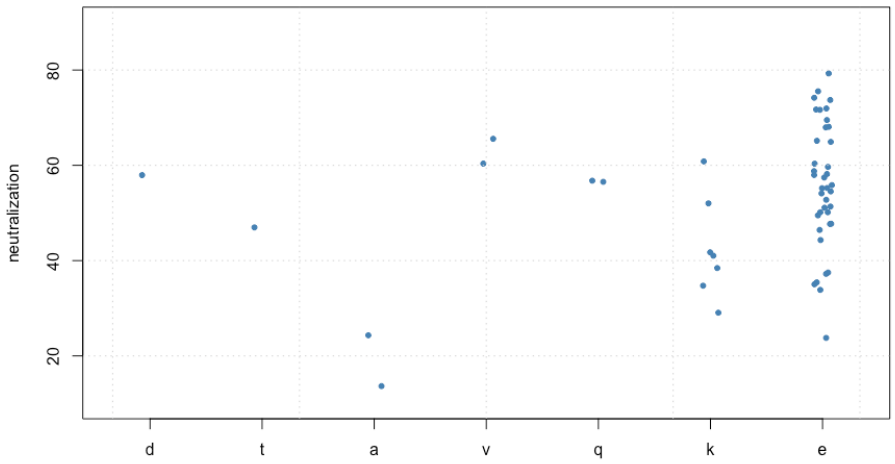
To experimentally assess the phenotypic impact of the polymorphisms associated with neutralisation, site-directed mutants were generated at positions 501 and 533. The reference sequence UKN3A13.6 carried threonine (T) at site 501 – a residue linked to lower levels of neutralization – so a mutant was designed to carry asparagine (N), the residue associated with higher neutralization sensitivity at this position (T501N; Figure 2.13A, Table 2.6). Likewise, as the reference encoded glutamic acid (E) at site 533, a non-E variant was introduced. Based on the most frequent non-E residues at this position, lysine (K) was selected, resulting in the E533K mutant (Figure 2.13B, Table 2.6). The impact of the two polymorphisms (T501N and E533K) on viral fitness were evaluated using an HCV pseudotype model. However, both mutants exhibited extremely low levels of infectivity, making it impossible to assess their relevance to antibody-mediated neutralisation. This highlights the importance of these sites for viral entry (Figure 2.14).

A



N	18
S	23
T	6
G	2
D	2
Q	1
H	1
E	1

B



E	39
K	7
Q	2
V	2
A	2
T	1
D	1

Figure 2.13: Neutralization sensitivity distribution across amino acid variants at site 501 and site 533.

Distribution at site 501 (A) and 533 (B). The left panels represent the distribution of neutralization, and the right panels show the frequency of each amino acid variant at positions 501 and 533, respectively. Rows labelled in blue indicate the residues associated with higher levels of neutralization.

Impact on:	Neut	Neut
Amino acid site:	501	533
Trends:	**N**: higher neut	**E**: higher Neut
UKN3A.1.3.6	T	E
Theoretical Mutants:	T - N	E - nonE
Consider AA frequency:	T501N	E533K

Table 2.6: The mutant design strategy for E2 sites 501 and 533.
This table summarizes the rationale for designing mutant at sites 501 and 533 in the HCV E2 region, based on their association with neutralization sensitivity and AA frequency.

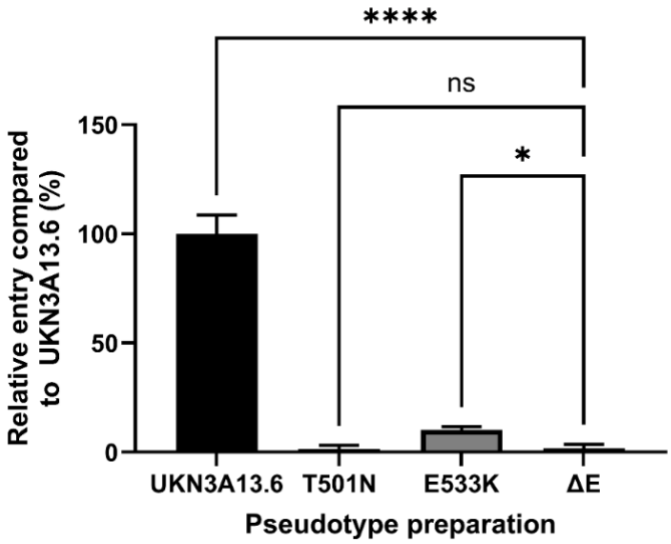


Figure 2.14: Infectivity of HCV mutants using E1/E2 Pseudo-particles.
Infection was performed using HuH7 cells, infected with pseudotypes bearing the wild-type UKN3A13.6 variant glycoprotein, or single aa mutants T501N or E533K. A preparation possessing pseudotypes created in the absence of E1/E2 (ΔE) was used as a negative control. Statistical comparisons were performed using one-way

ANOVA with Dunnett's correction for multiple comparisons. **** $p < 0.0001$, * $p < 0.05$, n.s. not significant.

2.4.8 Viral glycosylation sequons associated with antibody responses

To understand the relationship between viral glycosylation and antibody response, Linear regression analysis was conducted to test for association between the changes in glycosylation motifs in E1 and E2 and antibody responses. Antibody neutralisation and binding were used as the outcome variables, while the presence or absence of different glycosylation sequon motifs at each glycosylation site was used as the exposure variable. To control the possible confounding factors, sex, cirrhosis status, prior IFN-based treatment, *IFNL4* haplotypes and $\log_{10}$ of baseline viral load were included as covariates in the model.

The number and location of N-linked glycosylation sequons in this study.

Initially, the total number of N-linked glycosylation motifs and their locations in the E1/E2 protein were calculated. Most of the sequences (48/54) carried 15 glycosylated sites in their E1/E2 proteins (Table 2.7), which is characteristic of gt3 viruses. At some of the sites, glycosylation motifs were occasionally shifted by one, two or three positions (Figure 2.15). Of the 15 conserved glycosylated sites, six sites carried at least two different motifs with frequencies greater than 10. The further linear regression was only performed at these six sites.

Total number of potential glycosylation sites	Patients
13	2
14	1
15	48
16	3

Table 2.7: The total count of potential N-linked glycosylation sites detected in each patient.

Most patients (48) exhibit 15 potential glycosylation sites, followed by a smaller number with 16, 13, or 14 sites.

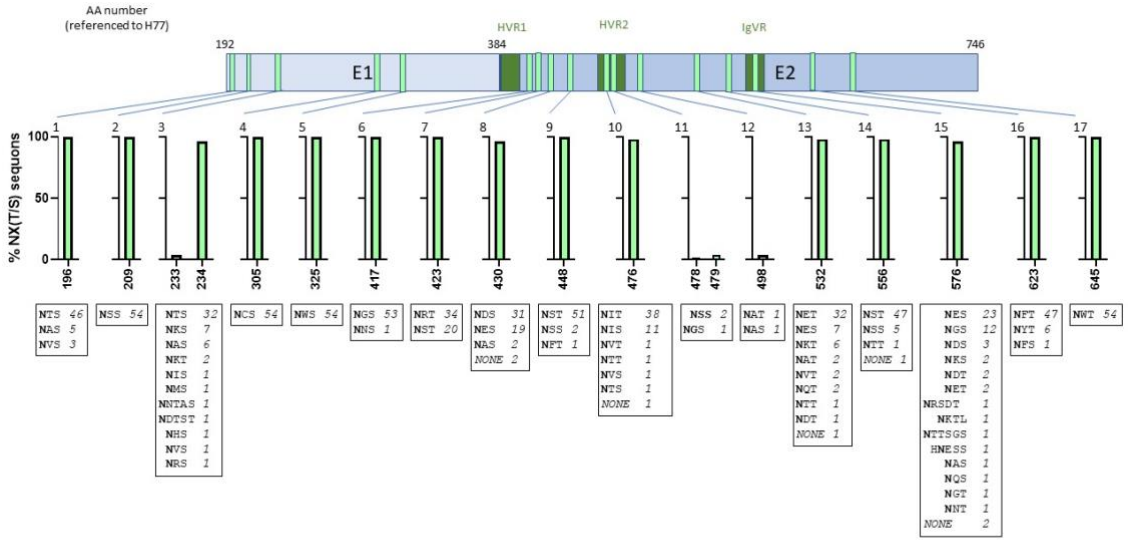


Figure 2.15: Locations and amino acid sequences of glycosylation sequons in the E1/E2 proteins in HCV genotype 3a samples in the study sequences.

Illustrated with a graphical representation of the E1 and E2 genes, the proportion of sequences that possessed N-linked glycosylation sequons (NX(T/S)) at each of the sites is indicated. Amino acid sequences of the glycosylation sequons and their frequency are listed, as represented in the dataset. The locations of the hypervariable region 1 (HVR1), hypervariable region 2 (HVR2) and Intergenotypic variable region (IgVR) are highlighted, as well as the amino acid numbers of the start of E1, the E1/E2 boundary, and the end of E2 (referenced to strain H77, Genbank accession number AAB67036.1)

Antibody response associated glycosylation sites.

As the number of tests were limited, the p values were combined from the binding and neutralisation tests to calculate FDR. No significant associations at a 5% FDR were observed. After increasing the threshold to 20%, motifs at three glycosylation sites were found to be associated with binding and/or neutralisation (Table 2.8). Respectively, motif NTS at site N234 (in E1 Protein, numbering relative to H77 polyprotein) was found associated with reduced neutralisation sensitivity, relative to non-NTS motifs (Figure 2.16A, P = 0.016). The NTS motif at this site was associated with lower binding response, however the effect was not statistically significant (Figure 2.16B, P = 0.095). Additionally, motif NIT at site N476 (in E2 protein, numbering relative to H77 polyprotein) was significantly associated with higher levels of binding and neutralisation response compared to non-NIT motifs (Figure 2.17A $P_{binding}$ = 0.042 and 2.17B $P_{neutralisation}$ = 0.011). Lastly, the NET motif at N532 (in E2 protein) associated with increased neutralisation sensitivity versus non-NET motifs. This N532 sequon notably includes amino acid position 533, where glutamic acid (E), part of the NET motif, independently associated with higher neutralisation sensitivity in our separate amino acid analysis (Figures 2.10 and 2.11).

Binding/ Neutralization	N-linked glycosylation sites	Most associated Motif	p value	Effect size	Standard error	q value
binding	**N476**	**NIT**	**0.042**	**0.564**	**0.270**	**0.170**
binding	N234	NTS	0.095	-0.430	0.252	0.228
binding	N532	NET	0.243	0.307	0.259	0.407
binding	N576c	NGS	0.280	-0.326	0.298	0.407
binding	N423	NRT	0.959	-0.014	0.272	0.991
binding	N430	NES	0.991	0.003	0.268	0.991

neutralization	**N476**	**NIT**	**0.011**	**0.775**	**0.291**	**0.094**
neutralization	**N234**	**NTS**	**0.016**	**-0.677**	**0.269**	**0.094**
neutralization	**N532**	**NET**	**0.058**	**0.545**	**0.280**	**0.173**
neutralization	N576c	NGS	0.220	0.409	0.328	0.407
neutralization	N423	NRT	0.306	-0.308	0.297	0.407
neutralization	N430	NES	0.568	-0.162	0.282	0.682

Table 2.8: Association between tested glycosylation motifs at 6 sites and antibody response.

The table indicates the *p* values, effect size, standard error from the linear regression for each test of the association between the absence or presence of E1/E2 N-linked glycosylation motifs and antibody binding and neutralization response. The location of the tested sites is relative to H77 polyprotein numbering. Associations at 20% FDR are highlighted in bold.

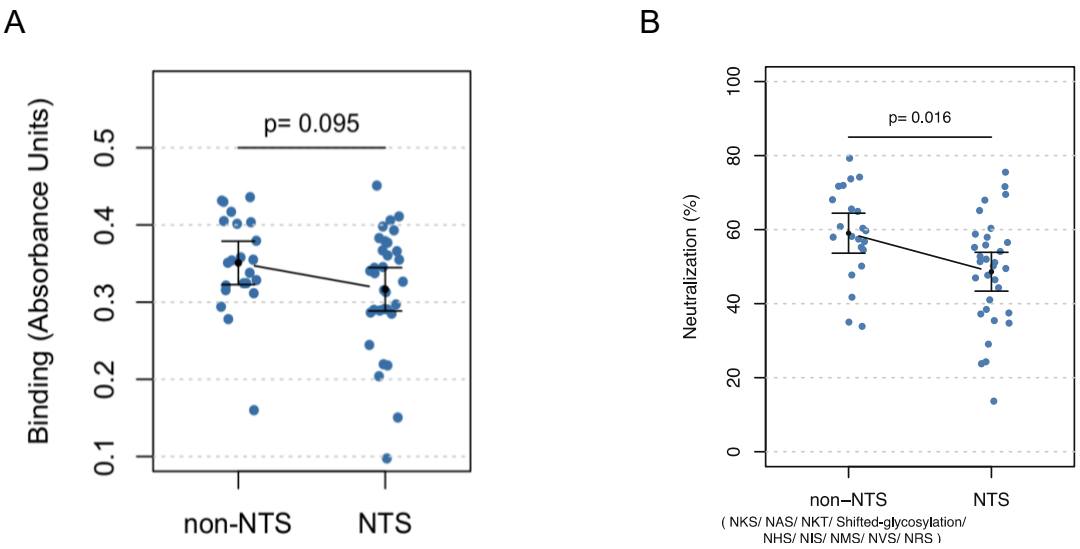


Figure 2.16: Association between variants in glycosylation site N234 and antibody response.

Association with binding (A) and neutralization (B). Site N234 is only associated with antibody neutralization at 20% false discovery rate (FDR). In patients whose virus carried NTS motif at this location, the antibody binding was lower relative to patients whose virus carried non-NTS motifs but not statistically significant (*P* = 0.095 from linear regression), while the neutralisation sensitivity was lower relative

to those with non-NTS motifs (P = 0.016 from linear regression). The motifs listed in the non-NTS group are ordered in decreasing frequency.

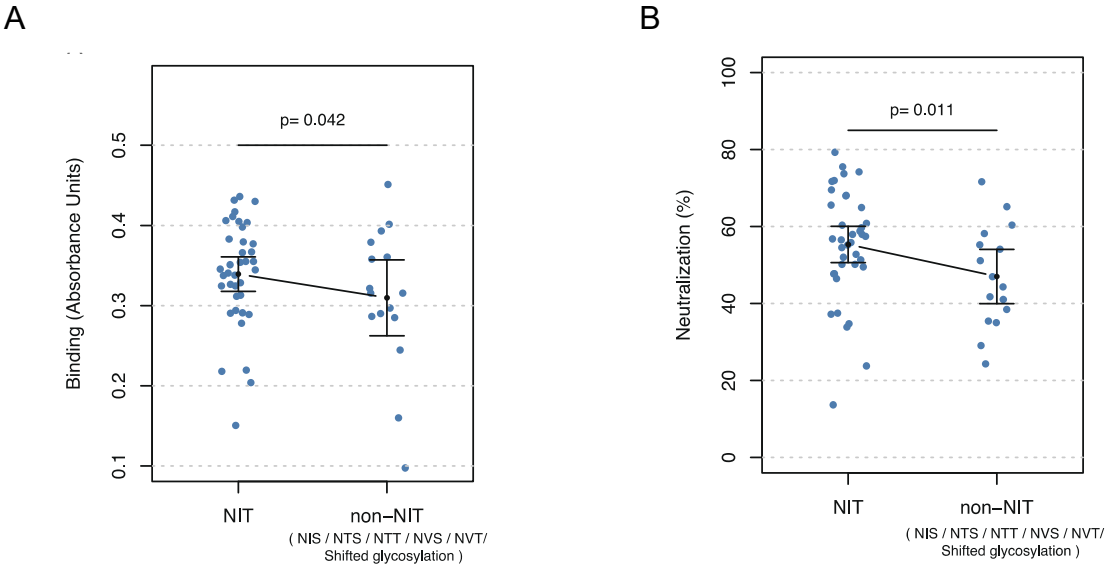


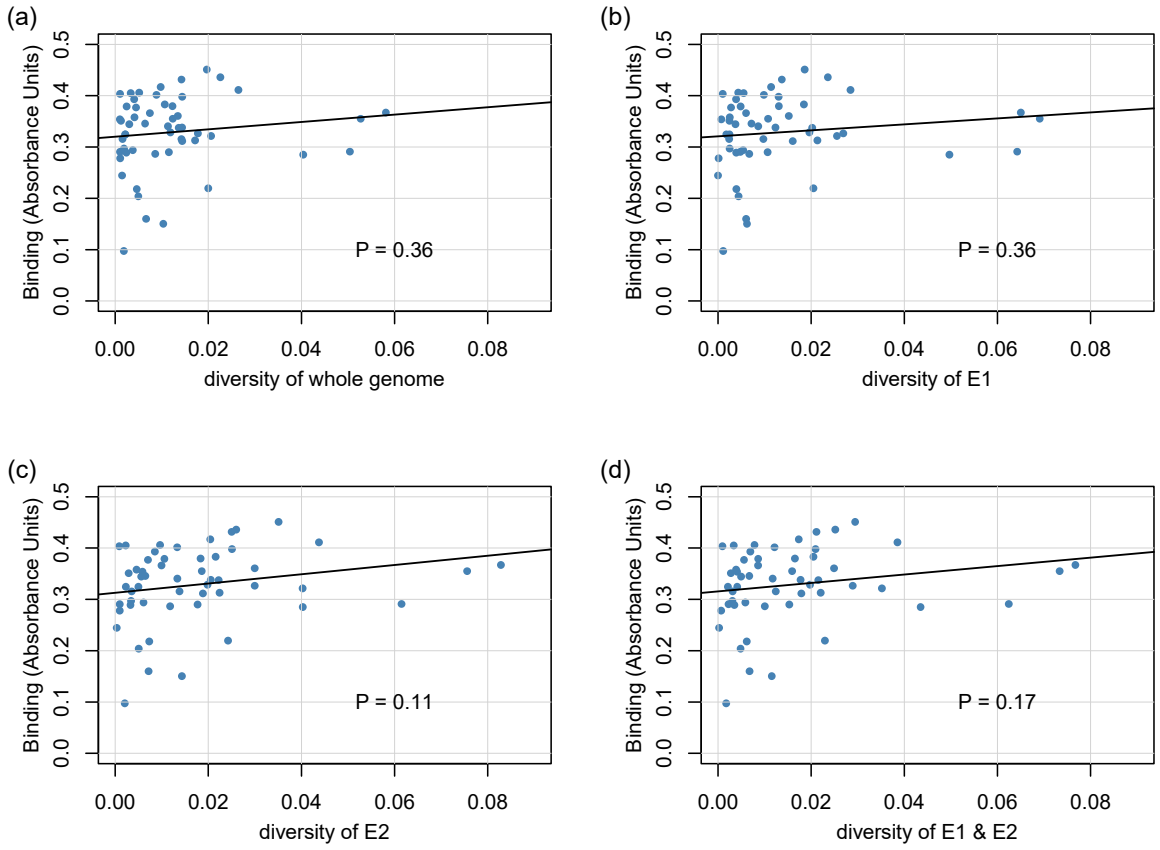
Figure 2.17: Association between variants in glycosylation site N476 and antibody response.

Association with binding (A) and neutralization (B). Site N476 is significantly associated with antibody binding response at 20% false discovery rate (FDR). In patients whose virus carried NIT motif at this location, the antibody responses were higher relative to patients whose virus carried non-NIT motifs ($P_{binding}$ = 0.042 and $P_{neutralization}$ = 0.011 from linear regression). The motifs listed in the non-NIT group are ordered in decreasing frequency.

2.4.9 The association of intra-patient viral nucleotide diversity and antibody responses

The relationship between intra-patient viral diversity at the nucleotide level and antibody responses was investigated as well. Nucleotide diversity, calculated from next-generation sequencing reads, was used to measure intra-patient viral

diversity. Binding and neutralization were regarded as the outcome variable and intra-patient diversity as the exposure variable in linear regression to test for associations between antibody responses and the mean nucleotides diversity across the entire genome, in E1 and E2 genes (Figure 2.18), and the hypervariable region 1 (HVR1). A significant association was only observed between intra-patient viral diversity in the HVR1 region and antibody binding response (Figure 2.18(e); P = 0.0016, r = 0.42), indicating that higher HVR1 diversity within individuals corresponded to increased binding levels. However, there were no significant associations between intra-patient viral diversity and neutralisation sensitivity (Figure 2.19).



(e)

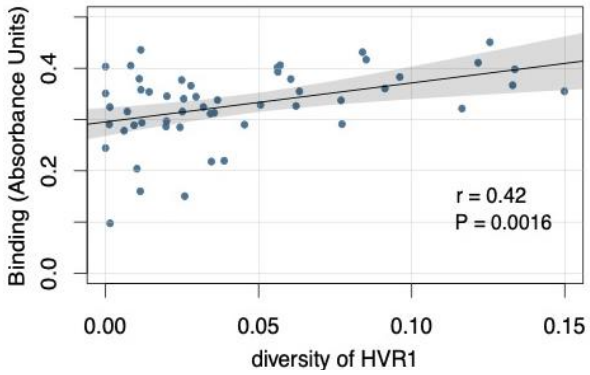


Figure 2.18: Correlates of intra-patient viral nucleotide diversity and binding. a) Correlation between diversity in whole viral sequences, b) E1 protein region, c) E2 protein region, d) E1 and E2 region together, e) HVR1 region and baseline binding. The solid black lines show the best fit linear regression line and grey area indicates its 95% confidence interval. *P* values for the slopes are from the linear regression model.

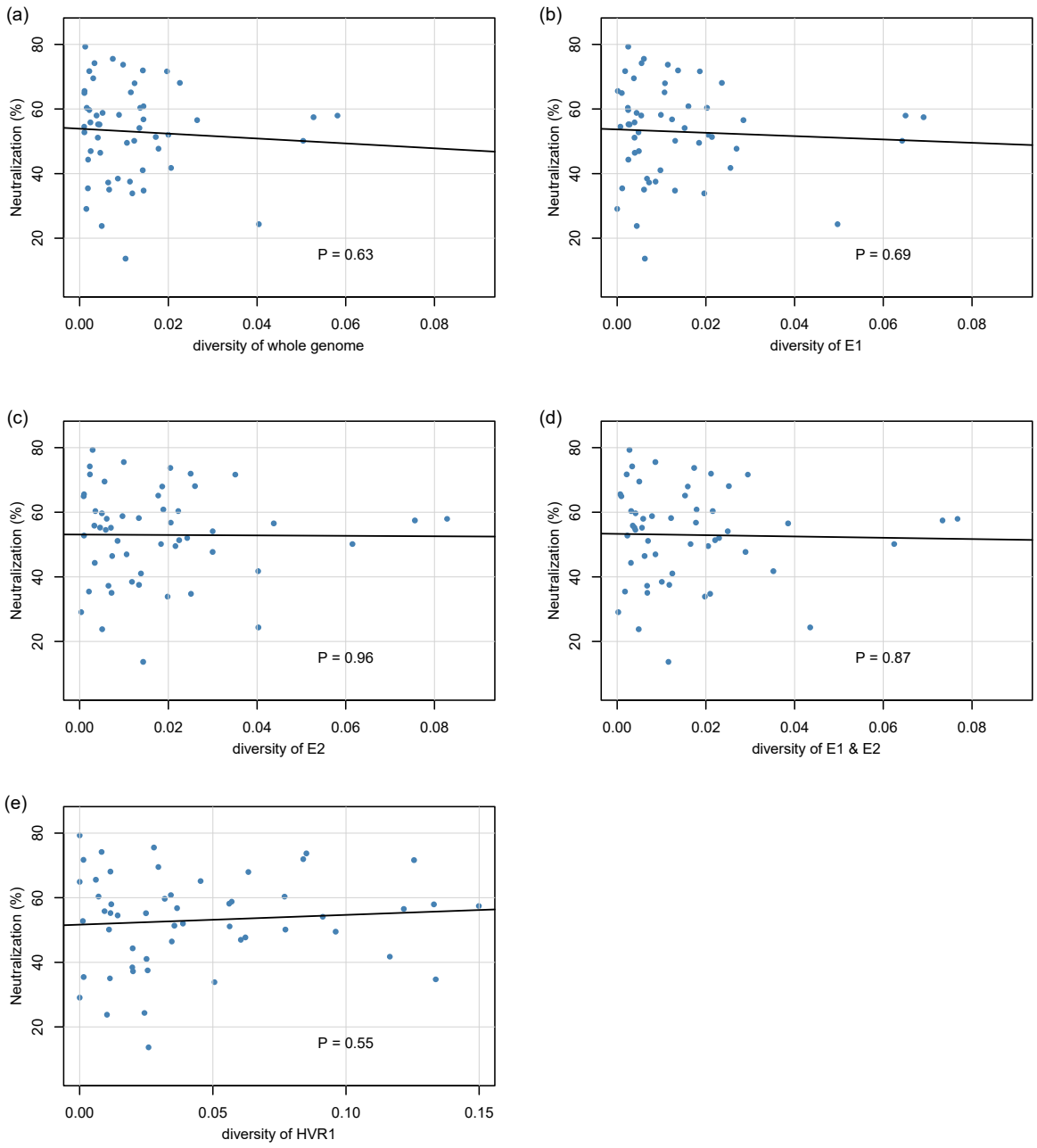


Figure 2.19: Correlates of intra-patient viral nucleotide diversity and neutralization.

a) Correlation between diversity in whole viral sequences, b) E1 protein region, c) E2 protein region, d) E1 and E2 region together, e) HVR1 region and neutralization binding. The solid black lines show the best fit linear regression line and grey area indicates its 95% confidence interval. *P* values for the slopes are from the linear regression model.

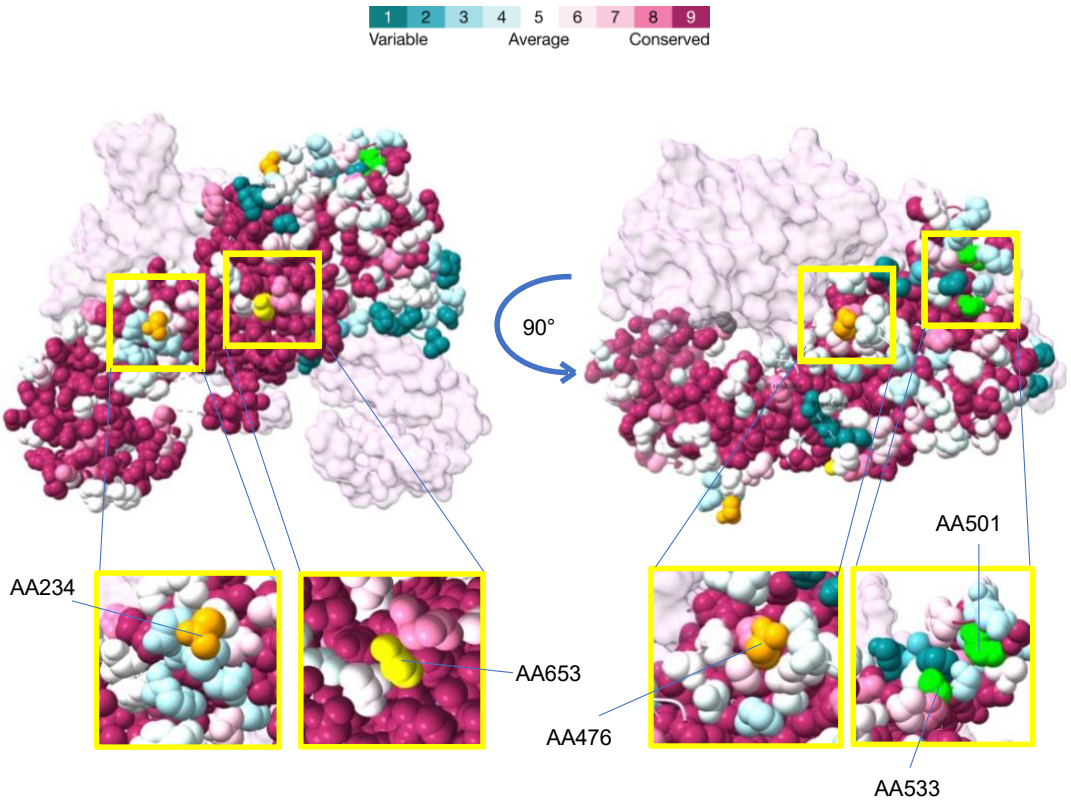
2.4.10 Mapping antibody-associated polymorphisms to surface-exposed E1/E2 regions

To evaluate whether polymorphisms associated with antibody responses localise to structurally accessible and immunologically relevant regions of the HCV E1/E2 glycoproteins, both sequence variability and protein structure using genotype 3a sequences were analysed. A multiple sequence alignment of E1/E2 sequences was generated and analysed using ConSurf [180] to compute residue-specific evolutionary conservation scores. These scores were then mapped onto the surface of the recently resolved genotype 3a E1/E2 heterodimer structure (PDB ID: 8RJJ) [158], allowing visualisation of variability across the three-dimensional conformation of the glycoprotein.

This analysis revealed that the polymorphism at residue 653–associated with differential antibody binding–resides on a highly conserved region of the E2 surface (Figure 2.20A). In contrast, glycosylation sites at positions 234 (in E1), 476 and 533 (in E2), along with amino acid residue 501 (in E2), all of which were associated with antibody neutralisation sensitivity, mapped to regions exhibiting elevated sequence variability (Figure 2.20A). These findings suggest that antibody-binding epitopes under limited immune pressure (e.g., residue 653) tend to be structurally conserved, whereas sites involved in neutralisation are more variable and likely subject to immune-driven selection. To complement these structural insights, sequence variation across the linear E1/E2 protein was visualised, highlighting major peaks in amino acid diversity that correspond to

known immunogenic domains, including hypervariable region 1 (HVR1), hypervariable region 2 (HVR2), and the inter-genotypic variable region (IgVR) (Figure 2.20B).

A



B



Figure 2.20: Structural and sequence context of antibody-associated amino acid polymorphisms in HCV E1/E2.

(A) Structural mapping of antibody-associated polymorphisms onto the E1/E2 heterodimer (PDB: 8RJJ), with ConSurf analysis used to visualise site-specific amino acid variability. One copy of the E1/E2 heterodimer is shown with a space-filling representation of side chains coloured by evolutionary conservation. Residues associated with neutralisation sensitivity (501 and 533) are highlighted in green; the residue associated with antibody binding (653) is shown in yellow. N-linked glycosylated asparagine residues at positions 234 and 476 are shown in orange. The second E1/E2 heterodimer is rendered as a surface to illustrate the quaternary interface. The left panel presents a side view of the E1/E2 dimer- of-heterodimers; the right panel shows the same structure rotated 90° along the horizontal axis. (B) Linear sequence representation of E1 and E2 proteins, highlighting regions of sequence conservation and variability. Hypervariable regions (HVR1, HVR2), the inter-genotypic variable region (IgVR), and antibody-associated polymorphic sites are annotated for reference.

To explore the potential structural impact of the identified polymorphisms, this study generated AlphaFold3-based structural models of E1/E2 incorporating the non-reference residues for each of the five key polymorphic sites (Figure 2.21A). While subtle local rearrangements cannot be ruled out, these variants were not predicted to induce major conformational changes relative to the reference structure. This supports the interpretation that the observed differences in antibody binding and neutralisation are more likely driven by epitope-specific surface alterations or glycan structure or accessibility rather than gross structural remodelling. Finally, the spatial overlap between the identified polymorphisms and known monoclonal antibody (mAb) epitopes (Figure 2.21B) was examined. Mapping of epitope footprints revealed that three of the four polymorphic sites associated with neutralization (501, 533, and 476 in E2) lie within or are proximal

to the binding sites of well-characterized neutralizing mAbs including 9/75, 2/64a, CBH7 and 6/41a. E1 site 234, also associated with neutralization, overlaps epitopes recognized by mAbs J81 and J82. The only polymorphism associated with differential antibody binding, 653 (in E2), overlaps with the non-neutralising antibody ALP1 and is adjacent to others such as the non-neutralizing antibody ALP98 and the neutralising antibody IGH526[181]. These spatial overlaps support the immunological relevance of the identified sites and suggest that natural polymorphisms at these positions may alter antigenicity through disruption or masking of key antibody contact points and support their potential relevance as targets in vaccine design.

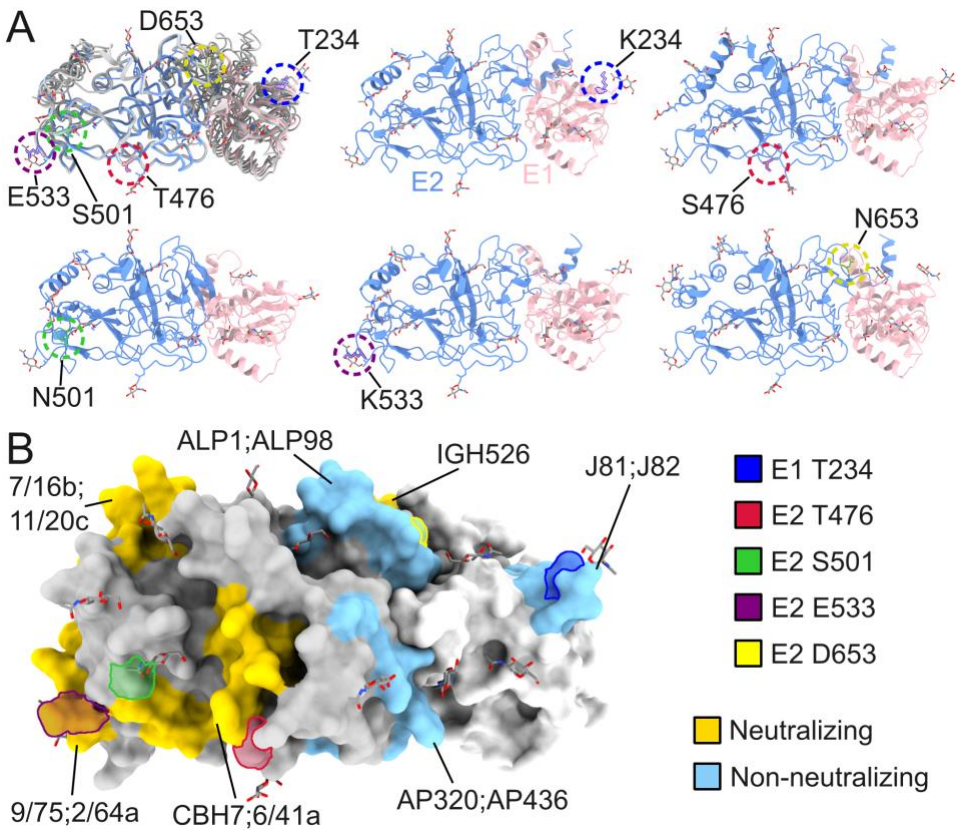


Figure 2.21: Structural localisation and epitope context of antibody-associated polymorphisms in HCV E1/E2.

(A) Structural models of the HCV E1/E2 ectodomain were generated using AlphaFold3 based on consensus genotype 3a sequences. Models were created for the consensus sequence (E1 in pink, E2 in blue) as well as for five individual variants identified in this study: T234K (E1), T476S, S501N, E533K, and D653N (all in E2). Overlaying each mutated structure (shown in grey) onto the consensus model revealed no major conformational changes, suggesting that these polymorphisms are unlikely to induce large-scale structural rearrangements. However, given that Alphafold3 is unlikely to be sensitive to localised structural differences, we cannot preclude the possibility that the observed polymorphisms introduce subtle structural effects.

(B) Surface rendering of the HCV E1/E2 heterodimer in the same orientation as in panel A, with E1 shown in white and E2 in grey. Polymorphic sites are indicated as coloured outlines corresponding to the residue-specific key. Known neutralising antibody epitopes (yellow) and non-neutralising epitopes (blue) are mapped onto the surface. Polymorphisms at sites 501, E533, and 476 (E2) overlap with the footprints of neutralising monoclonal antibodies (e.g., 9/75, 2/64a, CBH7, 6/41a, 7/16b, 11/20c), while polymorphisms at 234 (E1) and 653 (E2) lie within regions targeted by non-neutralising antibodies (e.g., ALP1, ALP98, J81, J82). These spatial overlaps support the immunological relevance of the identified sites. Colour key for polymorphic sites: blue outline: E1 234 (neutralization-associated), red outline: E2 476, green outline: E2 501, purple outline: E2 533, yellow outline: E2 653 (binding-associated), gold shading: Neutralising antibody epitopes, blue shading: non-neutralising antibody epitopes.

2.5 Discussion

In this chapter, linear regression was used to test for association between virus amino acid polymorphisms in E1 and E2 and antibody responses in a well characterized patient cohort, identifying three amino acid sites in the gt3a E2 protein that impact on the antibody responses. Additionally, distinct *N*-glycan motifs present at three specific *N*-glycosylation sites (one overlapping with an identified amino acid site) in E1/E2 were associated with antibody binding and/or neutralisation. Furthermore, the relationships between the intra-patient virus nucleotide diversity and host antibody response were examined, indicating that HVR1 region was associated with host antibody binding against E1/E2. This work also reported host *IFNL4* gene polymorphisms were associated with antibody responses as well as other non-genetic factors such as sex and cirrhosis status of the patient.

Three sites in the E2 protein possessed amino acid polymorphisms associated with differential anti-E1/E2 antibody responses in gt3a virus-infected patients. Among these sites, the strongest association was at site 653 which was significant at 5% FDR. The asparagine (N) and aspartic acid (D) residues at position 653 were found to be associated with antibody binding, which is part of a region previously reported to be targeted by non-neutralizing antibodies [182]. This polymorphism is located in a region of the E2 stem (Figure 2.22) known to be immunogenic in experimental immunizations, with murine monoclonal antibodies ALP11, AP266, ALP1, ALP98, and H52 specifically targeting this region [183]. The reference E1/E2 antigen utilized in the assays possessed an aspartic acid residue at this site, and

higher binding was observed in patients whose infecting virus carried aspartic acid compared to those with asparagine at this site. Structural modelling using AlphaFold3 suggested that the D653N substitution does not induce large-scale conformational changes in the E1/E2 heterodimer (Figure 2.21). However, given its surface accessibility and overlap with non-neutralizing epitope footprints such as ALP1 and ALP98, it is plausible that local side-chain variation or residue-specific recognition underlies the observed association, consistent with a model of variant-specific antibody binding shaped by minor surface alterations rather than global conformational shifts.

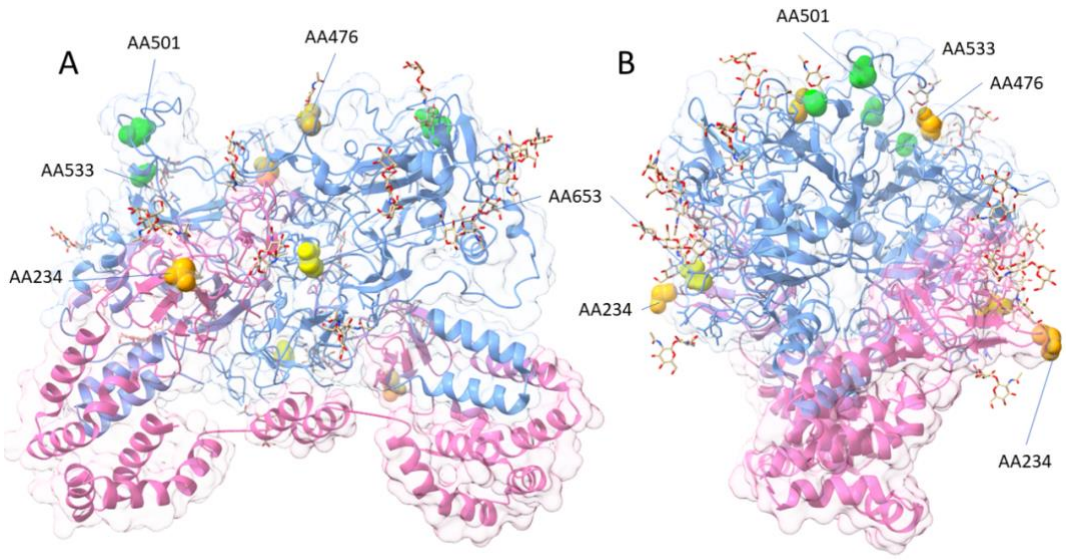


Figure 2.22: Location of sites associated with antibody responses in a structure of the E1/E2 heterocomplex.

In this structure of E1/E2 (PDB 8RJJ), with E1 in pink and E2 in blue. Neutralization-associated residues (AA501 and AA533) are shown in green, the binding-associated residue (AA653) in yellow, and glycosylated asparagine (N234 and N476) in orange. Carbohydrates are depicted as wireframes. Two views of the complex are provided: a side-on view and a 90° rotated perspective.

The amino acid polymorphisms observed to influence the neutralizing antibody response (at 20% FDR) located at positions 501 and 533 are presented at the neutralizing face of the E2 ectodomain [184] overlapping with the CD81 binding site. Site 533 is located in a relatively conserved region (533-535) critical for CD81 receptor interaction [185]. The reference antigen encoded a glutamic acid residue at site 533 and higher levels of neutralisation sensitivity were observed in plasma from patients whose virus possessed a glutamic acid at this site relative to other residues. Site 533 is also part of the N-linked glycosylation motif starting at N532; concordantly, glycosylation analysis in this work showed that the NET motif associated with increased neutralisation sensitivity compared to non-NET motifs, potentially due to distinct glycosylation efficiencies or glycan structures impacting antibody recognition.

Regarding site 501, its proximity to the CD81 binding site in the tertiary structure of E2 (Figure 2.22), suggests that substitutions (e.g., N to other residues) might alter the presentation of conformational neutralisation epitopes without disrupting receptor binding. Indeed, mutants with the Q501A or S501A substitutions at this site have been implicated to influence monoclonal antibody (mAb) binding [186, 187], and the N501S mutation altered mAb binding despite having little effect on virus entry [61].

When individual glycosylation sequons were analysed in the same manner as individual amino acids, it was revealed that three sites (N234, N476 and N532) were associated with the antibody response in this patient cohort. Beyond the NET sequon at site N532 which was associated with higher levels of antibody

neutralisation sensitivity, sequon variations at N234 and N476 were also associated with antibody response. N234 and N476 were previously implicated in resistance to neutralisation, where the removal of these N-linked glycans in HCV gt1a mutants individually reduced sensitivity to polyclonal serum antibodies [188]. It is plausible that the alternative sequons identified here modify the efficiency of glycosylation or the resulting glycan structure at these sites, impacting the sensitivity to neutralisation. Different NX(S/T) sequons are known to be glycosylated with varying efficiencies [189], a phenomenon demonstrated for other viral glycoproteins [190], and as such it is plausible that selection of different sequons during evolution of different virus strains will be associated with different carbohydrate modifications. Further investigations are required to determine if differences in the patterns of glycosylation at sites 234, 476 and 532 influence the efficiency or structure of carbohydrate modifications, as predicted for other virus glycoproteins [191].

The observation that residues and glycosylation motifs associated with antibody neutralisation sensitivity (234, 476, 501, and 533) map to regions of high sequence variability in this cohort, as visualised on the genotype 3a E1/E2 heterodimer structure (Figure 2.20), supports the notion that these sites are subject to immune selection pressure. This aligns with previous findings showing that viral escape from neutralising antibodies during acute HCV infection is often mediated by amino acid substitutions at exposed, immunogenic sites [192]. These findings extend this concept to chronic infection, suggesting that sustained antibody pressure may continue to shape viral sequence diversity at these neutralisation-sensitive loci. In contrast, the polymorphism at residue 653 (associated with binding but not

neutralisation) resides in a more conserved region, implying it may not be under the same selective pressure. This divergence highlights that potentially neutralising, but not necessarily binding, antibody responses are key drivers of antigenic variation in E1/E2 during natural infection.

The analyses in this chapter revealed instances, particularly concerning variations in amino acids and glycosylation motifs, where patient sera exhibited stronger binding or neutralisation against reference antigens that shared the same variant found in the patient's infecting virus. While potentially viewed as an experimental confounder, this work interpreted this concordance as strong evidence supporting a key conclusion: specific variations in E1/E2 elicit variant-specific antibody responses during natural infection, characterized by preferential recognition of homologous epitopes. This study proposed this reflects the host developing tailored humoral immunity specific to the infecting viral variant, underscoring the functional impact of viral diversity on antibody recognition and neutralisation. This finding has direct implications for HCV vaccine development. It highlights the challenge posed by viral diversity and demonstrates that antibodies elicited by one viral variant may be less effective against others even within the same virus subtype. Therefore, designing a broadly effective HCV vaccine likely requires strategies that can induce immune responses capable of overcoming this variant-specific recognition, targeting conserved epitopes or utilizing multi-valent formulations to cover key antigenic variations observed in circulating strains.

The intra-patient viral diversity in HVR1 region was also investigated and showed a significant association with antibody binding response. The causal direction is

difficult to determine as higher antibody response could result in more selective pressure on virus, encouraging it to develop escape mutations. These mutations, in turn, could lead to greater viral genetic diversity. Alternatively, higher viral sequence diversity could lead to a larger antibody response targeting multiple viruses.

In addition to the virus genetic variation, host *IFNL4* activity may shape the antibody response. Previous studies have shown that IFNλ4-P70 and IFNλ4-S70 variants of IFNλ4 protein have distinct phenotypes both *in vivo* and *in vitro* [170, 193]. Assuming a dominant effect (IFNλ4-P70 > IFNλ4-S70 > IFNλ4-Null), individuals with IFNλ4-P70 variant had significantly lower levels of antibody binding response relative to individuals with IFNλ4-S70 and IFNλ4-Null variants in this work. The same trend for neutralisation sensitivity was observed, but the effects were not statistically significant. The IFNλ4-Null variant (CC genotype at SNP rs12979860) was previously reported to associate with higher viral loads, increased spontaneous clearance, improved response to interferon treatment, and lower hepatic interferon-stimulated gene (ISG) expression. The observation of higher anti-E1/E2 antibody levels in individuals with the IFNλ4 -Null variant in this data suggests a potential mechanism: while individuals with the IFNλ4-P70 variant may experience lower viral loads due to higher hepatic ISG expression (which exert antiviral activity), IFNλ4-P70 itself may dampen the adaptive humoral immune response, potentially contributing to lower rates of spontaneous clearance and treatment response. A potential mechanism for the observed link between *IFNL4* polymorphisms and antibody response may involve ER stress-mediated impairment of MHC II presentation, affecting T helper cell function and therefore

alter HCV-specific antibody responses [194]. These findings suggest a role for host *IFNL4* genetic variation in modulating the antibody response to HCV infection.

While this study offers key insights into the interplay between host/viral genetics and antibody responses in patients infected with HCV genotype 3a, several limitations should be considered. First, the use of in vitro assays, including HCVpp with a single E1/E2 strain (UKN3A13.6) and GNA-ELISA binding assays, may not fully represent the diversity or function of in vivo responses. Additionally, our assays do not capture other potentially relevant antibody functions like ADCC. Furthermore, the small sample size (n=54), cross-sectional design, and potential confounders limit statistical power and causal inference. Finally, this study employed an exploratory statistical approach using a 20% false discovery rate (FDR) threshold for identifying associations between genetic variants and antibody responses. This threshold prioritizes sensitivity for hypothesis generation but inherently increases the risk of false positive findings. Therefore, while highlighting promising candidates, these findings require independent validation in larger cohorts before definitive biological conclusions can be drawn. Statistical significance alone does not guarantee biological effect, emphasizing the need for replication and further functional studies.

In conclusion, this work provided a detailed investigation of the impact of genetic variation in HCV and in the host *IFNL4* gene on humoral immune responses. Three novel sites associated with antibody responses were identified in HCV E1 and E2 proteins. Changes in glycosylation motifs in three sites (including one site common to both analyses) were observed to be associated with antibody responses. Given

that antibody responses are a critical component of most vaccine induced immunity, the results suggest that individuals with the IFNλ4-P70 variant may exhibit lower responses to an HCV vaccine, underscoring the need for tailored vaccine strategies. Moreover, naturally occurring amino acid polymorphisms within single HCV subtype (gt3a) impact neutralizing antibody responses emphasizes the importance of considering such variations when designing a broadly protective, pan-genotypic HCV vaccine. Such vaccines must address naturally occurring polymorphisms to enhance cross-reactivity against diverse variants and elicit robust humoral immune responses. Finally, the methodology employed in this study is broadly applicable to other viral infections, offering a framework to design vaccines that more effectively elicit immune responses by accounting for both host and viral genetic variability.

Chapter 3

Viral factors associated with liver disease in patients chronically infected with genotype 3a and 1a hepatitis C virus

3.1 Summary

Background and Aims: HCV-related liver disease is a significant contributor to the global burden of liver disease. However, the viral genetic factors associated with HCV-related liver diseases are not well understood. This chapter reports novel viral polymorphisms associated with liver disease identified among subjects with HCV genotype 3a and genotype 1a infection from two large cohorts (gt3a, n = 1347; gt1a, n = 1181).

Methods: Patients involved in this study were from the STOP-HCV-1 trial, the BOSON clinical trial and the HCV Research UK biobank, which also provided clinical records. Whole consensus viral genomes were generated from plasma samples for 1347 HCV gt3a and 1181 HCV gt1a samples after excluding low-quality data. A viral genome-wide association study (V-GWAS) was undertaken to investigate the effect of viral amino acids across the full-length of the viral polyprotein on liver disease, including cirrhosis and HCC. The impact of liver-diseases-associated polymorphisms identified in the V-GWAS on baseline HCV RNA levels was also examined. Interactions between host genetic variants – such as the *IFNL4* gene and HLA molecules – and viral polymorphisms were investigated to reveal potential underlying mechanisms.

Results: The V-GWAS identified five novel polymorphisms in HCV gt3a and three in gt1a significantly associated with the presence of cirrhosis, which were termed Cirrhosis Associated Polymorphisms (CAPs). Further analysis of these CAPs in combination revealed an additive effect on cirrhosis prevalence. Patients with more CAPs had higher rates of cirrhosis, with a significant increase in odds ratios

correlating with the number of CAPs present. The impact of CAPs on viral load was also additive. Furthermore, the interplay between two gt3a CAPs and *IFNL4* genotype suggested that the effect of the *IFNL4* gene on cirrhosis status was restricted to viruses with specific amino acid residues. No polymorphisms were found to be significantly associated with HCC in this study.

Conclusions: For the first time, this chapter applies a comprehensive V-GWAS to identify viral determinants of liver disease progression in chronic HCV infection across different genotypes, aiming to uncover mechanisms of pathogenesis and provide a foundation for improved risk stratification and patient management.

3.2 Background

Chronic HCV infection is a leading cause of liver disease, driving persistent inflammation and metabolic dysfunction and thereby progressing to fibrosis, cirrhosis, and ultimately hepatocellular carcinoma (HCC). Cirrhosis begins in a compensated state, where liver function remains largely preserved. As the disease progresses, it can transition to decompensated cirrhosis, characterized by hepatic dysfunction and an increased risk of complications [195]. Cirrhosis from any aetiology represents the most significant risk factor for HCC. As an RNA virus, HCV does not integrate into the host genome, meaning that the risk of HCV-related HCC is largely restricted to those with HCV-induced cirrhosis or chronic liver damage accompanied by bridging fibrosis [196]. With the great success of direct-acting antiviral (DAA) therapy for HCV infection, the overall risk of liver disease progression has decreased [197]. However, patients with HCV-induced cirrhosis remain cirrhotic even after achieving SVR over years, and therefore continue to face a high risk of developing HCC [125, 198, 199].

The genome-wide association studies (GWAS) are helpful for uncovering genetic factors that affect disease risk, disease progression, and how people respond to treatment, thereby supporting improved diagnostics, prevention and therapeutics. Genetic variation in both the host and the virus can influence the liver disease outcomes by shaping immune recognition, affecting viral replication and modulating tissue damage. Host genetic variation has been extensively investigated. The impact of the polymorphisms in *IFNL4* gene on viral clearance has been described in detail in Chapter 2. Recent evidence showed that *PNPLA3*

SNP rs738409 (C>G) and *IFNL4* SNP rs12979860 (C>T) were associated with the risk of cirrhosis in patients with chronic hepatitis C infection [129, 130], and these polymorphisms were strong predictors of HCC occurrence among cirrhotic patients. HLA genes are known to play critical roles in the regulation of immune responses through controlling antigen presentation to T cells. Prior studies reported that three variations (rs910049, rs3135363, HLA-DQA1*0601) in the HLA class II region were significantly associated with progression from chronic hepatitis C to liver cirrhosis. Recently, a Japanese cohort [200] demonstrated that the interaction between killer cell immunoglobulin-like receptors (KIR) and HLA class I ligands through specific pair (KIR3DL1+HLA-Bw4) was associated with increased HCC risk. These findings underscore the importance of HLA not only as a mediator of antiviral immunity but also as a genetic factor contributing to liver disease progression.

In studying viral genetic determinants of liver disease in HCV chronic infection, previous studies have reported associations between specific viral amino acid polymorphisms, particularly within the core, NS3, and NS5A regions of HCV gt1b, and the risk of progressing to cirrhosis and HCC (Table 3.1) [131, 132, 201]. However, these analyses have been largely focused on a limited set of preselected regions of the HCV polyprotein, leaving much of the viral genome unexplored. As a result, the role of viral genetic variation across the entire HCV polyprotein in liver disease outcomes remains poorly defined. This gap is particularly evident for HCV genotypes 3a and 1a, which predominate in many global populations but for which systematic studies of viral sequence variation and disease progression are lacking. To address this gap, this chapter leveraged detailed clinical records with full-length HCV genome sequencing data from 1,347 patients with gt3a and 1,181 with gt1a

virus infection. This enabled genotype-specific analyses of how viral amino acid variants relate to the development of liver disease, specifically cirrhosis and HCC. Furthermore, by examining the interplay between host and viral genetic factors, this study provided insights into potential pathways involved in liver disease progression among patients chronically infected with HCV.

Genotype	Cohort size (n)	Location	Polymorphisms	Associated phenotype
HCV gt1b	49	Core	Gln70	HCC
HCV gt1b	1097	Core	Met91	HCC
HCV gt1b	49	NS3	Tyr1082/Gln1112	HCC
HCV gt1b	42	NS3	Thr1072	Cirrhosis
HCV gt1b	42	NS3	Val1074	Cirrhosis
HCV gt1b	42	NS3	Asn/Ile1098	Cirrhosis
HCV gt1b	49	NS5A	Asn2218	HCC

Table 3.1: Viral polymorphisms associated with liver disease reported in prior studies.

The numbering of position was according to the polyprotein of the prototype of HCV gt1b; HCV-J (H77).

3.3 Methods

3.3.1 Patients and samples

The study was conducted in collaboration with STOP-HCV, a consortium sponsored by the Medical Research Council in the United Kingdom, which contributed to the study design. It combined three cohorts of patients infected with HCV chronically, including the HCV Research UK clinical database and biobank (HCVRUK) [202], the BOSON clinical trial [173] and the STOP-HCV-1 trial [203].

The HCVRUK patients were enrolled by consent into the HCV Research UK registry. Ethics approval for HCV Research UK was given by NRES Committee East Midlands – Derby 1 (Research Ethics Committee reference 11/EM/0314). It includes 10804 subjects in total with past or current HCV chronic infection from 56 UK specialist clinics during March 2012-July 2015. Participants provided complete epidemiological and clinical data at baseline and during follow-up, along with blood samples stored in a biorepository ('biobank'). The cohort demonstrated prolonged HCV infection, with a mean duration of 23.4 years at enrolment. Genotypes 1 and 3 are the most prevalent HCV genotypes in the cohort. Due to the persistent infection, the prevalence of liver cirrhosis in this cohort almost reached a quarter. Across the cohort, patients exhibited either no liver disease, compensated (mild) liver disease, or decompensated cirrhosis (advanced liver disease). A small proportion later developed hepatocellular carcinoma (HCC). Cirrhosis was identified using histology (liver biopsy), imaging techniques, or validated serum biomarkers. Hepatocellular carcinoma (HCC) was defined mainly based on

hospital admission records, death certificates, or cancer registry entries indicating HCC (ICD-10 code C22.0; ICD-9 code 155.0).

The second cohort comprised patients with no or mild liver disease from the BOSON phase 3 trial conducted across five countries [173], mostly infected with HCV gt3a. The ethics approval for BOSON trial has been previously described in chapter 2. Cirrhosis in BOSON was diagnosed via liver biopsy or FibroScan (>12.5 kPa).

The STOP-HCV-1 trial (Strategic Treatment Optimization for HCV) was designed to understand the factors for selecting patients likely to achieve sustained virologic response with shorter direct-acting antiviral (DAA) courses, in accordance with World Health Organization treatment guidelines, as patients still find 8-12 weeks of DAA therapy challenging to complete. This trial was approved by the Cambridgeshire South Research Ethics Committee (15/EE/0435) and registered with ISRCTN (37915093, 11th April 2016) and EudraCT (2015-005004-28, 31st December 2015). A total of 202 eligible participants enrolled. All were adults (>= 18 years) with chronic HCV gt1a/1b/4 infection, had no evidence of significant liver fibrosis (FibroScan <= 7.1 kPa), and had not previously received DAA exposure for their current infection (previous pegylated-interferon/ribavirin allowed) [203].

As not all patients had both reliable clinical records and viral sequencing data, the final analyses included 2956 patients across the three study cohorts whose clinical and sequencing data met the eligibility criteria. This dataset comprised 1347

individuals infected with gt3a, 1181 with gt1a, 228 with gt1b, and 200 with other genotypes. The analyses in this chapter focused specifically on gt3a and gt1a.

3.3.2 Viral sequencing

A detailed description of the how the patient samples were sequenced and the bioinformatic pipeline used to process the sequencing data can be found in chapter 2. Of note, plasma samples obtained from the three cohorts were sent to Oxford, University of Glasgow and UCL to sequence for full length HCV genome data. The outputs from the bioinformatics pipeline were analysed using custom code and the R statistical package. Briefly, high quality whole consensus viral genomes from plasma samples were generated from 2,956 participants after filtering.

3.3.3 Host genome-wide genotyping and imputation

Blood specimens collected at enrolment were genotyped using the Affymetrix UK Biobank array, with standard quality-control filters applied prior to imputation. Imputation was performed on the TOPMed Imputation Server, and classical HLA alleles were further imputed using SNP2HLA with the UK Biobank reference panel. Only HLA alleles and *IFNL4* variants were included in the analyses.

3.3.4 Host and virus genetic principal component analysis

Principal component analysis (PCA) on host and virus genetic data was conducted to evaluate population substructure. For host, Kinship coefficients were estimated using PC-Relate (as implemented in the R package GENESIS) [204] to select unrelated individuals. The R package SNPRelate [205] was then used for PCA on pruned common genotyped SNPs from a set of unrelated individuals, defined as those with pairwise kinship coefficients less than $2^{-9/2}$. These two estimation procedures were iterated to ensure that the kinship coefficients were unbiased in admixed individuals and that the PCs were computed on an unrelated set. For the viral data, PCA was performed on the nucleotide data as follows. Bi, tri- and quad-allelic nucleotide sites were converted to binary variables and were standardised to have mean zero and unit variance. MATLAB (release 2015a, The MathWorks) was used to perform the PCA using the singular value decomposition function.

3.3.5 Statistical analysis

Phylogenetics and association of a phenotype with tree structure

Maximum-likelihood phylogenetic trees from full-length HCV genome sequences were constructed using IQ-TREE v2.2.2.6 [206] under the GTR+I+G nucleotide substitution model, with branch support estimated using 1000 ultrafast bootstrap replicates and 1000 SH-aLRT replicates. TreeBreaker (v1.0) software [207], which can infer the evolution of a discrete phenotype distribution on a phylogenetic tree and identify clades where the distribution of the phenotype is different to the rest of the tree, was then applied to investigate whether cirrhosis or HCC status were associated with specific clades. If such associations were detected, logistic regression was then repeated with an additional indicator variable for the samples

including the clade was added to the regression model as a covariate to ensure that viral population structure was not confounding the association.

Viral genome-wide association study (V-GWAS) on liver disease

To investigate the effect of individual viral amino acid polymorphisms on liver disease, a V-GWAS was performed using a logistic regression model. Each liver disease served as the response variable, with the presence and absence of each amino acid as the explanatory variable. To ensure adequate statistical power, the analysis was restricted to amino acid residues present in sequences from at least 50 patients. The model also included age, sex, ethnicity, sequencing centre, $\log_{10}$ viral load, viral population structure (viral PCs) and viral sequence clades that were associated with liver disease as covariates to account for confounding. Analyses were conducted separately in a genotype-specific manner.

Covariation and interaction analysis

Pearson correlation and logistic regression were used to assess covariation among any novel liver-disease-associated polymorphisms identified. Logistic regression was further used to test whether interactions between covarying sites influenced liver disease outcomes.

The impact of identified polymorphisms on other clinical outcomes

Logistic regression was conducted to examine the association between cirrhosis associated polymorphisms (CAPs) and HCC status. Linear regression was performed to assess the relationship between liver-disease-associated polymorphisms and viral replication.

Interplay between* CAPs and *IFNL4 SNP

To investigate the impact of *IFNL4* gene on CAPs, logistic regression was performed first to examine the association between *IFNL4* genotypes (CC vs. non-CC) and individual CAP, adjusting for viral population structure as covariates. Conditional logistic regression including *IFNL4* genotypes was then used to evaluate whether the effect of *IFNL4*-associated CAPs on cirrhosis was mediated by the *IFNL4* SNP. Interaction between *IFNL4* genotypes and CAPs was tested using logistic regression including interaction terms.

Interplay between CAP and HLA molecules

To explore the impact of HLA alleles on CAPs, logistic regression was performed to examine the association between 148 two-field HLA alleles and each CAP, adjusting for viral and host principal components, age, sex, ethnicity and sequencing centre. Conditional logistic regression including the associated HLA alleles was used to evaluate whether the effect of CAPs on cirrhosis was influenced by the HLA allele. Finally, interaction between HLA alleles and CAPs was tested using logistic regression containing interaction terms.

3.4 Results

3.4.1 Cohort description

Patients in our study came from the BOSON clinical trial [173], STOP-HCV-1 trial [203] and HCVRUK [202]. The cohort was primarily composed of individuals infected with HCV genotype 3a (gt3a; n = 1347, 45.6%) and genotype 1a (gt1a; n = 1181, 40%)(Table 3.2). Since other HCV genotypes and subtypes accounted for 428 samples, this study focused on the gt3a and gt1a, which together constitute more than 85% of the dataset.

For the 1347 patients infected with HCV gt3a, the median age was 60 years (range 29 – 87), and males constituted 70.4% (n = 947) of the cohort. The predominant ethnicity was White (81.0%, n = 1076), followed by Asian (16.7%, n = 225) and other ethnicities (1.9%, n = 26). A total of 647 patients had no diagnosis of cirrhosis, while 682 patients had cirrhosis (412 compensated and 270 decompensated). Hepatocellular carcinoma (HCC) was diagnosed in 13.1% (n = 176) of gt3a patients, all of whom had a prior diagnosis of cirrhosis.

In the HCV gt1a cohort of 1181 patients, the median age was also 60 years. The majority were White males (n = 884 males, n = 1106 White). Cirrhosis status was similarly to that observed in gt3a patients, with 622 patients having no cirrhosis diagnosis and 548 patients diagnosed with cirrhosis (322 compensated, 226 decompensated). All 122 HCC patients (10.3%) within the gt1a group also had a documented history of cirrhosis. Notably, the proportions of cirrhosis and HCC in both cohorts are higher than expected in unselected HCV populations, as the study

was intentionally enriched for patients with advanced liver disease rather than representing a random sample.

	HCV Gt3a	**HCV Gt1a**
Total patients: n	1347	1181
Age: medium (range)	60 (29-87)	60 (25-88)
Sex: male n (%)	947 (70.4%)	884 (75.0%)
Ethnicity: n (%)[a] *White* *Asian* *Black/ Mixed/ Other*	 1076 (79.9%) 225 (16.7%) 26 (1.9 %)	 1106 (93.6%) 24 (2.0%) 34 (2.9%)
Cirrhosis status: n (%) [b] *no-cirrhosis(cHCV)* *Compensated* *Decompensated*	 647 (48.7%) 412 (31.0%) 270 (20.3%)	 622 (53.2%) 322 (27.5%) 226 (19.3%)
HCC status: yes n (%)	173 (13.1%)	122 (10.3%)

Table 3.2: Demographic and clinical characteristics of the study cohort.

[a]missing information for 20 gt3a patients and 17 gt1a patients

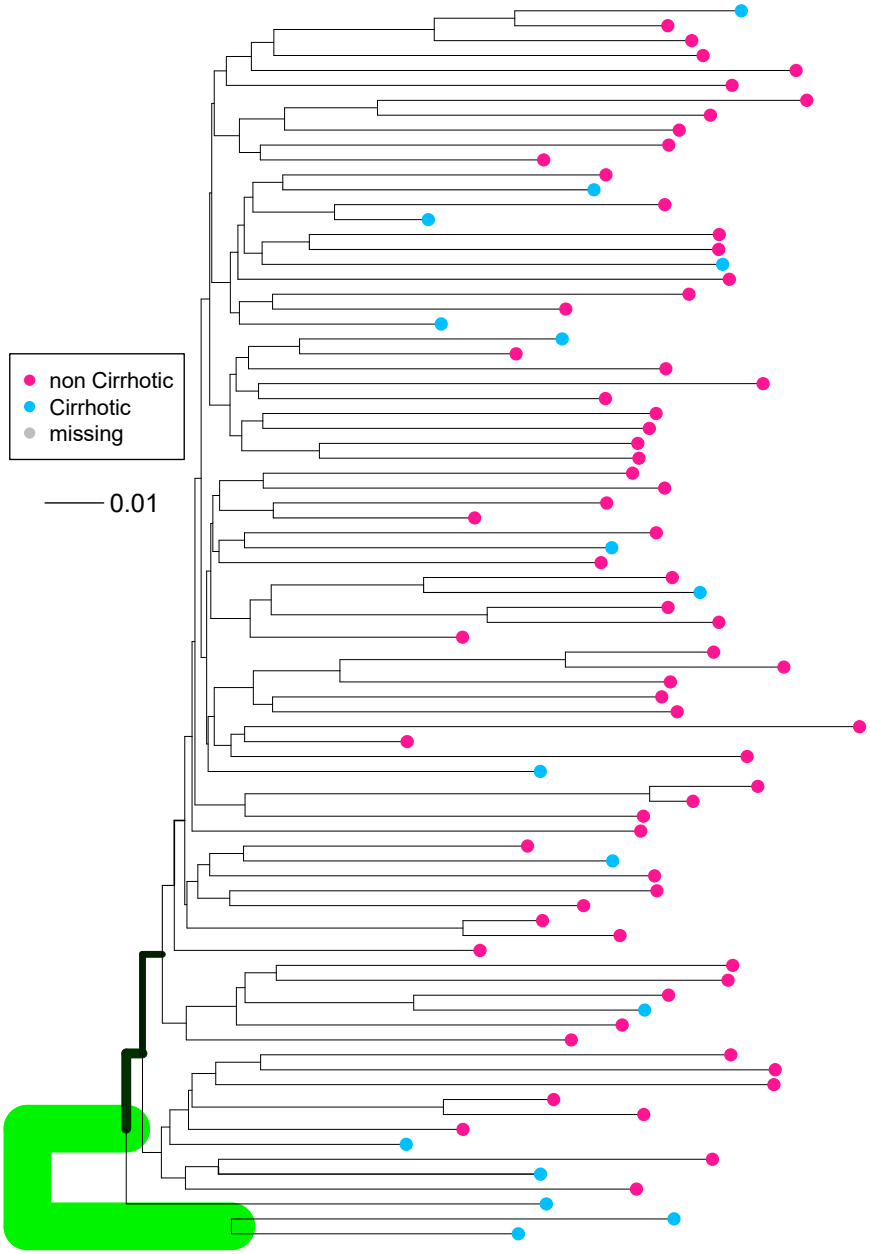
[b]missing information for 18 gt3a patients and 11 gt1a patients

3.4.2 Association between HCV clades and Cirrhosis

As a first step, the association between HCV lineages and cirrhosis status was assessed using treeBreaker software [207] to determine whether cirrhosis clustered within particular viral lineages. This step was important to evaluate potential confounding by viral population structure before testing for associations at the level of individual amino acid polymorphisms. Bayes factors calculated from treeBreaker close to 1 indicate no evidence for an association, values greater than 3 suggest moderate evidence, and values greater than 10 are considered strong evidence. For patients infected with HCV gt3a, analysis with treeBreaker did not

identify any clades with cirrhosis frequencies differing from the rest of the tree (Bayes factor = 1.09) (Supplementary Figure B.1). In contrast, for HCV gt1a, two clades, designated as clades A and B, showed significantly lower proportions of cirrhotic patients (Bayes factor = 61.5) (Figure 3.1, Supplementary Figure B.2). Clade A comprised 81 sequences (14 with cirrhosis; cirrhosis rate = 17.3%, Table 3.3), and Clade B comprised 35 sequences (5 with cirrhosis; cirrhosis rate = 14.3%, Table 3.3). Both clades had lower rates of cirrhosis relative to the rest of the tree, where the average rate of cirrhosis was 49.7% (529 of 1065 patients). Notably, patients in these clades were, on average, younger, and Clade B in particular was largely drawn from STOP-HCV-1 trial, which did not enrol patients with cirrhosis. These demographic and cohort differences likely account for the lower cirrhosis rates observed in the two clades. To control for this potential confounding effect, two indicator variables in subsequent analyses were included to denote whether each sample belonged to clade A or B.

A



B

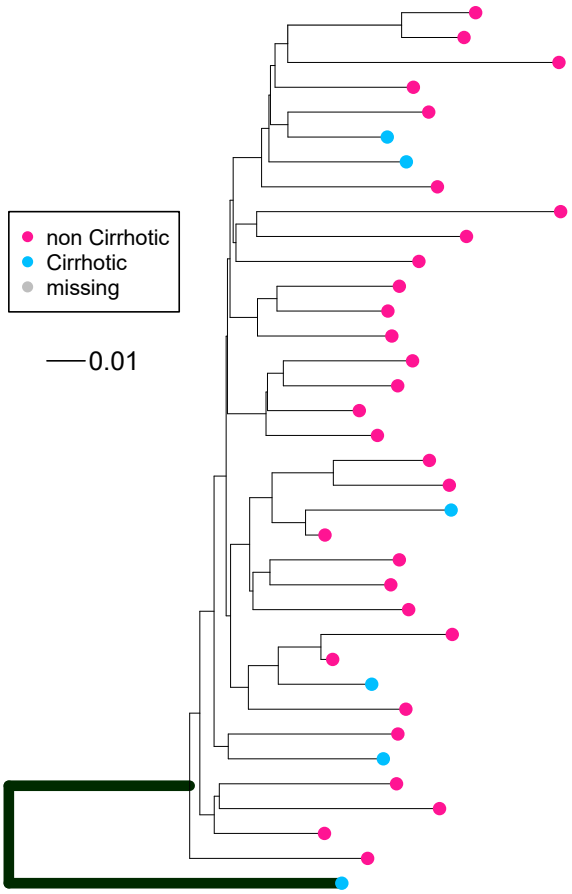


Figure 3.1: Two gt1a clades associated with low cirrhosis prevalence from the HCV gt1a phylogeny.

Detailed views of (A) Clade A and (B) Clade B, representing lineages with significant associations to cirrhosis distribution. The full maximum likelihood phylogenetic tree was constructed using whole HCV gt1a genome sequences using IQ-TREE (see full tree in Supplementary Figure B.2). The colour of the tips shows if the patient diagnosed with cirrhosis (blue) or not (pink). Line thickness and colour are proportional to the posterior probability of a change point on the branch. Green branches indicate an extreme high confidence change point with significant Bayes factors, identifying clades where the distribution of cirrhosis significantly diverges from the rest of the tree. The change point analysis conducted using treeBreaker. The scale bar represents 0.01 nucleotide substitutions per site.

	Clade A	**Clade B**	**The rest**
Total patients: n	81	35	1065
Age: medium (range)	55 (38-79)	55 (32-74)	60 (25-88)
Sex: male n (%)	59 (73.8%)	34 (97.1%)	791 (74.4%)
Ethnicity: n (%) *White* *Asian* *Black/ Mixed/ Other*	 79 (97.5%) 0 (0%) 2 (2.5%)	 30 (88.2 %) 0 (0%) 4 (11.8%)	 997 (95.0 %) 24 (2.3%) 28 (2.7%)
Cohort: n (%) *STOP-HCV-1* *BOSON* *HCVRUK*	 6 (7.4%) 0 (0%) 75 (92.6%)	 24 (68.6%) 0 (0%) 11 (31.4%)	 105 (9.9 %) 0 (0 %) 960 (90.1%)
Cirrhosis: n (%)	14 (17.3%)	5 (14.3%)	529 (50.2%)
HCC status: yes n (%)	5 (6.2%)	0 (0%)	117 (11.0%)
Mean log_{10} viral load	6.19	5.95	6.24

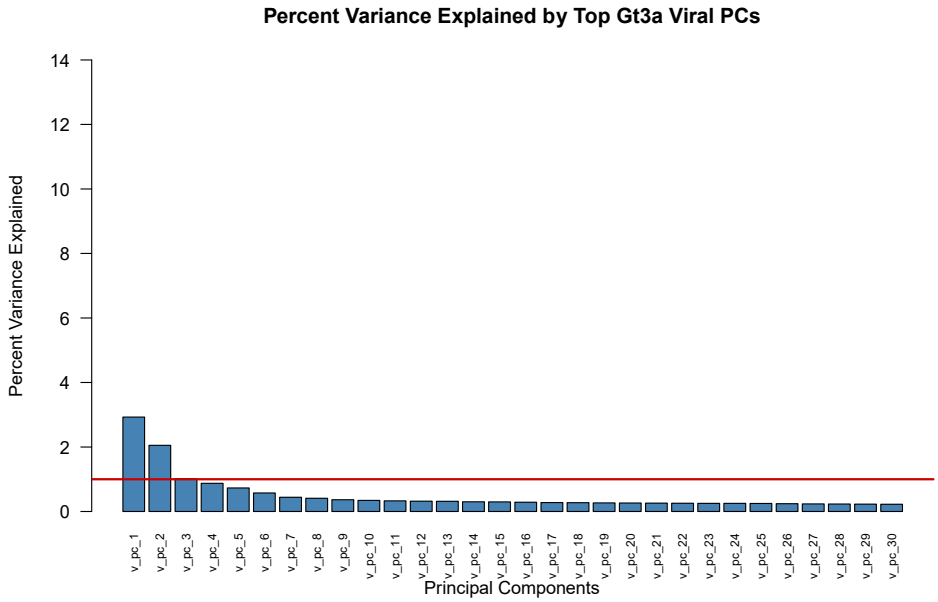
Table 3.3: Demographic and clinical characteristics of different clades.

3.4.3 Viral Genome-Wide Association Study (V-GWAS) for Cirrhosis

A V-GWAS was performed to test for associations between amino acid variation across the viral polyprotein and the presence or absence of cirrhosis using logistic regression. The model included age, sex, ethnicity, sequencing centre, and log_{10} viral load as covariates. Principal component analysis of viral genetic variation showed greater genetic diversity in gt1a than in gt3a (Figure 3.2). In gt3a, the first three viral principal components (PCs) together explained most of the between-sample variation, whereas additional PCs explained little variance. By contrast, in gt1a the first five viral PCs each accounted for >1% of variance and together

captured substantially more of the population structure. Based on these patterns, the first three viral PCs were added as covariates for the HCV gt3a analysis and the first five viral PCs for gt1a analysis to adjust for viral population structure. The gt1a model also included two indicator variables accounting for the confounding clades identified in 3.4.2. In all models, cirrhosis status was the outcome variable, and the presence or absence of specific amino acid residues at variable sites in the viral polyprotein served as the predictor. To ensure adequate statistical power, the analysis was restricted to amino acid residues present in sequences from at least 50 patients.

A



B

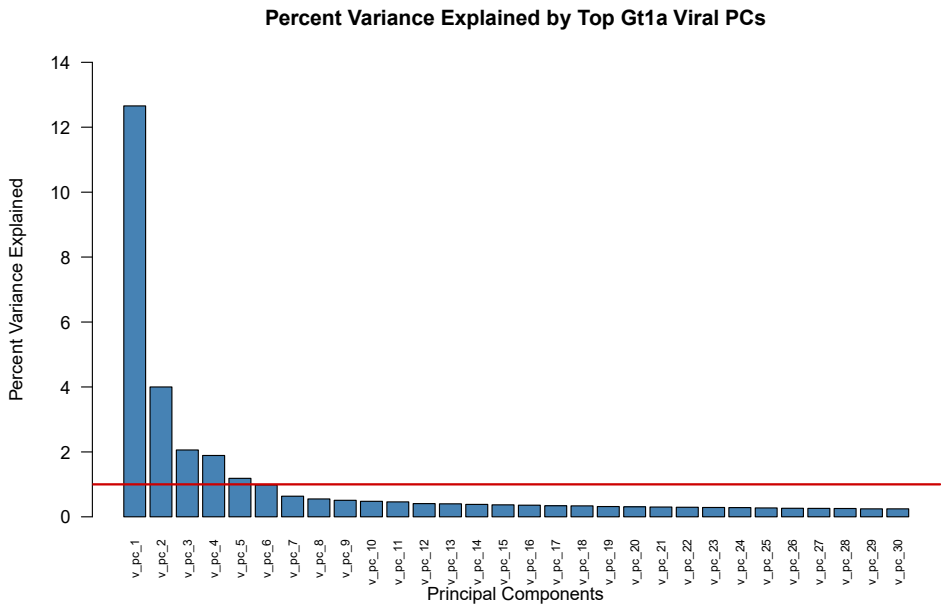


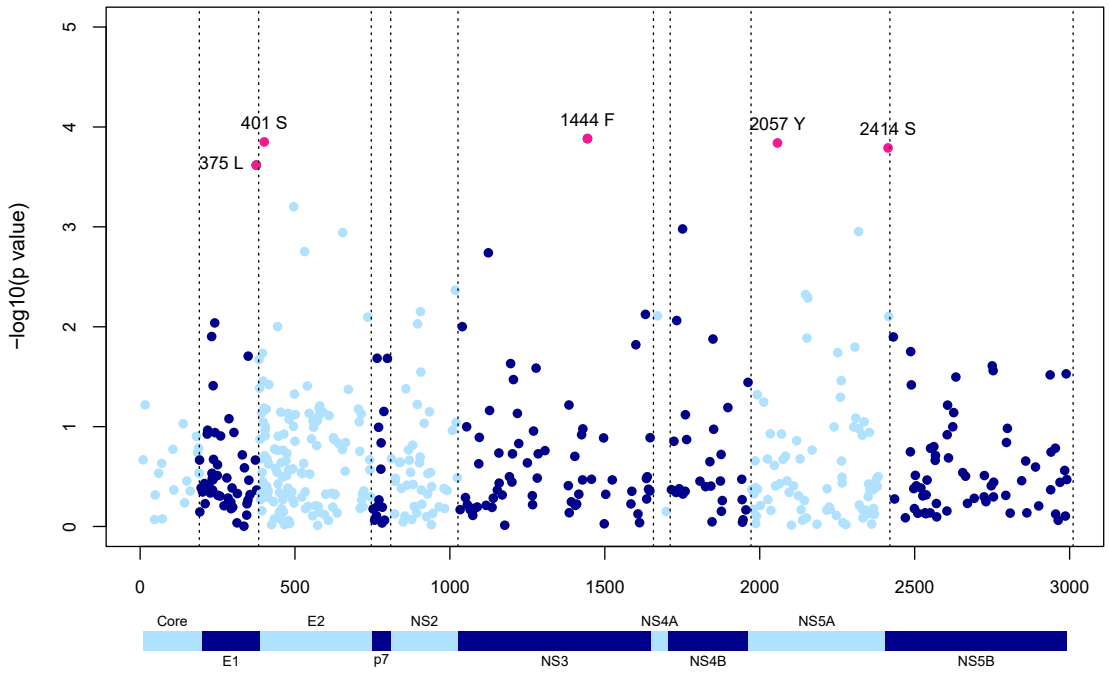
Figure 3.2: Percentage of variance explained by HCV viral PCs.
(A) Barplot of percent variance explained by the first 30 HCV gt3a viral PCs.
(B) Barplot of percent variance explained by the first 30 HCV gt1a viral PCs. Each bar represents the variance explained by each PC, ordered from most to least explanatory. The red line represents the 1% variance threshold. Three HCV gt3a viral PCs and five HCV gt1a viral PCs exceeded the threshold.

CAPs Identified in HCV Genotype 3a

For patients infected with HCV gt3a, 1089 tests at 508 unique sites were conducted. At a false-discovery rate (FDR) of 5%, five viral polymorphisms were identified significantly associated with presence or absence of cirrhosis (Figure 3.3 A&B, Table 3.4). Specifically, leucine (L) at position 375 in the E1 protein ($P = 2.41\times10^{-4}$), serine (S) at position 401 in the E2 protein ($P = 1.41\times10^{-4}$) were associated with higher rates of cirrhosis. The presence of phenylalanine (F) at position 1444 in NS3 ($P = 1.31\times10^{-4}$), tyrosine (Y) at position 2057 ($P = 1.44\times10^{-4}$) and serine (S) at position 2414 ($P = 1.62\times10^{-4}$) in NS5A protein were associated with lower rates of

cirrhosis relative to individuals carrying alternative residues at these positions (i.e., Y at 1444, non-Y at 2057, and non-S at 2414, respectively). The association between cirrhosis status and amino acid variation at position 375 in E1 (comparing L versus non-L residues) was primarily driven by the contrast between methionine (M) and valine (V) against leucine (L) ($P_{\text{L vs non-L}} = 2.4 \times 10^{-4}$, $P_{\text{L vs M}} = 1.4 \times 10^{-4}$, $P_{\text{L vs V}} = 2.8 \times 10^{-3}$, Figure 3.3B, Figure 3.4A). At position 401 in E2, most non-serine residues were associated with lower rates of cirrhosis including glycine (G), arginine (R), asparagine (N), Lysine (K) and threonine (T) (Figure 3.4B). At position 1444, tyrosine (Y) is the only alternative residue observed, with significant association with higher rate of cirrhosis ($P_{\text{Y vs F}} = 1.3 \times 10^{-4}$, Figure 3.4C). Similarly, at position 2057, histidine (H) is the only residue present at count of greater than 20 ($P_{\text{H vs Y}} = 1.8 \times 10^{-4}$, Figure 3.4D). Asparagine (N) is second most common variant at position 2414 after serine and the difference between the two residues is the driver of the signal ($P_{\text{S vs non-S}} = 1.6 \times 10^{-4}$, $P_{\text{S vs N}} = 9.7 \times 10^{-4}$, Figure 3.4E). Notably, the associations between each amino acid residue and cirrhosis status at each of the five sites were further used to guide the selection of mutants for in vitro validation (Table A.1), which will be discussed in chapter 5. (Table A.1 included the positions relative to H77 reference, the amino acid sequences surrounding the identified polymorphisms, and the viral strain used to produce the mutants.) For the rest of the chapter, Cirrhosis-Associated Polymorphisms or CAPs will be used to refer to the amino acid residues associated with a higher rate of cirrhosis. Thus, E1 375L, E2 401S, NS3 1444Y, NS5A 2057non-Y and NS5A 2414non-S are identified as CAPs in gt3a infection.

A



B

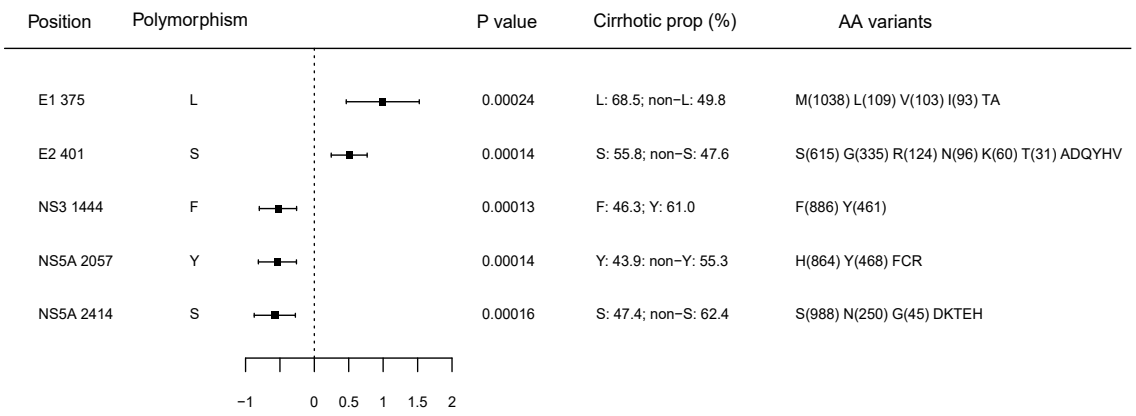


Figure 3.3: Association between HCV gt3a amino acids and cirrhosis status. (A) Manhattan plot of association tests between gt3a HCV amino acids and cirrhosis status. For each site only the most associated p-value is plotted. The pink dots with amino acid labelled indicate the five sites significantly associated with the presence or absence of cirrhosis at 5% FDR. The schematic of HCV polyprotein is under the Manhattan plot. (B) Forest plot of the impact of identified amino acid

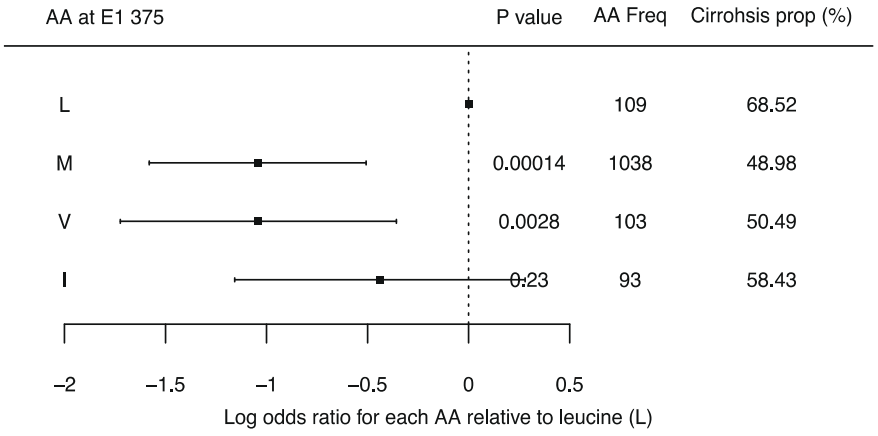
residues on cirrhosis status. The squares show the logistic regression log odds ratio for each tested amino acid residue and the lines show its 95% confidence interval (CI). The p values, proportions of cirrhosis and amino acid composition at each site are shown on the right. Amino acids with fewer than 20 occurrences are not showing their exact numbers.

The tested site	The most associated amino acid	P value	Effect size	Standard error	q value
1444	**f**	**0.00013**	**-0.528**	**0.138**	**0.040**
401	**s**	**0.00014**	**0.506**	**0.133**	**0.040**
2057	**y**	**0.00014**	**-0.537**	**0.141**	**0.040**
2414	**s**	**0.00016**	**-0.575**	**0.152**	**0.040**
375	**l**	**0.00024**	**0.994**	**0.271**	**0.048**
496	i	0.00063	0.511	0.149	0.089
1751	t	0.00105	-0.674	0.206	0.103
2319	q	0.00112	0.594	0.182	0.103
654	i	0.00114	-1.010	0.311	0.103
531	a	0.00177	-0.485	0.155	0.139

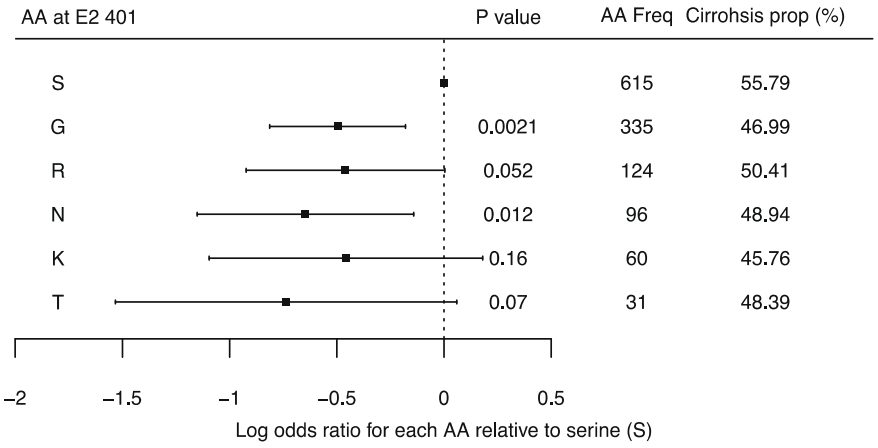
Table 3.4: Top ten associations between gt3a HCV amino acid variations and cirrhosis status.

The table presents the first ten significant results from V-GWAS testing the association between presence or absence of cirrhosis and polymorphisms at 508 sites (in total 1089 residues were tested at these 508 sites, but only the most associated residue per site is shown in the table). *P* values, effect sizes, and standard errors are reported, along with q-values representing FDR-corrected p-values to account for multiple testing. Associations at 5% FDR are highlighted in bold.

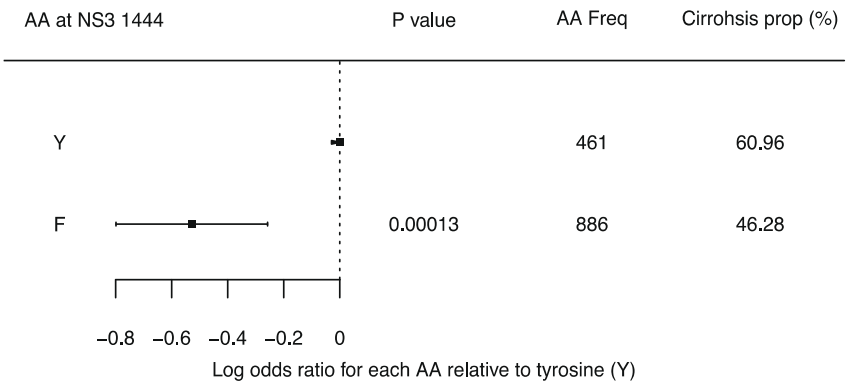
A


AA at E1 375
P value
AA Freq
Cirrohsis prop (%)
L
109
68.52
M
0.00014
1038
48.98
V
0.0028
103
50.49
I
0.23
93
58.43
–2
–1.5
–1
–0.5
0
0.5
Log odds ratio for each AA relative to leucine (L)

B


AA at E2 401
P value
AA Freq
Cirrohsis prop (%)
S
615
55.79
G
0.0021
335
46.99
R
0.052
124
50.41
N
0.012
96
48.94
K
0.16
60
45.76
T
0.07
31
48.39
–2
–1.5
–1
–0.5
0
0.5
Log odds ratio for each AA relative to serine (S)

C


AA at NS3 1444
P value
AA Freq
Cirrohsis prop (%)
Y
461
60.96
F
0.00013
886
46.28
–0.8
–0.6
–0.4
–0.2
0
Log odds ratio for each AA relative to tyrosine (Y)

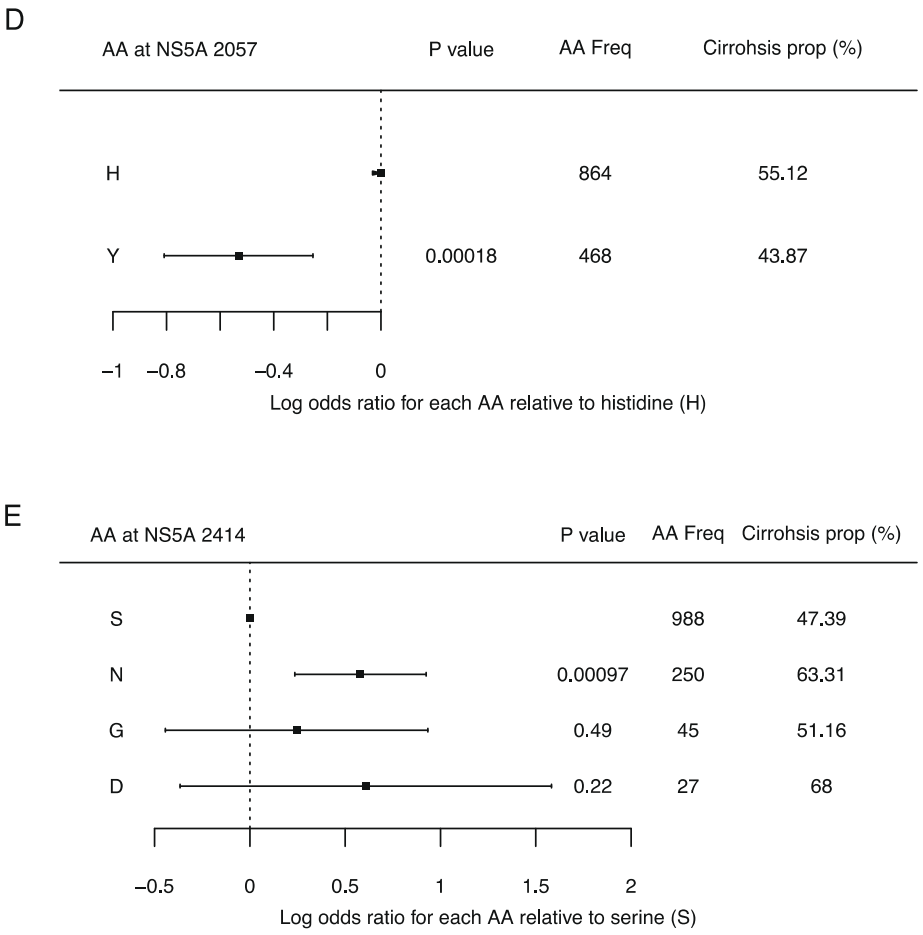


Figure 3.4: Association between each amino acid residue and cirrhosis status at each of the five sites identified within HCV gt3a infection.

(A) Amino acid variants at site 375. (B) Amino acid variants at site 401.

(C) Amino acid variants at site 1444. (D) Amino acid variants at site 2057.

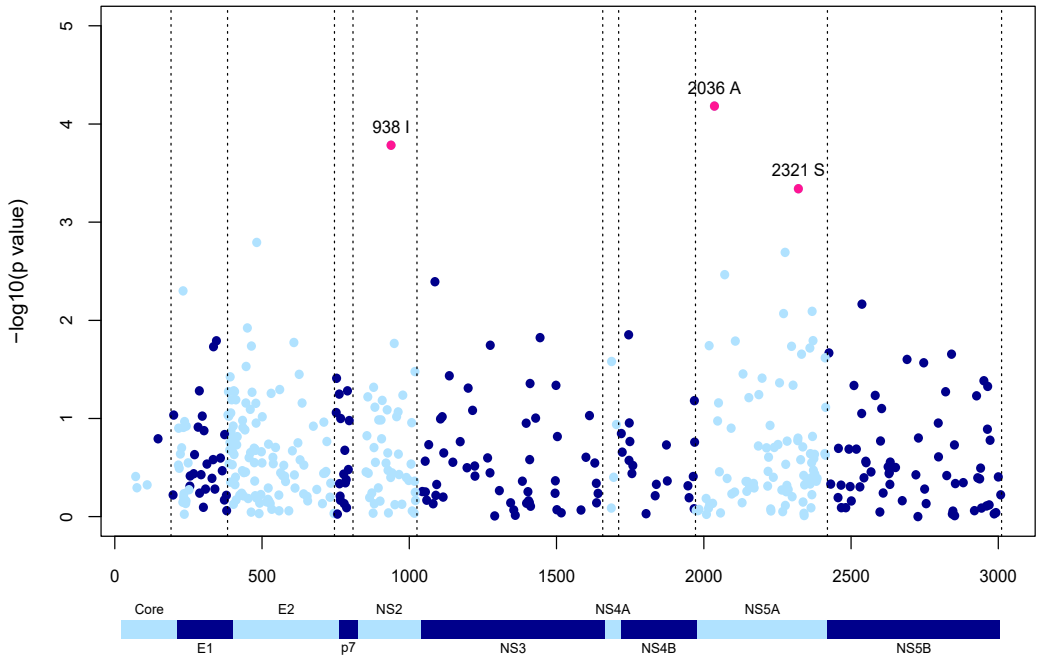
(E) Amino acid variants at site 2414.

Only amino acids with a frequency greater than 20 at each reported site are shown in the forest plot. The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression. *P* values were calculated using logistic regression model.

CAPs Identified in HCV Genotype 1a

For patients with HCV gt1a infection, 919 tests were performed across 458 unique sites. At a 5% FDR, no significant associations were observed between any variable amino acid sites and cirrhosis status (Figure 3.5 A and B and Table 3.5). To improve statistical power and minimize false negatives, the significance threshold was subsequently relaxed to 10% FDR. Under this criterion, three amino acid polymorphisms were found to be significantly associated with cirrhosis. The most significant association with cirrhosis was at position 2036 in the NS5A protein, where alanine (A) was associated with higher rate of cirrhosis compared to non-alanine residues at this position ($P = 6.56\times10^{-5}$). Isoleucine (I) at position 938 in the NS2 protein ($P = 1.64\times10^{-4}$) and serine (S) at position 2321 in the NS5A protein ($P = 4.57\times10^{-4}$) were also associated with higher rates of cirrhosis. The CAPs in HCV gt1a are NS2 938I, NS5A 2036A and NS5A 2321S. As with gt3a, associations between individual residues at those three reported sites and cirrhosis status were also assessed, indicating which specific residue comparisons drove the observed association. (Figure 3.6A-C). These associations were used to guide the production of mutants for in vitro validation in chapter 5 (Table A.2).

A



B

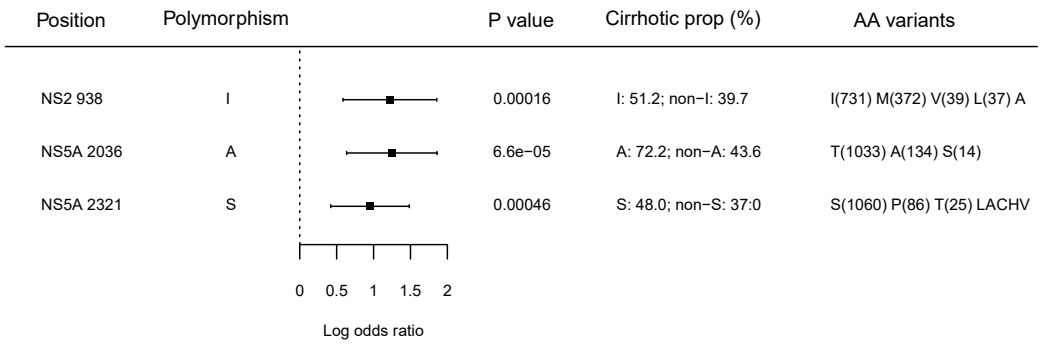


Figure 3.5: Association between HCV gt1a amino acids and cirrhosis status. (A) Manhattan plot of association tests between gt1a HCV amino acids and cirrhosis status. For each site only the most associated p-value is plotted. The pink dots with amino acid labelled indicate the three sites significantly associated with the presence or absence of cirrhosis at 10% FDR. The schematic of HCV polyprotein is under the Manhattan plot. (B) Forest plot of the impact of identified amino acid residues on cirrhosis status. The squares show the logistic regression log odds ratio for each tested amino acid residue and the lines show its 95% confidence interval (CI). The p values, proportions of cirrhosis and amino acid

composition at each site are shown on the right. Amino acids with fewer than 20 occurrences are not showing their exact numbers.

The tested site	The most associated amino acid	P value	Effect size	Standard error	q value
2036	**a**	**6.56E-05**	**1.248**	**0.313**	**0.053**
938	**i**	**0.00016**	**1.221**	**0.324**	**0.066**
2321	**s**	**0.00046**	**0.952**	**0.272**	**0.092**
482	q	0.00161	1.031	0.327	0.258
2276	l	0.00203	-1.274	0.413	0.272
2071	t	0.00342	0.801	0.274	0.393
1087	t	0.00405	-1.046	0.364	0.407
232	e	0.00501	1.155	0.412	0.447
2537	a	0.00685	-0.902	0.334	0.483
2368	s	0.00810	0.725	0.274	0.483

Table 3.5: Top ten associations between gt1a HCV amino acid variations and cirrhosis status.

The table presents the first ten significant results from V-GWAS testing the association between presence or absence of cirrhosis and polymorphisms at 458 sites (in total 919 residues were tested at these 458 sites, but only the most associated residue per site is shown in the table). P values, effect sizes, and standard errors are reported, along with q-values representing FDR-corrected p-values to account for multiple testing. Associations at 10% FDR are highlighted in bold.

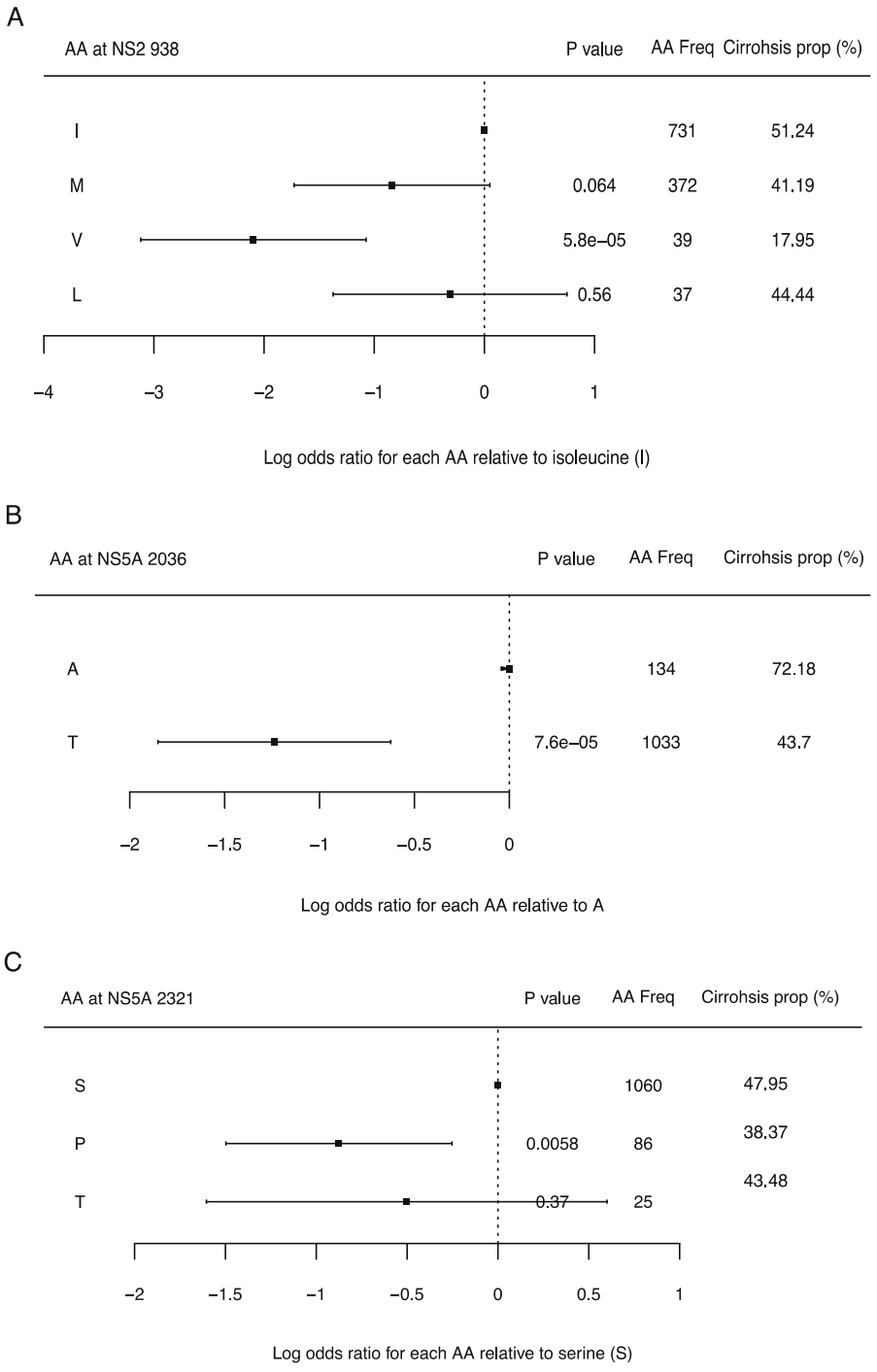


Figure 3.6: Association between each amino acid residue and cirrhosis status at each of the three sites identified within HCV gt1a infection.

(A) Amino acid variants at site 938. (B) Amino acid variants at site 2036. (C) Amino acid variants at site 2321.

Only amino acids with a frequency greater than 20 at each reported site are shown in the forest plot. The black squares and the lines indicate the estimated effect size

(log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression. *P* values were calculated using logistic regression model.

3.4.4 Covariation of CAPs and their impact on cirrhosis

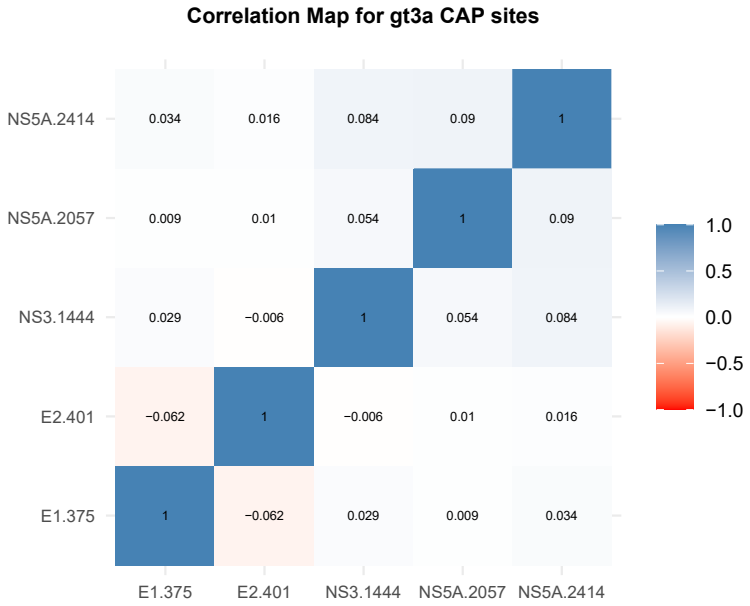
To assess the combined effect of the cirrhosis-associated polymorphisms (CAPs) on cirrhosis development, logistic regression was performed to test the association between cirrhosis status and combinations of CAPs in viral sequences, incorporating the same covariates used in the V-GWAS analysis.

Covariation of gt3a CAPs

First, covariation between pairs of CAPs was assessed using Pearson correlation analysis (Figure 3.7A). Correlations were generally weak, with most values close to zero, indicating little covariation among sites. The highest correlation was between sites 2057 and 2414 (r = 0.09), while the only negative correlation was between sites 375 and 401 (r = -0.062). These results suggested that the observed associations with cirrhosis at these positions are largely independent of one another. Logistic regression was then performed to assess associations among CAPs while adjusting for viral population structure by including the first three viral PCs as covariates. A nominal association was observed between sites E1 375 and E2 401 (P = 0.02; Figure 3.7B). Conditional analysis showed that the association of each site with cirrhosis remained significant after adjusting for the other, indicating independent effects. Specifically, the log odds ratio (logOR) for E1 375 was 0.99 ($P = 2.41 \times 10^{-4}$) and increased slightly to 1.10 ($P = 9.57 \times 10^{-5}$) after adjusting for E2 401. Similarly, the logOR for E2 401 was 0.51 ($P = 1.41 \times 10^{-4}$)

and remained consistent after adjusting for E1 375 (logOR = 0.54, $P = 5.97 \times 10^{-5}$) (Figure 3.8A). In addition, the CAP at NS5A 2414 showed significant associations with CAPs at NS3 1444 and NS5A 2057 ($P = 0.003$ and $P = 0.0002$, respectively; Figure 3.7B). Despite these associations, conditional analyses indicated that each variant contributed independently to the presence of cirrhosis (Figure 3.8B).

A



B

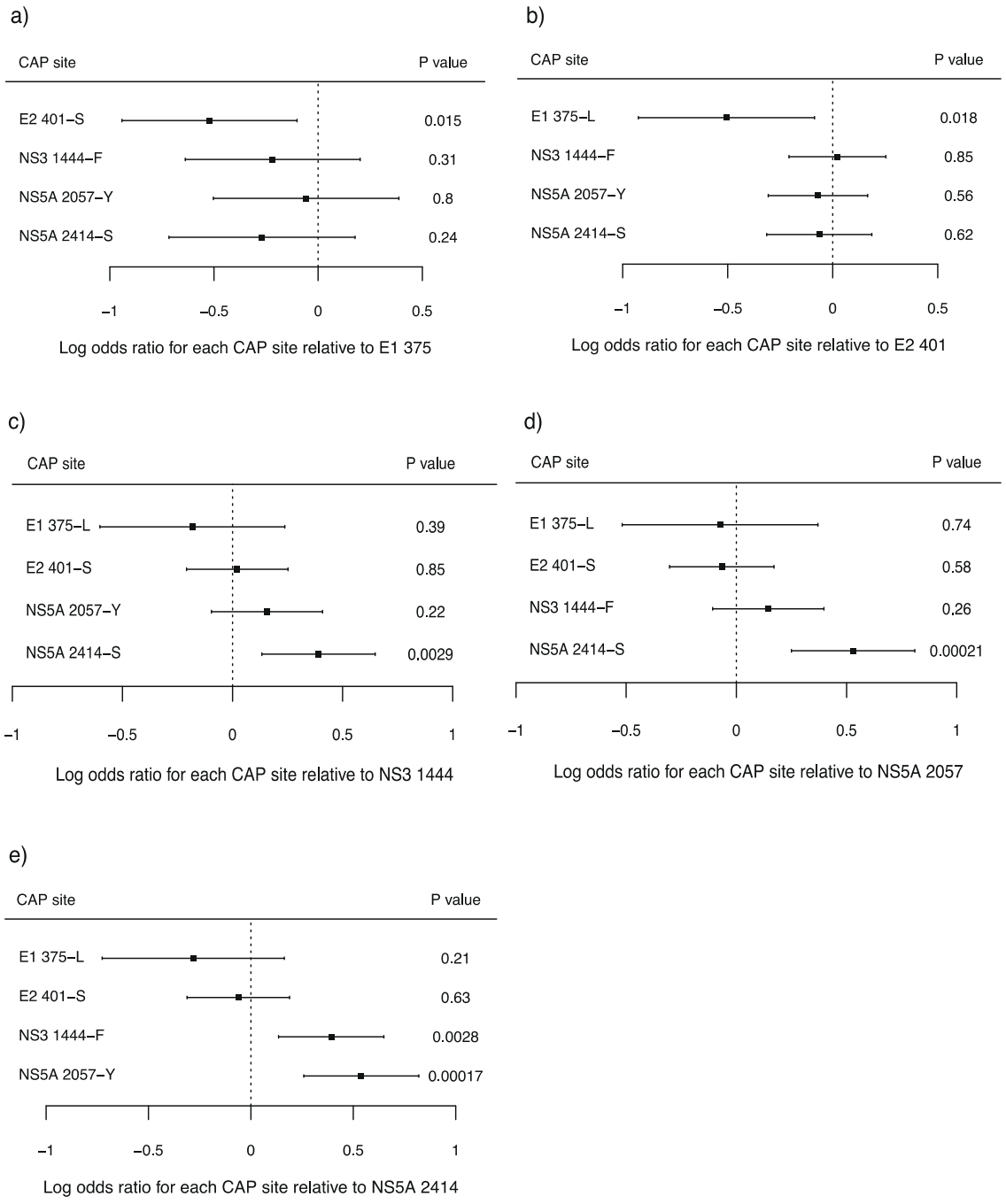
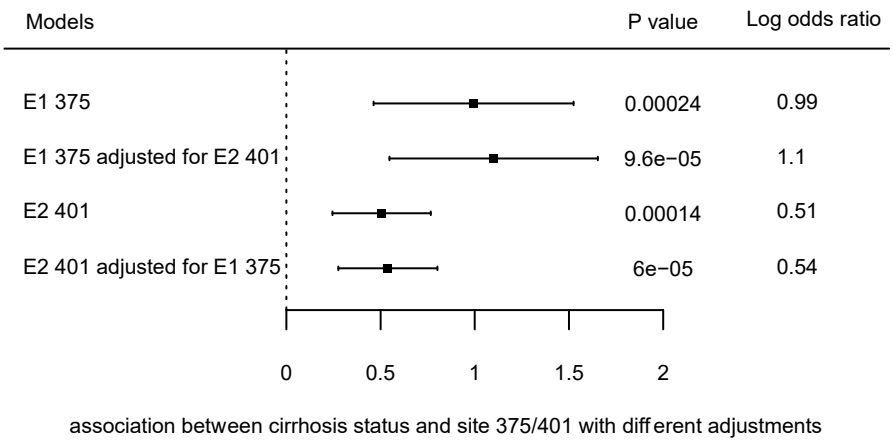


Figure 3.7: Covariation between pairs of gt3a CAPs.

(A) Pairwise Pearson correlation matrix of binary CAP residues. Color intensity demonstrates the power and direction of correlation, with blue indicating positive and red indicating negative correlations. Only weak correlations were observed among those five pairs.

(B) Forest plots of association between the five reported sites using logistic regression. We set each site as the reference group to compare against all others, presenting the resulting log odds ratios and p values. The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

A



B

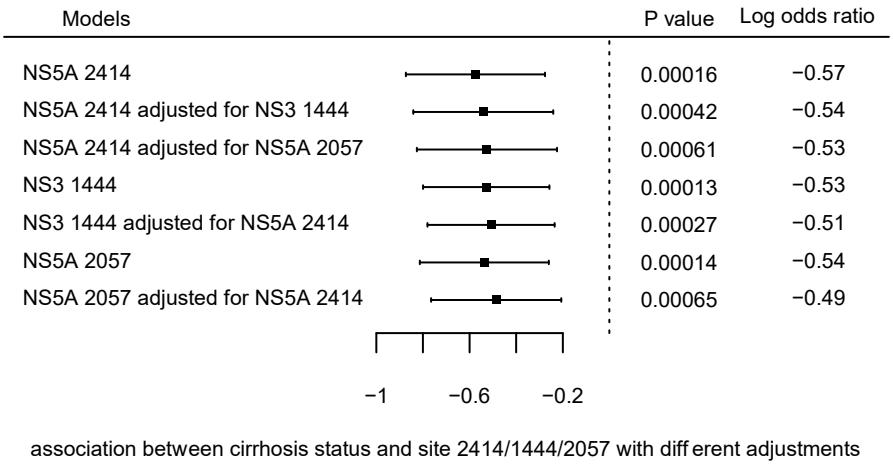


Figure 3.8: Conditional analysis of cirrhosis status after covariate adjustment.

(A) First two lines indicated association test between cirrhosis status and site 375, with and without adjustment for its associated site 477. The latter two lines

indicated association test between cirrhosis status and site 477, with and without adjustment for its associated site 375.
(B) Association test between cirrhosis status and site 2414, with and without adjustment for its associated site 1444 or site 2057; association test between cirrhosis status and site 1444, with and without adjustment for its associated site 2414; association test between cirrhosis status and site 2057, with and without adjustment for its associated site 2414.
The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

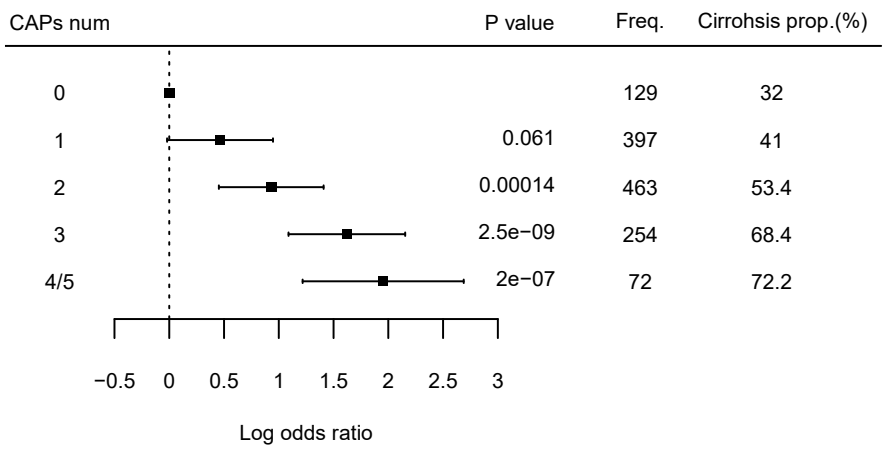
Cumulative effect of gt3a CAPs on cirrhosis

To assess the cumulative effect of cirrhosis-associated polymorphisms (CAPs), patients were stratified by the total number of CAPs present in their consensus viral sequences. Logistic regression analysis revealed a strong positive association between the number of CAPs and cirrhosis status (logOR = 0.52, P = 1.14×10^{-14}), indicating that each additional CAP significantly increased the rate of cirrhosis. The number of CAPs was also modeled as a categorical variable to explore how cirrhosis prevalence varied across groups. Compared to patients with no CAPs, cirrhosis prevalence increased progressively with CAP burden: 41.0% for those with one CAP (P = 0.061), 53.4% with two (P = 1.4×10^{-4}), 68.4% with three (P = 2.5×10^{-9}), and 72.2% with four or five CAPs (P = 2×10^{-7}). The latter two groups were combined due to limited sample size. These results suggested a stepwise increase in cirrhosis prevalence with a rising number of CAPs (Figure 3.9A). However, it should be noted that these findings reflected cirrhosis prevalence within the study sample rather than in the general population. Because this study was conducted using a case-control design, with enrichment for

individuals with liver disease, the observed rates of cirrhosis were generally higher than would be expected in a population-based cohort.

To explore this further, patients were further categorized based on the specific presence or absence of the five individual CAPs, resulting in 31 unique CAP combinations. Logistic regression was used to assess the association between each combination and cirrhosis status (Figure 3.9B; Table 3.6). To ensure sufficient statistical power and stable estimates, the analysis was restricted to combinations observed in more than 20 individuals. Among patients whose viral sequences lacked all five CAPs (E1 375 non-L, E2 401 non-S, NS3 1444F, NS5A 2057Y, and NS5A 2414S), the cirrhosis prevalence was 32.0%. In contrast, all other CAP combinations were associated with higher cirrhosis rates, with the proportion increasing in tandem with the number of CAPs present. Notably, combinations with more than three CAPs had a minimum cirrhosis prevalence of 63%, indicating the additive impact of CAPs on disease risk.

A



B

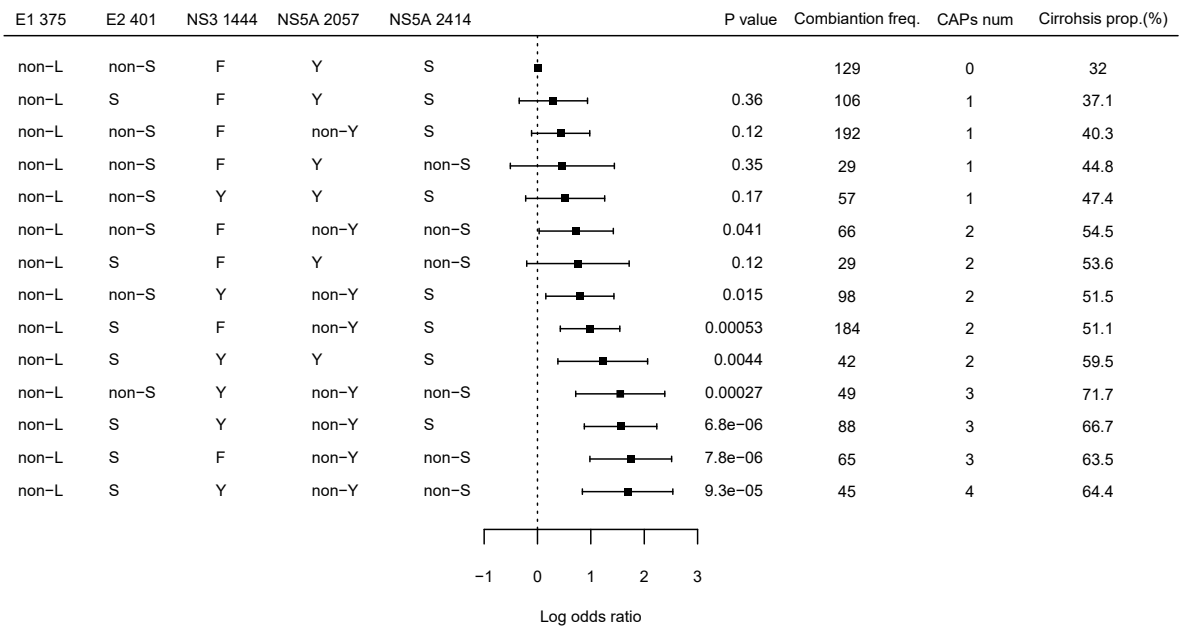


Figure 3.9: Cumulative impact of gt3a CAPs on cirrhosis status.

(A) Cirrhosis proportion for patients infected with HCV gt3a with different numbers of CAPs in baseline sequences. The total number of CAPs present at each group is shown on the left of the table. The p-values (logistic regression), frequency of each group and the cirrhosis rate for each group are shown on the right.

(B) Covariation between the five gt3a CAPs and their associations with cirrhosis. Combinations with a frequency greater than 20 involved in tests are listed in the forest plot on the left. The p-values (logistic regression), frequency of each tested combination, number of CAPs carried by each combination and the cirrhosis rate for each combination are shown on the right.

The black squares and the lines indicate the estimated effect size (log (odds ratio)) and its 95% confidence interval estimated from logistic regression.

CAP combinations (n = 31)					Regression outcomes				Combination characteristics		
E1 375	E2 401	NS3 1444	NS5A 2057	NS5A 2414	logOR	SE	z	p value	Combination freq.*	CAPs num	Cirrhosis prop. (%)
non-L	non-S	F	Y	S	0	0	0	NA	129	0	32.0%
non-L	non-S	Y	Y	S	0.519	0.377	1.376	0.169	57	1	47.4%
non-L	S	F	Y	S	0.296	0.326	0.907	0.365	106	1	37.1%
non-L	non-S	F	non-Y	S	0.435	0.278	1.566	0.117	192	1	40.3%
non-L	non-S	F	Y	non-S	0.466	0.498	0.935	0.350	29	1	44.8%
L	non-S	F	Y	S	2.242	0.754	2.973	0.003	13	1	46.2%
non-L	S	Y	Y	S	1.223	0.429	2.850	0.004	42	2	59.5%
non-L	non-S	Y	non-Y	S	0.794	0.325	2.441	0.01	98	2	51.5%
non-L	S	F	non-Y	S	0.985	0.284	3.466	0.001	184	2	51.1%
non-L	non-S	Y	Y	non-S	1.559	0.816	1.910	0.056	13	2	69.2%
non-L	S	F	Y	non-S	0.758	0.489	1.551	0.121	29	2	53.6%
non-L	non-S	F	non-Y	non-S	0.726	0.355	2.047	0.041	66	2	54.5%
L	non-S	Y	Y	S	-15.738	782.147	-0.020	0.984	3	2	0.0%
L	S	F	Y	S	2.127	0.917	2.320	0.020	10	2	50.0%
L	non-S	F	non-Y	S	1.465	0.702	2.087	0.037	18	2	66.7%
non-L	S	Y	non-Y	S	1.556	0.346	4.499	6.83E-06	88	3	66.7%
non-L	S	Y	Y	non-S	1.453	0.688	2.113	0.035	18	3	66.7%
non-L	non-S	Y	non-Y	non-S	1.551	0.426	3.642	0.000	49	3	71.7%
non-L	S	F	non-Y	non-S	1.747	0.391	4.470	0.000	65	3	63.5%
L	S	Y	Y	S	15.040	1013.739	0.015	0.988	2	3	100.0%
L	non-S	Y	non-Y	S	2.271	0.931	2.440	0.015	10	3	80.0%
L	S	F	non-Y	S	1.244	0.895	1.390	0.165	7	3	71.4%
L	non-S	Y	Y	non-S	15.647	800.939	0.020	0.984	3	3	100.0%
L	S	F	Y	non-S	15.074	1455.398	0.010	0.992	1	3	100.0%
L	non-S	F	non-Y	non-S	1.140	0.869	1.313	0.189	11	3	70.0%
non-L	S	Y	non-Y	non-S	1.688	0.432	3.907	9.35E-05	45	4	64.4%
L	S	Y	non-Y	S	1.700	0.796	2.136	0.033	9	4	66.7%
L	S	Y	Y	non-S	15.921	748.720	0.021	0.983	3	4	100.0%
L	non-S	Y	non-Y	non-S	2.959	1.140	2.596	0.009	9	4	88.9%
L	S	F	non-Y	non-S	15.837	708.915	0.022	0.982	4	4	100.0%
L	S	Y	non-Y	non-S	14.929	961.776	0.016	0.988	2	5	100.0%

* *Results for combinations with low frequency (n < 20) are shown for descriptive purposes only and should be interpreted cautiously. Particularly, estimates for combinations with n < 5 are extremely unstable and not statistically meaningful.*

Table 3.6: Associations between gt3a CAP combinations and cirrhosis status.

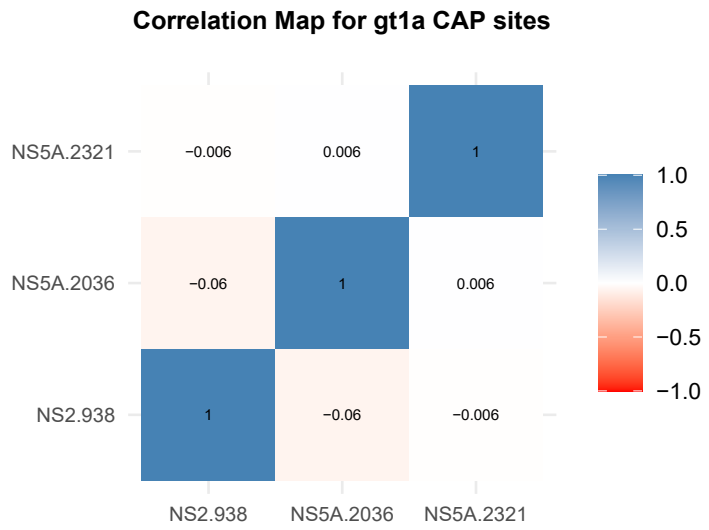
The table presents the total 31 unique combinations based on the specific presence or absence of the five individual CAPs in patients' sequences. The

logistic regression outcomes (p values, effect sizes, standard errors, et al) and characteristics of these combinations were also shown in the table.

Covariation of gt1a CAPs

A similar analysis was conducted for gt1a CAPs to assess their combined effect on cirrhosis. Pearson correlation analysis showed that correlations between sites were minimal, with values ranging from -0.06 to 0.006, indicating that amino acid variation at these positions occurs independently (Figure 3.10A). These results suggested that the observed associations with cirrhosis are unlikely to be confounded by linkage between the sites. Logistic regression adjusted for the first five viral principal components detected a marginal association between NS5A 2321 and NS2 938 (P = 0.03, Figure 3.10B), although their effects on cirrhosis appeared to be independent. The log odds ratio (logOR) for NS5A 2321 was 0.95 ($P = 4.6 \times 10^{-4}$) and decreased slightly to 0.91 ($P = 9.3 \times 10^{-4}$) after adjusting for NS2 938 (Figure 3.11).

A



B

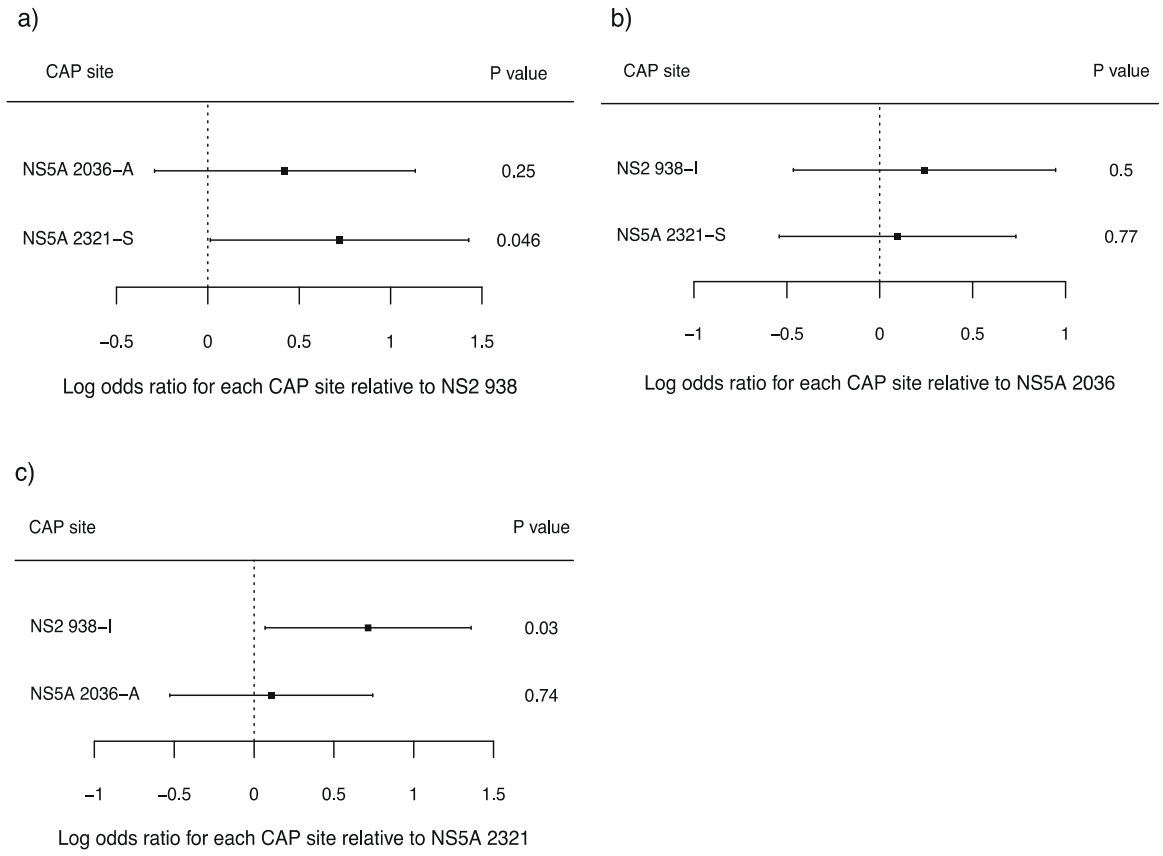


Figure 3.10: Covariation between pairs of gt1a CAPs.

(A) Pairwise Pearson correlation matrix of binary CAP residues. Color intensity demonstrates the power and direction of correlation, with blue indicating positive and red indicating negative correlations. Only weak correlations were observed among those five pairs.

(B) Forest plots of association between the three reported sites using logistic regression. We set each site as the reference group to compare against all others, presenting the resulting log odds ratios and p values. The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

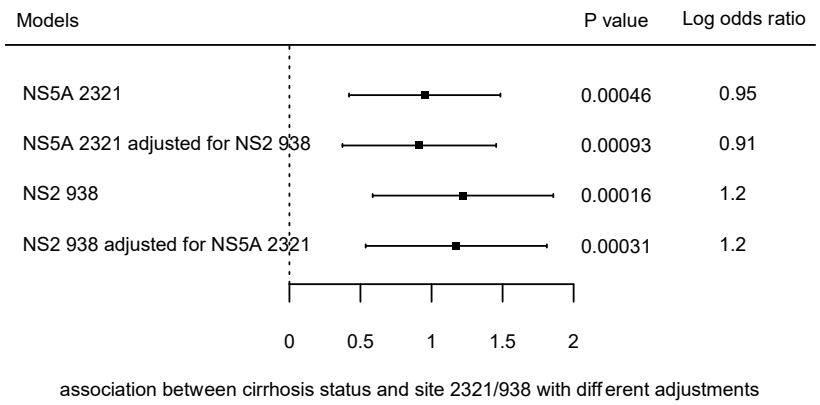


Figure 3.11: Association test between cirrhosis status and site 2321/938, with and without different adjustments.

The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

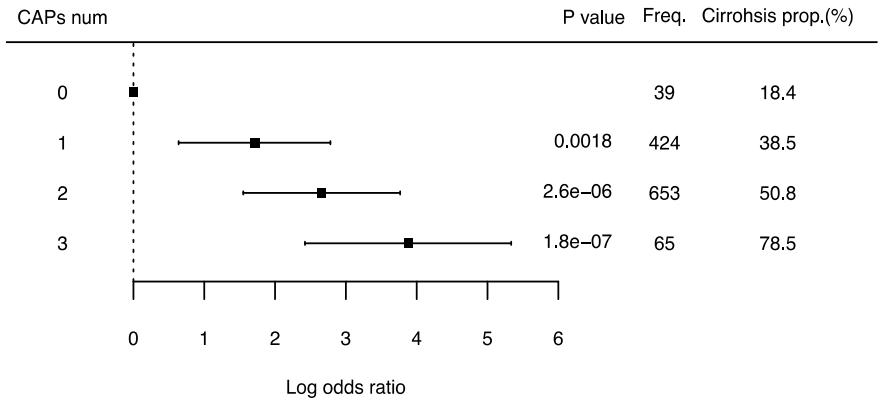
Cumulative effect of gt1a CAPs on cirrhosis

When patients were stratified by the number of gt1a CAPs present, a strong positive association was observed between CAP count and cirrhosis status (logOR = 1.11, $P = 4.95 \times 10^{-10}$). Cirrhosis prevalence rose progressively with CAP burden: 18.4% for patients carrying no CAPs, 38.5% for patients with one CAP, 50.8% with two, and 78.5% with all three CAPs (Figure 3.12A). These findings mirror the pattern observed in genotype 3a, which supports a stepwise increase in cirrhosis prevalence with the accumulation of CAPs in viral sequences.

To evaluate the cumulative impact of these variants, eight possible combinations were generated based on the presence or absence of the three gt1a CAPs (NS3 938 S, NS5A 2036 A, and NS5A 2321 S) across patient viral sequences (Table

3.7). Only combinations observed in more than 20 individuals were analysed to ensure stable estimates. The highest cirrhosis prevalence (78.5%) was seen in patients carrying all three CAP residues, while the lowest prevalence (18.4%) was observed in those with none. All CAP-containing combinations were associated with higher cirrhosis rates compared with the no-CAP reference group (Figure 3.12B).

A



B

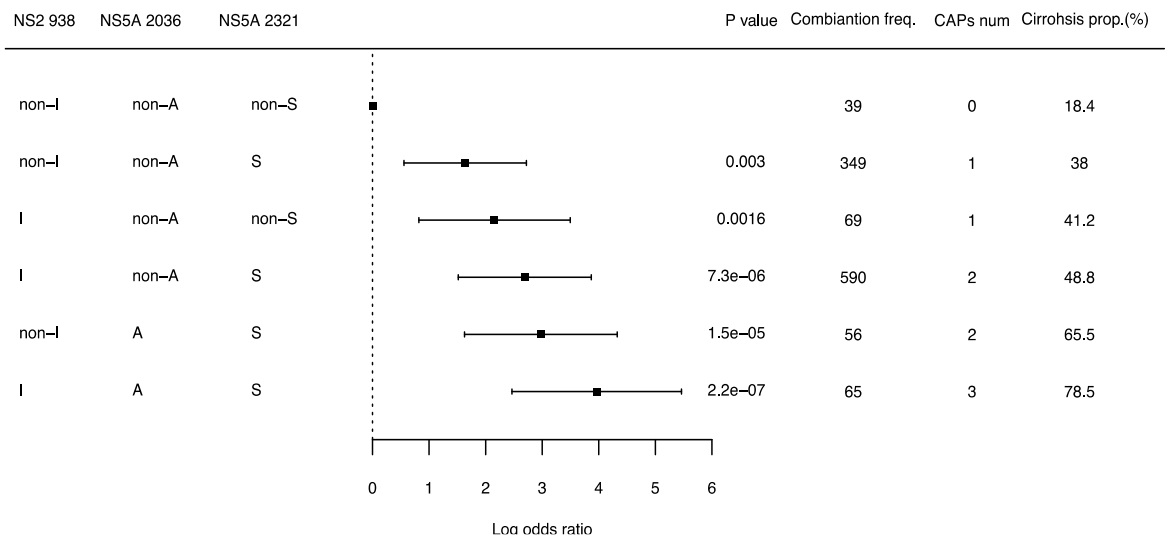


Figure 3.12: Covariation of gt1a CAPs and their impact on cirrhosis status.

(A) Cirrhosis proportion for patients infected with HCV gt1a with different numbers of CAPs in baseline sequences. The total number of CAPs present at each group is shown on the left of the table. The p-values (logistic regression), frequency of each group and the cirrhosis rate for each group are shown on the right.
(B) Covariation between the three gt1a CAPs and their associations with cirrhosis. Combinations with a frequency greater than 20 involved in tests are listed in the forest plot on the left. The p-values (logistic regression), frequency of each tested combination, number of CAPs carried by each combination and the cirrhosis rate for each combination are shown on the right.
The black squares and the lines indicate the estimated effect size (log (odds ratio)) and its 95% confidence interval estimated from logistic regression.

CAP combinations (n = 8)			**Regression outcomes**				**Combination characteristics**		
NS2 938	NS5A 2036	NS5A 2321	logOR	SE	z	p value	Combination freq.*	CAPs num	Cirrhosis prop. (%)
non-I	non-A	non-S	0	0	0	NA	39	0	18.42%
I	non-A	non-S	2.157	0.682	3.161	0.002	69	1	41.18%
non-I	non-A	S	1.638	0.552	2.966	0.003	349	1	38.04%
non-I	A	non-S	0.494	1.167	0.423	0.672	6	1	33.33%
non-I	A	S	2.977	0.689	4.323	1.54E-05	56	2	65.45%
I	non-A	S	2.691	0.600	4.483	7.33E-06	590	2	48.80%
I	A	non-S	16.491	528.270	0.031	0.975	7	2	100.00%
I	A	S	3.963	0.765	5.183	2.19E-07	65	3	78.46%

* *Results for combinations with low frequency (n < 20) are shown for descriptive purposes only and should be interpreted cautiously.*

Table 3.7: Associations between gt1a CAP combinations and cirrhosis status.

The table presents the total 8 unique combinations based on the specific presence or absence of the three individual CAPs in patients' sequences. The logistic regression outcomes (p values, effect sizes, standard errors, et al) and characteristics of these combinations were also shown in the table.

3.4.5 Effects of CAPs on hepatocellular carcinoma (HCC)

Effects of gt3a CAPs on HCC

The association between individual CAPs and the presence of hepatocellular carcinoma (HCC) was first examined, using logistic regression models that included sex, age, the first three viral PCs, $\log_{10}$ viral load, sequencing center, and cirrhosis status. Cirrhosis included as a covariate was essential given that nearly all HCC cases occur in the context of cirrhosis. Among the CAPs, only the variant at NS5A position 2414 (non-S) showed a borderline association with increased HCC rate ($P = 0.05$; Figure 3.13A), while no significant associations were observed for the other CAPs. Following that, the total number of CAPs carried by a patient was then tested for association with HCC status. Although not statistically significant, there was a trend towards increased odds of HCC with a greater number of CAPs (logOR = 0.179, $P = 0.062$). HCC prevalence increased with CAP burden: 5.4% in patients with no CAPs, 8.8% with one CAP, 13.6% with two, and 21.7% with three. Among patients carrying four or five CAPs, HCC prevalence was 16.7% (Figure 3.13B).

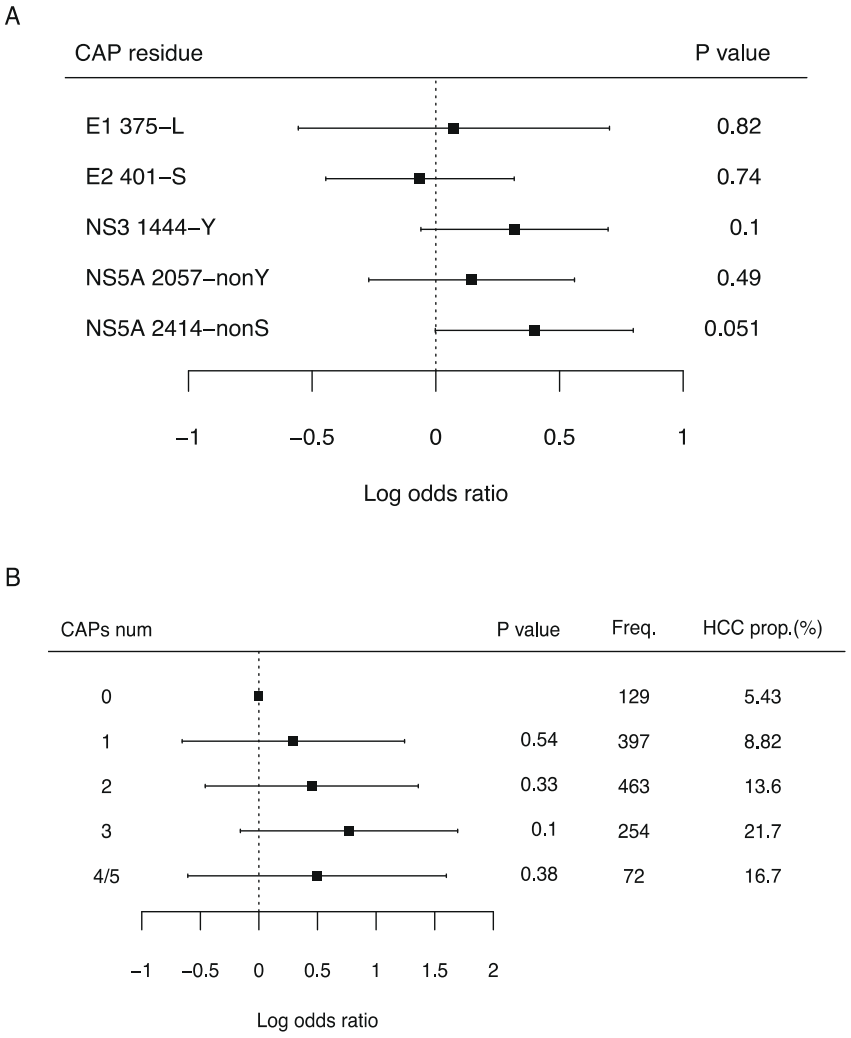


Figure 3.13: Association between gt3a CAPs and hepatocellular carcinoma (HCC) status.

(A) Association observed from individual gt3a CAP. Each gt3a CAP residue is listed in the table on the left. The p-values calculated by logistic regression are shown on the right.

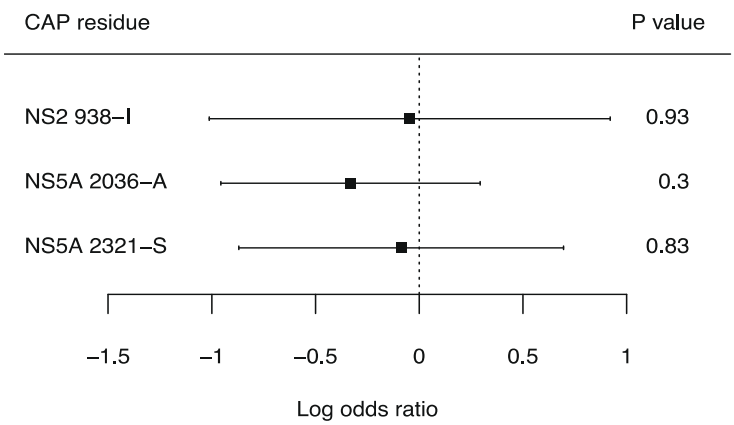
(B) Association between different number of CAPs and HCC status. The total number of CAPs present at each group is shown on the left of the table. The p-values (logistic regression), frequency of each group and the HCC rate for each group are shown on the right.

The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

Effects of gt1a CAPs on HCC

The association between individual gt1a CAP residues and hepatocellular carcinoma (HCC) status was also assessed using logistic regression. The model included sex, age, the first five viral principal components, $\log_{10}$ viral load, sequencing center, and cirrhosis status to account for potential confounders. None of the gt1a CAP residues showed a significant association with HCC (Figure 3.14). In the cumulative analysis, the presence of different numbers of gt1a CAPs was also not significantly associated with HCC (logOR = -0.19, *P* = 0.37).

A



B

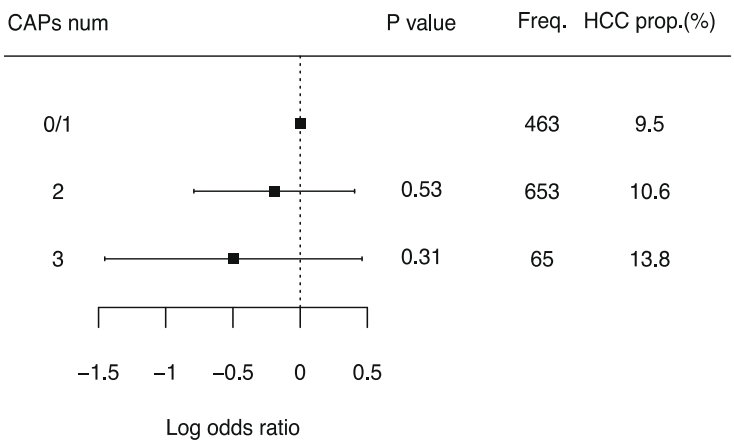


Figure 3.14: Association between gt1a CAPs and hepatocellular carcinoma (HCC) status.

Each gt1a CAP residue is listed in the table on the left. The p-values calculated by logistic regression are shown on the right. The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

3.4.6 Effects of CAPs on viral load (VL)

Effects of gt3a CAPs on log_{10} VL

Among patients infected with HCV gt3a, the impact of individual CAPs on viral load was assessed using linear regression. The model included sex, age, the first three viral principal components, sequencing centre and cirrhosis status. As the *IFNL4* gene is known to influence HCV viral load, it was also incorporated as a covariate to account for host genetic effects. Notably, incorporating *IFNL4* SNP genotype data reduced the sample size from 1,347 to 1,028 patients.

Of all the CAPs, only the presence of non-serine at position NS5A 2414 was significantly associated with higher $\log_{10}$ viral load ($P = 3.1 \times 10^{-5}$; Figure 3.15E). The cumulative effect of CAPs was then investigated by examining the association between the total number of CAPs in each patient's viral sequence and viral load using the same covariates. Each additional CAP was associated with a 12% increase in mean viral load ($\beta = 0.051$, $P = 0.025$), indicating a significant dose-dependent effect. Patients with no CAPs had a mean viral load of 1,250,423 IU/mL, rising to 1,655,483 with one CAP, 1,735,985 with two CAPs, 1,759,976 with three CAPs, and 2,228,183 IU/mL in the combined four/five-CAP group (the four/five-

CAP groups were combined due to limited sample size). This suggested that CAPs may have an additive effect on viral replication or persistence, potentially contributing to the observed disease severity.

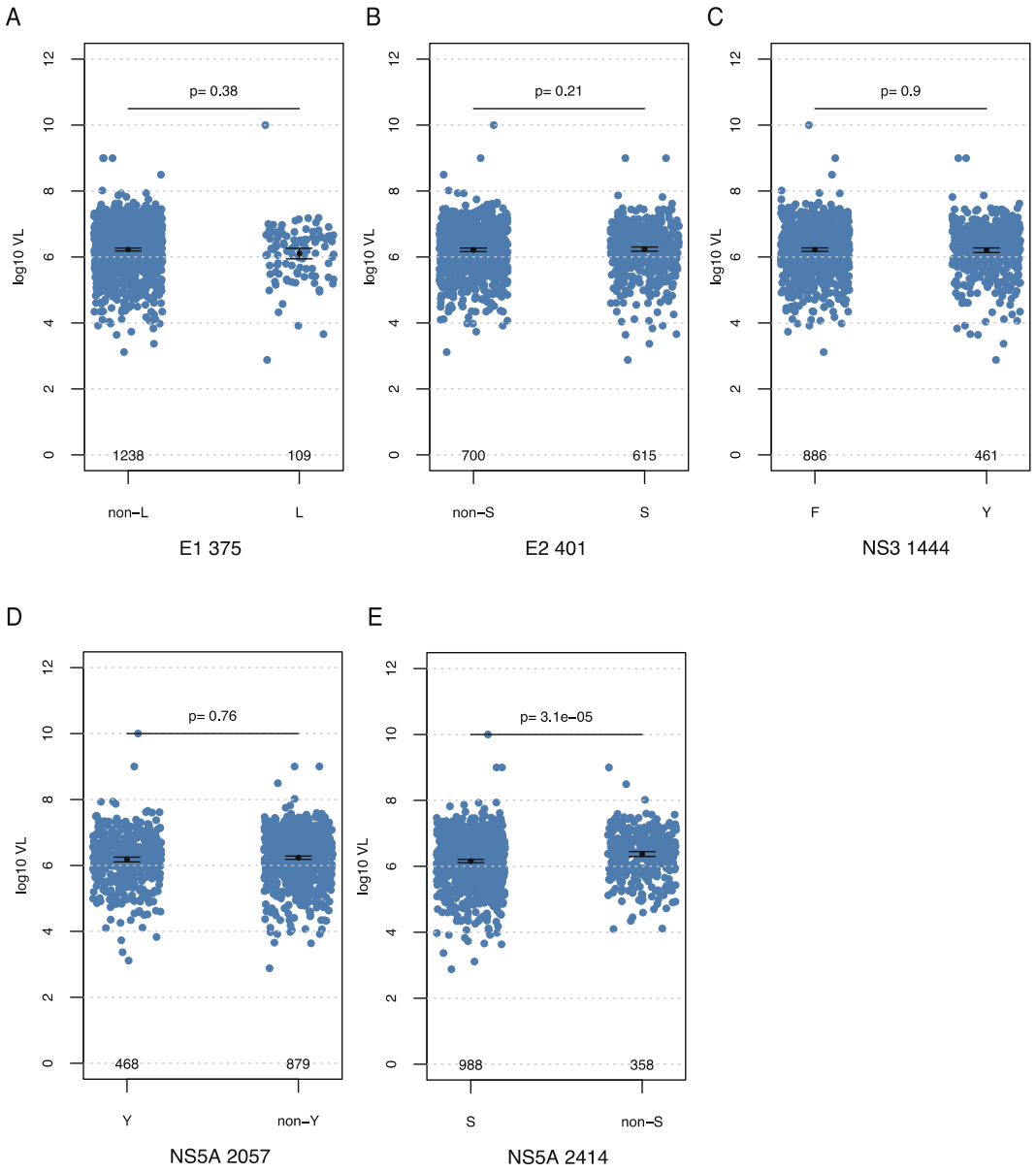


Figure 3.15: Association between gt3a CAPs and log_{10} viral load.
Figures (A) to (E) show the distribution of log_{10} viral load across patients at five reported sites, stratified by CAP and non-CAP groups. The p-values are calculated by linear regression. The mean change in log_{10} viral load is shown as a black dot and the lines indicate its 95% confidence interval.

Effects of gt1a CAPs on log$_{10}$ VL.

The impact of gt1a CAPs on viral replication was then assessed using a similar analytical pattern, incorporating *IFNL4* SNP genotypes along with the previously mentioned covariates. Including *IFNL4* genotypes reduced the sample size from 1,181 to 535 patients. Linear regression identified a significant association between higher viral load and the presence of serine at NS5A 2321 (P = 0.006, Figure 3.16C). The cumulative effect of CAPs on viral load indicated that each additional gt1a CAP was associated with an approximate 30% increase in mean viral load (β = 0.113, P = 0.013), suggesting a dose-dependent relationship. Compared with patients with no CAPs (mean viral load = 1,133,283 IU/mL), those with a single CAP had higher mean levels but the difference was not statistically significant (mean viral load = 1,661,058 IU/mL, β = 0.241, P = 0.25). Patients with two CAPs showed ~1.5-fold higher mean viral load (1,677,366 IU/mL, β = 0.417, P = 0.050), and those with three CAPs had ~2-fold higher levels (2,279,262 IU/mL, β = 0.496, P = 0.041). Together, these findings indicated a stepwise increase in viral load corresponding to the number of gt1a CAPs present in the viral sequence.

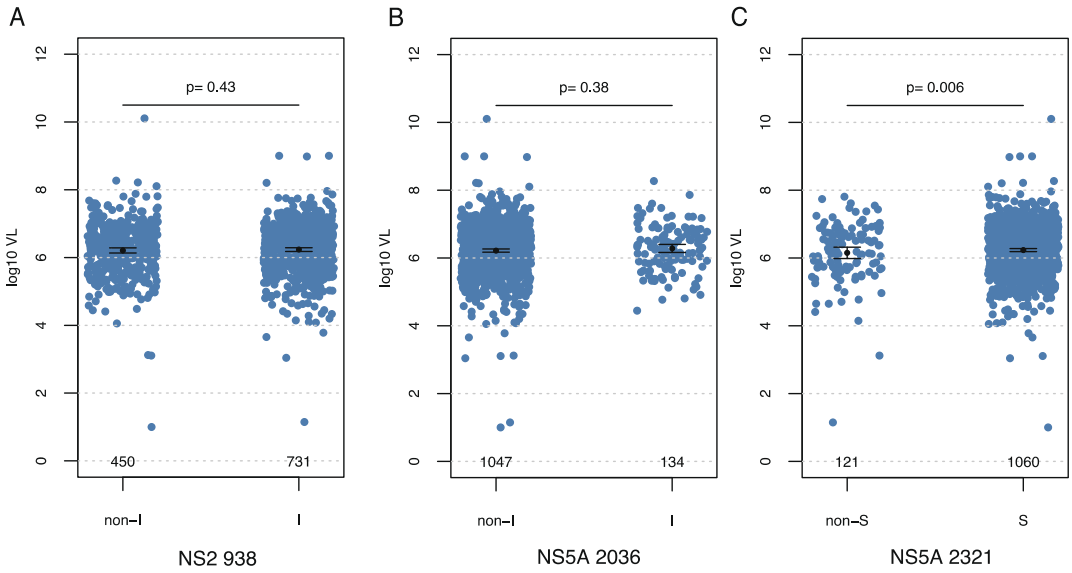


Figure 3.16: Association between gt1a CAPs and log_{10} viral load.
Figures (A) to (C) showed the distribution of log_{10} viral load across patients at three reported sites, stratified by CAP and non-CAP groups. The p-values are calculated by linear regression. The mean change in log_{10} viral load is shown as a black dot and the lines indicate its 95% confidence interval.

3.4.7 Interplay between host *IFNL4* gene and CAPs

From this section onward, the analysis focused on interactions between specific host genetic factors and the identified CAPs. The first factor examined was the *IFNL4* gene, a well-established host genetic determinant of HCV viral clearance and treatment response. The importance of this gene has been described in detail in chapter 2, sections 2.2 and 2.5. Host-virus interactions can shape viral evolution and influence disease progression, making it important to assess whether *IFNL4* genotypes are associated with the presence of CAP residues. Such analyses provide insights into whether CAPs arise independently of host background or are, at least in part, driven by host genetic pressures.

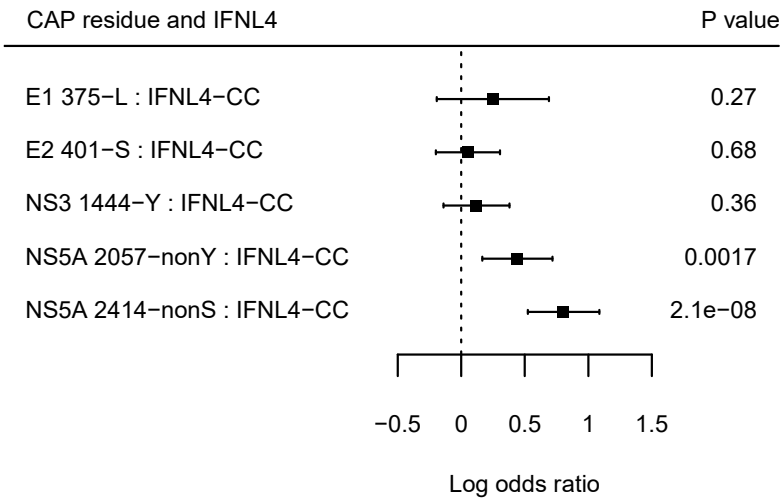
IFNL4 gene and gt3a CAPs

The relationship between gt3a CAP amino acid residues and the *IFNL4* SNP rs12979860 genotypes (CC vs. non-CC) was shown in Figure 3.18A. The strongest association was observed at NS5A position 2414 in gt3a, where patients with the CC genotype had a significantly higher prevalence of the non-serine (non-S, CAP residue) in their viral sequences ($P = 2.1 \times 10^{-8}$). Additionally, the presence of non-

tyrosine (non-Y, CAP residue) at NS5A position 2057 was more frequent in patients with CC genotypes (P = 0.0017). The *IFNL4* SNP did not show any association with the remaining CAP sites.

To determine whether these two CAPs mediated the effect of host *IFNL4* genetic variation on cirrhosis, conditional logistic regression including *IFNL4* genotype as a covariate was performed (Figure 3.17B). While the effect sizes of NS5A 2414 non-serine (original logOR = 0.57, $P = 1.62 \times 10^{-4}$; adjusted logOR = 0.52, $P = 3.06 \times 10^{-3}$) and NS5A 2057 non-tyrosine (original logOR = 0.54, $P = 1.4 \times 10^{-4}$; adjusted logOR = 0.46, $P = 3.93 \times 10^{-3}$) on cirrhosis were slightly reduced after conditioning on *IFNL4* genotype, they remained statistically significant. Independently, *IFNL4* CC genotype was also associated with cirrhosis in both models ($P_{CC \text{ in } 2057 \text{ model}}$ = 0.0015 and $P_{CC \text{ in } 2414 \text{ model}}$ = 0.0046, respectively) compared with non-CC.

A



B

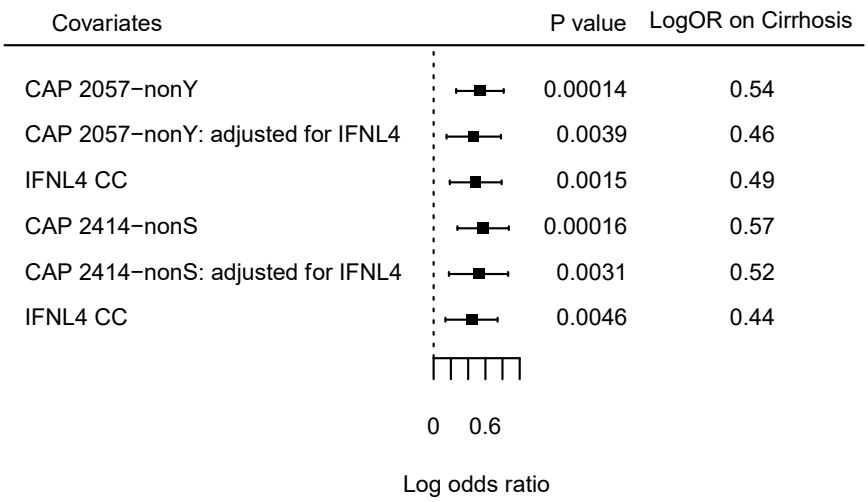


Figure 3.17: Association between *IFNL4* genotype and gt3a CAPs.

(A) Impact of *IFNL4* genotype-CC on each gt3a CAP residue (E1 375-L, E2 401-S, NS3 1444-Y, NS5A 2057-nonY and NS5A 2414-nonS). In logistic regression, the response variable was each gt3a CAP position (non-CAP residue as the reference group) and *IFNL4* genotypes were explanatory variable (non-CC as the reference group), including the first three viral PCs as covariates.

(B) Association between cirrhosis and site 2057 and site 2414 with and without adjustment for *IFNL4* genotypes, respectively.

The black squares and the lines indicate the estimated effect size (log (odds ratio)) and its 95% confidence interval estimated from logistic regression. The p-values calculated by logistic regression are shown on the right of the table.

Interaction between IFNL4 gene and gt3a CAPs

Further interaction analyses revealed whether the influence of *IFNL4* genotype on cirrhosis was dependent on the amino acid present at these two CAP sites (site 2057 and site 2414). The interaction term from the interaction analysis was not statistically significant ($P = 0.33$) between *IFNL4* gene and site 2057, indicating no strong evidence that the effect of *IFNL4* genotypes (CC vs. non-CC) on cirrhosis

status varied according to the presence or absence of the CAP residue (non-tyrosine) at position 2057. Nevertheless, when patients were stratified by the interaction variable combining *IFNL4* genotypes (CC vs. non-CC) and CAP status at position 2057 (Table 3.8), a pattern merged. Among those with non-tyrosine (CAP residue) at NS5A 2057, the *IFNL4* CC was associated with higher cirrhosis prevalence (66.1% vs. 56.3% in non-CC; P = 0.0025; Table 3.8). In contrast, no significant association between *IFNL4* genotype and cirrhosis was observed in patients carrying tyrosine (non-CAP residue) at this position (P = 0.27). Taken together, these findings indicated that while the formal interaction test did not reach significance, stratified analyses suggested a possible trend in which *IFNL4* genotype may have a stronger association with cirrhosis in patients carrying the CAP residue at NS5A 2057. This trend, although not statistically confirmed as an interaction, may point to a biologically relevant host-virus interplay that requires further investigation in larger cohorts.

***IFNL4* genotypes**	**Residues at site2057**	**Non-cirrhotic (n)**	**Cirrhotic (n)**	**Cirrhosis Prop. (%)**
CC	non-tyrosine (CAP)	101	197	66.11%
non-CC	non-tyrosine (CAP)	164	211	56.27%
CC	tyrosine	56	63	52.94%
non-CC	tyrosine	107	113	51.36%

Table 3.8: Interaction analysis of NS5A 2057 and *IFNL4* genotype.
The distribution of cirrhotic and non-cirrhotic individuals across the four groups defined by the interaction of *IFNL4* genotype (CC vs. non-CC) and residues at site 2057 (tyrosine vs. non-tyrosine).

A similar analytical approach was performed for the interaction between *IFNL4* genotypes and residues at NS5A 2414. Likewise, the interaction term was not statistically significant (P = 0.09), indicating no strong evidence that the effect of *IFNL4* genotypes on cirrhosis status differed by residue at this site. However, when patients were stratified by the combined *IFNL4* genotypes (CC vs. non-CC) and the presence or absence of a CAP residue (non-serine) at NS5A 2414, a different trend was observed. Within individuals with the non-serine residue at position 2414, cirrhosis prevalence was similar between genotypes: 68.4% in the CC group and 69.6% in the non-CC group (Table 3.9). In contrast, among individuals carrying a non-CAP (serine) at NS5A 2414, the CC genotype was associated with higher cirrhotic prevalence (58.5% vs. 50.4% in non-CC; P = 0.0014). Although the formal interaction test did not reach significance, the stratified analyses indicate a potential trend where the effect of *IFNL4* genotype on cirrhosis prevalence may depend on the viral amino acid present at NS5A 2414. This trend, while not statistically confirmed as an interaction, points to a possible host-virus interplay that could be explored further in larger cohorts.

***IFNL4* genotypes**	**Residues at site2414**	**Non-cirrhotic (n)**	**Cirrhotic (n)**	**Cirrhosis Prop. (%)**
CC	non-serine (CAP)	48	104	68.43%
non-CC	non-serine (CAP)	38	87	69.60%
CC	serine	109	156	58.87%
non-CC	serine	233	237	50.43%

Table 3.9: Interaction analysis of NS5A 2414 and *IFNL4* genotype.

The distribution of cirrhotic and non-cirrhotic individuals across the four groups defined by the interaction of *IFNL4* genotype (CC vs. non-CC) and residues at site 2414 (serine vs. non-serine).

IFNL4 gene and gt1a CAPs

The relationship between gt1a CAP amino acid residues and the *IFNL4* gene was also examined in the subset of patients with available host genetic data (n = 535) using logistic regression, incorporating the first five viral PCs as covariates. No association was observed between gt1a CAPs and *IFNL4* genotype (Figure 3.18).

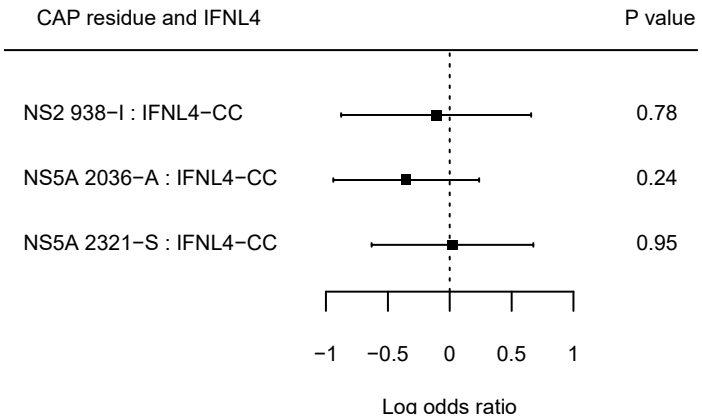


Figure 3.18: Association between *IFNL4* genotype and gt1a CAPs.
Impact of *IFNL4* genotype-CC on each gt1a CAP residue (NS2 938-I, NS5A 2036-A, NS5A 2321-S). In logistic regression, the response variable was each gt1a CAP position (non-CAP residue as the reference group) and *IFNL4* genotypes were explanatory variable (non-CC as the reference group), including the first five viral PCs as covariates.

3.4.8 Interplay between host HLA alleles and CAPs

Of the many factors previously implicated with the clearance or chronicity of viral HCV and subsequent disease progression, the host immunogenetic factors affecting components of the innate and adaptive immune system that modulate host-viral interactions are key players [208-210]. Numerous studies have demonstrated associations between the human leucocyte antigen (HLA) system and both beneficial and deleterious effects on viral infections, including HIV and HCV [211, 212]. In this study, the analyses were extended by examining, for the first time, the interactions between HLA variation and CAPs to provide insight into how host immune recognition might shape viral sequence diversity and contribute to liver disease risk.

HLA alleles and gt3a CAPs

To explore the potential role of HLA alleles in the CAP-cirrhosis associations, the logistic regression was used to test each CAP against 148 two-field HLA alleles, adjusting for viral and host PCs, age, sex, ethnicity and sequencing centre. Only the presence and absence of tyrosine (Y) at gt3a NS3 1444 showed a significant association with HLA alleles at a 5% FDR (Table 3.10), with the strongest link observed for HLA-A*01:01 ($P = 1.2 \times 10^{-26}$). This site has previously been reported as an HLA-associated escape mutation [146]. To assess whether the impact of the remaining of 31 significant associations at site1444 were driven by HLA-A*01:01, a conditional analysis was performed including all 32 HLA alleles alongside the covariates described above (Figure 3.19). The results indicated that most of the associations were attributable to linkage with HLA-A*01:01, except for HLA-

DRB1*04:04, which showed an independent positive associated with site1444. Conditional analysis adjusting for HLA-A*01:01 or HLA-DRB1*04:04 showed only marginal changes to the association between NS3 1444Y and cirrhosis (Figure 3.20). Furthermore, neither allele itself was directly associated with cirrhosis status. (Figure 3.20)

HLA	P values	Log OR	SE	FDR
HLA-A*01:01	1.20E-26	-3.5505417	0.33230059	1.56E-24
HLA-G*01:06	4.17E-13	-2.7169109	0.37475169	2.43E-11
HLA-C*07:01	5.60E-13	-1.3907736	0.19290053	2.43E-11
HLA-B*08:01	2.42E-10	-1.1568602	0.18269989	7.86E-09
HLA-E*01:01	9.68E-09	-0.5765268	0.10050646	2.52E-07
HLA-DRB1*03:147	9.66E-08	-0.9074773	0.17015963	2.09E-06
HLA-DRB3*01:01	1.17E-07	-1.0381608	0.19596549	2.18E-06
HLA-F*01:01	2.23E-07	-0.6408825	0.12374836	3.63E-06
HLA-F*01:03	2.52E-07	0.64689518	0.12546423	3.64E-06
HLA-E*01:03	3.22E-07	0.51887835	0.10154268	4.19E-06
HLA-DQA1*05:01	8.82E-07	-0.7687942	0.15637596	1.04E-05
HLA-DQB1*02:01	2.40E-06	-0.7408384	0.15707681	2.60E-05
HLA-G*01:01	1.14E-05	0.53911363	0.12285241	0.0001142
HLA-A*03:01	1.62E-05	0.6052233	0.14036246	0.00014542
HLA-C*04:01	1.68E-05	0.6530172	0.15172854	0.00014542
HLA-DRB3*99:99	1.89E-05	0.45755833	0.106977	0.00015374
HLA-DQB1*02:02	4.36E-05	0.65249546	0.15963747	0.00033352
HLA-B*35:01	4.69E-05	0.9095914	0.22344795	0.00033838
HLA-B*44:03	0.00020627	0.74010083	0.199423	0.00141078
HLA-DRB4*01:01	0.00030542	0.7823296	0.21667235	0.00198449
HLA-C*06:02	0.00044334	-0.5753918	0.16379657	0.00274343
HLA-DRB4*99:99	0.00046872	-0.4611714	0.13183763	0.00276863
HLA-A*11:01	0.00066695	0.56927631	0.16729565	0.00376829

HLA-DRB1*04:04	0.0012011	0.87611363	0.27052081	0.00650349
HLA-E*01:06	0.00194356	1.10965468	0.35810047	0.01010264
HLA-C*16:01	0.00229127	0.84098386	0.27576627	0.01145197
HLA-B*37:01	0.00300347	-1.324731	0.44643075	0.01445562
HLA-A*32:01	0.00336039	0.73846991	0.25180633	0.01559587
HLA-A*29:02	0.00950811	0.72383296	0.27912484	0.04117637
HLA-DRA*01:02	0.00982273	-0.2687668	0.10409217	0.04117637
HLA-DRA*01:01	0.00982273	0.26876684	0.10409217	0.04117637
HLA-B*57:01	0.01019678	-0.6220752	0.24213894	0.04140863

Table 3.10: Top associations (FDR < 5%) between HLA alleles and the residue present at NS3 position 1444.

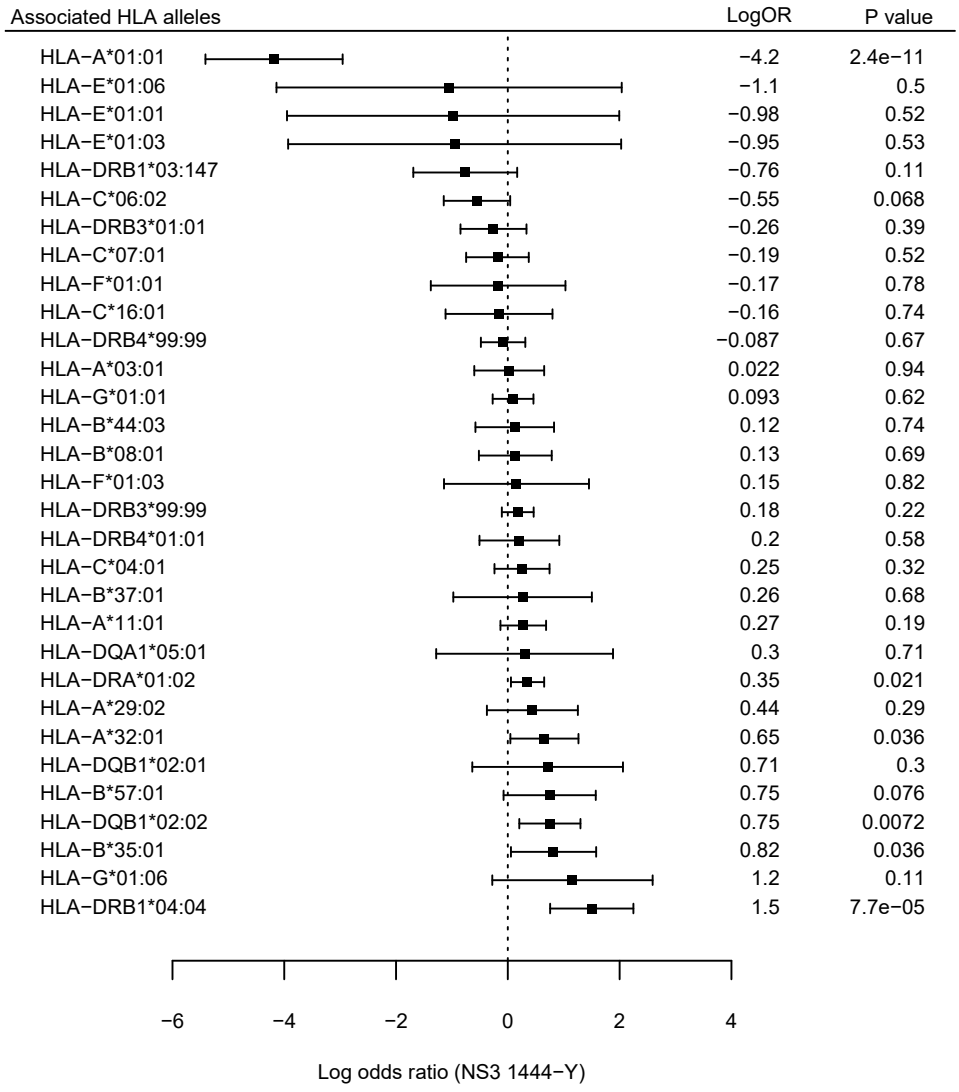


Figure 3.19: Conditional association analysis of HLA alleles with site1444. Logistic regression models were used to test associations between HLA alleles and the presence of tyrosine (Y) at NS3 site1444, adjusting for all 32 HLA alleles and other covariates. The forest plot indicated whether the impact of HLA alleles on site1444 was independent. Except for HLA-DRB1*04:04, the effects of other HLA alleles were largely captured by the allele A*01:01. The black squares and the lines indicate the estimated effect size (log (odds ratio)) on site1444 and its 95% confidence interval estimated from logistic regression. The p-values calculated by logistic regression are shown on the right of the table.

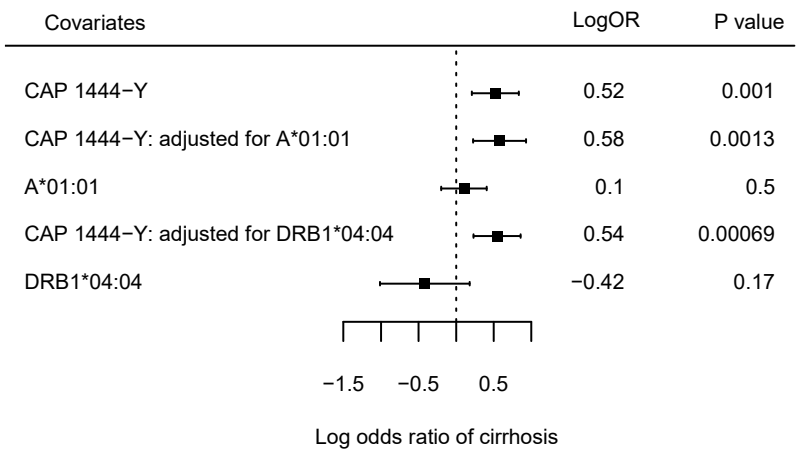


Figure 3.20: Association between cirrhosis status and residues at site1444 with or without adjusting for HLA A*01:01 or DRB1*04:04.

The forest plot indicates the estimated associations between the residue tyrosine (Y) at site 1444 and cirrhosis, both unadjusted and adjusted for the presence of HLA-A*01:01 or HLA-DRB1*04:04. The associations of the two HLA alleles with cirrhosis are also displayed. The black squares and the lines indicate the estimated effect size (log (odds ratio)) of cirrhosis and its 95% confidence interval estimated from logistic regression. All estimates and p-values were obtained from logistic regression.

Interaction between HLA alleles and gt3a CAPs

The strong associations between viral site 1444 and both HLA-A*01:01 (P = 1.2 × 10^{-26}) and HLA-DRB1*04:04 (P =0.001) suggested that site 1444 was under host genetic control and that its impact on cirrhosis may reflect not only a direct viral effect alone but also HLA-mediated immune pressure. After stratifying the patients by the interaction variable combining HLA-A*01:01 (presence vs. absence) and the residues at 1444 (tyrosine vs. phenylalanine), Table 3.11 showed differences in

cirrhosis prevalence across subgroups. Among individuals lacking HLA-A*01:01, those with tyrosine (CAP residue) at site 1444 had a higher cirrhosis prevalence (66.1%) compared to those with phenylalanine (50.0%). Similarly, in individuals carrying at least one copy of HLA-A*01:01, cirrhosis was more frequent in those with tyrosine (100%) than phenylalanine (55.2%). Notably, the subgroup with both HLA-A*01:01 and tyrosine at site 1444 included only 8 individuals, all of whom were cirrhotic. However, the formal interaction test did not reveal a significant interaction between HLA-A*01:01 copy number and cirrhosis in either phenylalanine or tyrosine carriers ($P_{\text{site1444-Y : A*01:01}}$ = 0.99, logOR = 14.5, SE = 372.3). This lack of significance may reflect limited statistical power due to the absence of non-cirrhotic patients carrying both HLA-A*01:01 and tyrosine at NS3 1444 (0/8 non-cirrhotic with A*01:01 vs. 119/351 without A*01:01), restricting the ability to detect effects in this subgroup.

HLA-A*01:01	Residues at site1444	Non-cirrhotic (n)	Cirrhotic (n)	Cirrhosis Prop. (%)
absence (0 copy)	phenylalanine	158	158	50.00%
absence (0 copy)	tyrosine (CAP)	119	232	66.10%
presence (1/2 copy)	phenylalanine	151	186	55.19%
presence (1/2 copy)	tyrosine (CAP)	0	8	100%

Table 3.11: Interaction analysis of NS3 1444 and HLA-A*01:01.
The distribution of cirrhotic and non-cirrhotic individuals across the four groups defined by the interaction of HLA-A*01:01 (absence vs. presence) and residues at site 1444 (phenylalanine vs. tyrosine).

A similar interaction analysis was performed for position 1444 and HLA-DRB1*04:04, and the interaction term did not show any association ($P_{\text{site1444-Y:DRB1*04:04}} = 0.24$). As shown in Table 3.12, the presence of a CAP residue (tyrosine) at site 1444 was consistently associated with a higher prevalence of cirrhosis, regardless of DRB1*04:04 status. Among individuals without the DRB1*04:04 allele, cirrhosis prevalence was higher in those with tyrosine (66.8%) compared to those with phenylalanine (53.4%). A similar pattern was observed among individuals carrying HLA-DRB1*04:04, where cirrhosis was more common in tyrosine carriers (67.9%) than in phenylalanine carriers (37.0%). Notably, the proportion of cirrhotic individuals was markedly reduced in those carrying DRB1*04:04 and phenylalanine at site 1444, suggesting a possible protective effect in this subgroup. However, the absolute number of individuals in the HLA-DRB1*04:04-positive groups was relatively small, so these findings should be interpreted with caution.

HLA-DRB1*04:04	Residues at site1444	Non-cirrhotic (n)	Cirrhotic (n)	Cirrhosis Prop. (%)
absence (0 copy)	phenylalanine	292	334	53.35%
absence (0 copy)	tyrosine (CAP)	110	221	66.77%
presence (1/2 copy)	phenylalanine	17	10	37.04%
presence (1/2 copy)	tyrosine (CAP)	9	19	67.86%

Table 3.12: Interaction analysis of NS3 1444 and HLA DRB1*04:04.
The distribution of cirrhotic and non-cirrhotic individuals across the four groups defined by the interaction of HLA DRB1*04:04 (absence vs. presence) and residues at site 1444 (phenylalanine vs. tyrosine).

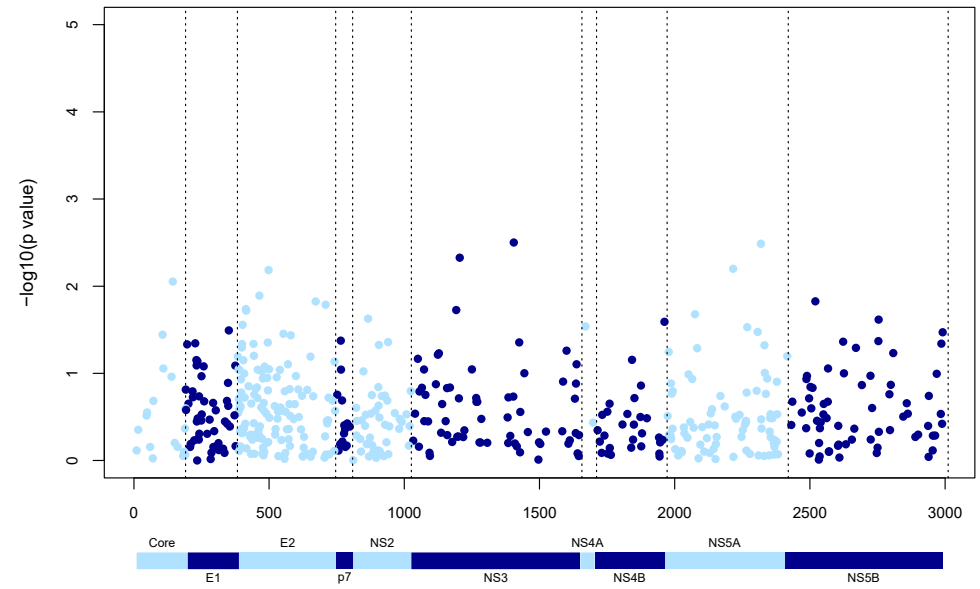
HLA alleles and gt1a CAPs

The association between gt1a CAP amino acid residues and HLA alleles was also assessed in the subset of patients with available host genetic data (n = 535) by using logistic regression, incorporating viral and host principal components, age, sex, ethnicity, sequencing centre as covariates in the model. No association was observed between gt1a CAPs and HLA alleles at a 5% FDR.

3.4.9 Viral Genome-Wide Association Study (V-GWAS) for HCC

To investigate viral determinants associated with HCC status, associations between amino acid variants across the entire HCV polyprotein and HCC status were examined. Phylogenetic analysis using treeBreaker did not indicate any HCV lineages associated with higher or lower rates of HCC in either genotype (Bayes factor for gt3a = 1; Bayes factor for gt1a = 1.03), suggesting a random distribution of HCC across the phylogenetic tree (Supplementary Figures B.3 and B.4). In the V-GWAS for HCC, cirrhosis status was included as an additional covariate along with those previously used for each genotype. At a 10% FDR, no viral polymorphisms were significantly associated with the presence or absence of HCC in either HCV gt3a or gt1a infections (Figure 3.21 A and B).

A



B

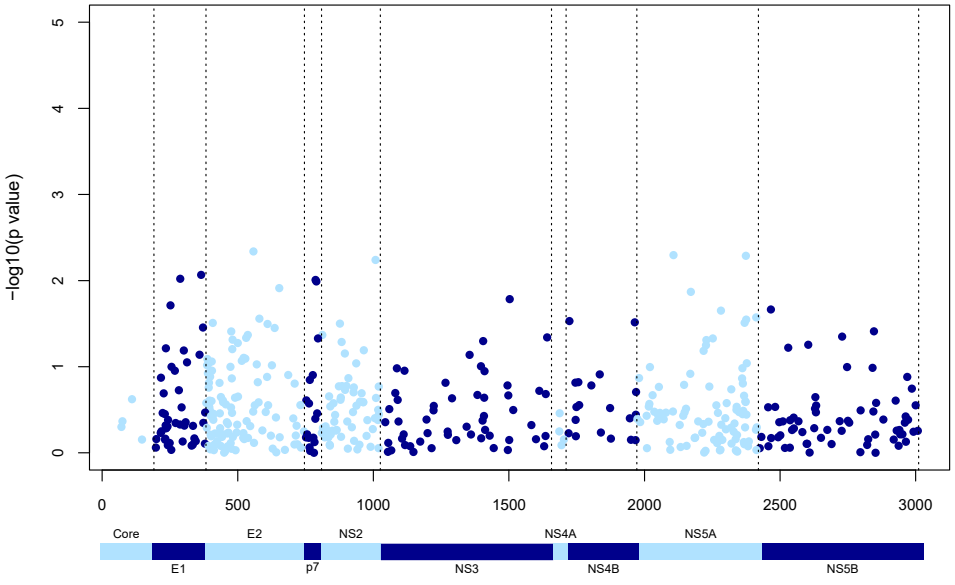


Figure 3.21: Association between HCV amino acids and HCC status.
Manhattan plot of association tests between HCC status and HCV gt3a (A) and gt1a (B) amino acids, respectively. For each site only the most associated p-value is plotted. And no sites are associated with the presence or absence of HCC at a 10% FDR.

3.5 Discussion

This is the first systematic investigation of the viral genetic determinants of HCV-induced liver disease within a large, genotype-specific cohort of patients infected with HCV gt3a and gt1a. The present study leveraged next-generation sequencing (NGS) technologies to generate full-length HCV genomes for 1347 gt3a-infected patients and 1181 gt1a-infected patients, enabling an unbiased, genome-wide association approach. This approach enabled a more comprehensive investigation of the relationship between liver disease progression and viral amino acid variation. It uncovered novel genotype-specific determinants of HCV pathogenicity that had not been detected in prior region-limited studies.

In this study, five viral amino acid polymorphisms within gt3a polyprotein (E1 375 leucine, E2 401 serine, NS3 1444 tyrosine, NS5A 2057 non-tyrosine and NS5A 2414 asparagine) and three within gt1a polyprotein (NS2 938 isoleucine, NS5A 2036 alanine, NS5A 2321 serine) showed significant associations with the presence of cirrhosis and were subsequently classified as CAPs. These associations remained significant after adjusting for age, sex, ethnicity, viral load, viral population structure, sequencing center and lineage effects. Moreover, a stepwise effect was observed when investigating CAP covariation: patients carrying more CAPs in their viral sequences had progressively higher rates of cirrhosis. For example, gt3a-infected patients with four or five CAPs had cirrhosis rates exceeding 70%, compared to 32% among those with none. This pattern suggests that multiple viral adaptations can additively worsen fibrotic outcomes.

Together, these findings identify specific viral genetic determinants of cirrhosis status that are distinct from known host risk factors.

When investigating the impact of these CAPs on additional phenotypes, the NS5A 2414 in gt3a virus was marginally associated with the occurrence of HCC (Figure 3.14A). Furthermore, among patients with fewer than four CAPs in their viral sequences, the rate of HCC appeared to rise incrementally with increasing CAP number, although this trend did not reach statistical significance (Figure 3.14B). Notably, the proportion of HCC cases among subjects with none of these gt3a CAPs was the lowest across all groups (5.43%). These findings suggest a potential cumulative effect of CAPs on HCC rate. While no single CAP strongly predicts HCC on its own, their combined presence may suggest an elevated risk of disease progression.

The presence of gt3a CAP NS5A 2414 non-serine and gt1a CAP NS5A 2321 serine showed significant associations with higher baseline HCV RNA levels (Figure 3.15 and 3.16). The amino acid variation at site 2414 has been previously reported to affect viral load and to modulate the impact of the *IFNL4* gene on viral replication [146]. These findings support a broader role for these variants in viral fitness. Linear regression between $\log_{10}$ viral load and CAPs count showed positive additive effects of CAPs on viral replication (P_{gt3a} = 0.025; P_{gt1a} = 0.013). These findings suggested that CAPs may not only mark advanced liver disease but also contribute directly to increased viral fitness, highlighting the potential role of persistent, high-level viral activity in promoting liver inflammation, fibrosis, and eventual cirrhotic progression. In this context, cirrhosis-associated polymorphisms

may function both as markers of advanced disease and as mediators of viral survival, promoting disease progression through prolonged immune stimulation and hepatocyte turnover.

Previously, the functional *IFNL4* genotype (non-CC) was reported to protect against hepatic inflammation and fibrosis in both infectious and non-infectious aetiologies [213]. This is in agreement with the findings in this work – patients with the *IFNL4* CC genotype showed a significantly higher prevalence of cirrhosis (Figure 3.17B). Moreover, the CC genotype was associated with gt3a CAP non-tyrosine at position 2057 and non-serine at position 2414, respectively (Figure 3.17A). The interplay between amino acid variation at NS5A 2057 and *IFNL4* genotype (Table 3.8) suggested that *IFNL4* CC genotype and the presence of the CAP (non-tyrosine) at site 2057 together conferred an increased prevalence of cirrhosis (Cirrhosis prop. in CC = 66.1%, in non-CC = 56.3%, P = 0.0025). In contrast, when tyrosine is present at site 2057, *IFNL4* genotype has little effect (Cirrhosis prop. in CC = 53.0%, in non-CC = 51.4%, P = 0.27). This implies that the protective effect of the non-CC *IFNL4* genotype may only manifest in the context of infection with a virus carrying the CAP residue at position 2057. At NS5A position 2414, the non-CC *IFNL4* genotype was significantly associated with a lower rate of cirrhosis among patients infected with viruses carrying non-CAP residue (serine) at this position (Cirrhosis prop. in CC = 58.9%, in non-CC = 50.4%, P = 0.0014). Previous work demonstrated that non-CC genotype was associated with reduced viral load, but only in patients whose viruses encoded a serine at position 2414 [146]. Consistent with this, this present study also found that serine at this site was associated with lower viral load (Figure 3.15E), *IFNL4* non-CC

genotype linked to lower viral load only among patients carrying serine at this position ($P_{\text{in non-serine}} = 0.61$; $P_{\text{in serine}} = 0.03$). Taken together, these findings suggested that the non-CC *IFNL4* genotype may play a protective role against progression to cirrhosis by modulating viral replication in a residue-dependent manner, specifically when serine is present at NS5A 2414.

This work also reported a significant negative association between HLA-A*01:01 and the CAP residue (Y) at position NS3 1444 (Table 3.10), consistent with previous findings [146, 214]. Prior studies showed that the HLA-A*01-restricted epitope NS3 1436-1444 (ATDALMTGY), located in the NS3 helicase, is a frequent CD8+ T cell target conserved across all genotypes [214]. Phenylalanine at position 1444 has been identified as a key escape mutation and is the predominant variant among HLA-A*01 positive subjects in European cohorts. This substitution impairs the CD8+ T cell responses, potentially through reduced binding to the MHC class I molecule [214, 215]. In the present study, the high-resolution allele HLA-A*01:01 was specifically associated with variation at NS3 1444, suggesting that it may contribute to disease progression by exerting selective pressure on viral evolution. Among HLA-A*01:01 carriers, 337 were infected with viruses carrying phenylalanine (non-CAP) at position NS3 1444. In contrast, only 8 individuals carried tyrosine (CAP) at this position, and all 8 were cirrhotic (Table 3.13A and Table 3.11). This distribution suggests phenylalanine at NS3 1444 in this study also represents an immune escape mutation under a strong immune selection pressure exerted by HLA-A*01:01, allowing the virus to evade A*01:01-restricted CD8+ T cell responses. In contrast, the viral peptide with tyrosine at this position may remain recognized by CD8+ T cells, potentially eliciting stronger adaptive

immune responses and increased hepatocyte killing. Hepatocyte death may subsequently lead to the release of damage-associated molecular patterns (DAMPs), which can directly activate HSCs [119, 120] – the predominant precursors of myofibroblasts in fibrotic liver [115].

While HLA-A*01:01 appears to exert robust selection pressure at this site, the evidence for HLA-DRB1*04:04-mediated selection is less compelling. Among carriers of DRB1*04:04, the nearly equal distribution of phenylalanine and tyrosine (Table 3.13B) suggests only a marginal selective advantage for the tyrosine variant. Consequently, this weak immune pressure alone may be insufficient to explain the observed association between the tyrosine residue (CAP) and increased cirrhosis prevalence. This discrepancy suggests that the impact of the site on cirrhosis may not be solely explained by immune pressure, but instead potentially involves altered viral fitness or other host genetic factors, rather than a simple immune escape mechanism.

A

pos 1444 / HLA-A*01:01	Phenylalanine	Tyrosine (CAP)
0 copy of A*01:01	316 (47.4%)	351 (52.6%)
1/2 copy of A*01:01	337 (97.7%)	8 (2.3%)

B

pos 1444 / HLA-DRB1*04:04	Phenylalanine	Tyrosine (CAP)
0 copy	626 (65.4%)	331 (34.6%)
1/2 copy	27 (49.1%)	28 (50.9%)

Table 3.13: The distribution of amino acid at position 1444 across the presence or absence of HLA alleles.

(A) HLA-A*01:01, (B) HLA-DRB1*04:04.

While these analyses demonstrate intriguing trends, the small number of individuals in several key subgroups limits the strength of the conclusions. For example, only eight individuals carried both HLA-A*01:01 and tyrosine at NS3 1444, and all of them were cirrhotic. At the same site, 28 individuals carried both HLA-DRB1*04:04 and tyrosine, of whom only nine were non-cirrhotic. Such small numbers greatly reduce statistical power and make effect estimates unstable. In addition, because no non-cirrhotic patients were observed in the subgroup carrying both A*01:01 and tyrosine, formal interaction tests could not adequately assess the relationship between allele-residue combinations and cirrhosis rate. These findings should therefore be interpreted with caution and regarded as exploratory, and they require validation in larger cohorts.

Chapter 4

Viral determinants of viral load in patients chronically infected with genotype 3a and 1a hepatitis C virus

4.1 Summary

Background and Aims: HCV viral load represents the number of virus particles in patient's serum. The measurement of viral load is essential for HCV infection diagnosis and guiding clinical treatment decisions. It is also closely related to the host immune response and reflects the balance between viral replication and immune response. This chapter reports novel viral amino acid variants that are associated with viral load in patients with HCV gt3a and gt1a infection from two large cohorts (gt3a, n = 1028; gt1a, n = 535).

Methods: Patients and sequences included in this study were the same as those described in chapter 3, originating from the STOP-HCV-1, the BOSON clinical trial and the HCV Research UK biobank. A viral genome-wide association study (V-GWAS) was used to identify the viral load associated amino acid polymorphisms across the full length of the HCV polyprotein. The impact of these polymorphisms on liver diseases, including cirrhosis and HCC, was also investigated. Furthermore, the interplay between host HLA molecules and the significant viral variants was examined to explore the potential mechanisms underlying differences in viral load.

Results: The V-GWAS identified five novel polymorphisms in HCV gt3a and six in gt1a that were significantly associated with higher viral load, which were termed High Viral Load associated Polymorphisms (HVLPs). Further analysis of these HVLPs in combination revealed an additive effect on differences in viral load: patients with more HVLPs had higher levels of viral load, with an increase in beta estimates correlating with the increasing number of HVLPs present. Furthermore,

three HLA alleles showed strong associations with HVLPs at a 5% FDR, and the interplay between them suggested viral replication may be also modulated by these HLA alleles.

Conclusions: This chapter presents, for the first time, a comprehensive V-GWAS to identify viral amino acid variations across entire HCV polyprotein that associated with viral replication in chronic HCV infection across different genotypes. The findings highlight the influence of viral polymorphisms on viral load and provide a foundation for understanding how these variations interact with host genetic factors to shape infection outcomes and disease progression.

4.2 Background

The recommended diagnostic strategy for chronic HCV infection begins with initial screening using an HCV antibody serological test, followed by laboratory-based molecular viral load testing for HCV RNA to confirm active HCV viraemia. Quantitative viral load testing is highly accurate, providing reliable guidance for treatment decisions and monitoring of treatment response [216, 217]. Beyond its clinical utility, viral load also serves as a marker of viral replication dynamics and has been investigated as a potential indicator of disease progression. HCV is an RNA virus with a compact genome encoding ten canonical proteins: three structural (core, envelope 1[E1], envelope 2 [E2]) and seven non-structural proteins (viroporin p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B). Each of these proteins plays a critical role in viral entry, replication, assembly, and immune evasion. Their functions are described in detail in chapter 1, section 1.1.2 (Life cycle) [42]. Because these proteins directly mediate essential steps of the viral life cycle, even minor amino acid changes could influence viral fitness. For example, substitutions in NS5A or NS5B can alter replication complex formation and RNA polymerase activity, thereby affecting replication efficiency. Changes in NS3 protease may influence polyprotein processing and replication competence. Mutations identified within E1/E2 in chapter 2 can change antibody response. These explain how amino acid polymorphisms within HCV proteins can modulate viral load altering replication capacity, immune recognition or viral particle production. Investigating viral load-associated amino acid polymorphisms

therefore provides a valuable framework for understanding how viral genetic variation contributes to persistence and disease outcomes.

Numerous studies have demonstrated that host genetic factors can drive HCV RNA levels within patients and exert selective pressure on the virus, with the most notable examples being polymorphisms in the *IFNL4* gene and variation within the human leukocyte antigen (HLA) system [193, 218, 219]. By contrast, much less is known about the contribution of viral genetic variation itself, and to date, no large-scale analyses have systematically examined viral amino acid polymorphisms associated with HCV replication. Understanding such viral determinants is particularly important, as changes in viral proteins that modulate replication directly affect viral load, which in turn influences immune escape and treatment response. Mutations in specific viral positions that enhance replication efficiency may allow the virus to persist at higher levels, thereby potentially amplifying immune-mediated liver inflammation and fibrosis. Nonetheless, the extent to which viral replication alone drives disease progression remains debated, as host immune responses and broader host-virus interactions may play a more important role. Importantly, identifying viral polymorphisms that influence replication has direct relevance for vaccine development, since vaccines need to target conserved regions that are essential for viral replication or entry, which are therefore less likely to undergo immune escape, and for improving diagnostic and prognostic strategies in chronic HCV infection.

Thus, this chapter first systematically investigated genotype-specific viral amino acid polymorphisms associated with HCV viral load within a well-characterized

cohort, providing new potential insights into the mechanisms of viral replication. This work also reported novel interactions between HLA genes and HVLPs, suggesting that HLA-restricted immune pressure may drive viral adaptation and potentially influence both viral load and clinical outcomes in chronic HCV infection.

4.3 Methods

4.3.1 Patients and samples

Patients from this work were enrolled from the BOSON, STOP-HCV-1 and HCVRUK, with further details provided in chapter 3, section 3.3.1. Briefly, whole genome HCV sequences were generated from 2956 patients after quality filtering for further analysis. After integrating host genetic data, 1028 patients infected with gt3a HCV and 535 with gt1a remained for further analysis. For each patient, the highest recorded viral load measurement available in the clinical dataset was used for analysis. This approach maximised the number of patients included, as viral load measurements were collected at different time points across the contributing clinics and pre-treatment samples were not consistently available for all individuals. Previous analyses within this cohort demonstrated a very high correlation between the highest recorded viral load and pre-treatment viral load, supporting the use of this measure as a reliable substitute for baseline viral replication.

4.3.2 Viral sequencing

A detailed description of how the patient samples were sequenced and the bioinformatic pipeline used to process the sequencing is provided in chapter 2, section 2.3.2 and chapter 3, section 3.3.2. The outputs from the bioinformatics pipeline were analysed using custom code and the R statistical package.

4.3.3 Host genome-wide genotyping and imputation

A detailed description of the how the host genome was genotyped and imputed can be found in chapter 3, section 3.3.3. The outputs from the bioinformatics pipeline were analysed using custom code and the R statistical package.

4.3.4 Host and virus genetic principal component analysis

A detailed description can be found in chapter 3, section 3.3.4.

4.3.5 Statistical analysis

Viral genome-wide association study on viral load

To investigate the effect of individual viral amino acid polymorphisms on viral load, a GWAS was performed using a linear regression model. Log_{10} viral load was used as the response variable and the presence and absence of each amino acid as the explanatory variable. To ensure adequate statistical power, the analysis was restricted to amino acid residues present in sequences from at least 50 patients. The model also included age, sex, ethnicity, sequencing centre, viral population structure (viral PCs) and *IFNL4* gene as covariates to account for confounding.

Incorporating *IFNL4* genotype data reduced the cohort size from 1347 to 1028 for gt3a patients and from 1181 to 535 gt1a patients. The analysis was conducted in a genotype-specific manner.

Covariation and interaction of HVLP analysis

Pearson correlation test was used to assess the covariation between pairs of HVLPs. Logistic regression was then used to test for association among HVLPs adjusting for viral population structure as covariates. The response variable was any of the individual HVLP and the explanatory variable was the rest of the HVLPs. Furthermore, the cumulative effect of HVLPs on cirrhosis status was calculated by linear regression. The response variable was the $\log_{10}$ viral load, and the explanatory variable was number of HVLPs or specific presence or absence of the HVLPs in combinations in the sequence. The non-viral factors previously indicated were included as covariates.

The impact of HVLPs on liver disease (Cirrhosis & HCC)

Logistic regression was conducted to examine the association between HVLP and liver disease status. The presence or absence of cirrhosis/HCC was treated as response variable and individual HVLP / total number of HVLPs were as explanatory variable. Age, sex, ethnicity, sequencing centre, $\log_{10}$ viral load, viral population structure (viral genetic PCs), liver disease associated viral sequence clades as well as the cirrhosis status (in the investigation of HCC) were included as covariates in the model.

Interplay between HVLP and HLA molecules

To explore the impact of HLA alleles on HVLP, logistic regression was performed to examine the association between 148 two-field HLA alleles and each HVLP, adjusting for viral and host principal components, age, sex, ethnicity, sequencing centre as covariates. Following that, the conditional logistic regression including the associated HLA allele was used to evaluate whether the effect of HVLPs on cirrhosis was mediated by the HLA allele. Interaction analysis between HLA alleles and HVLP was tested using logistic regression that included interaction terms to evaluate potential synergistic effects on clinical outcomes.

4.4 Results

4.4.1 Viral Genome-Wide Association Study (V-GWAS) on viral load

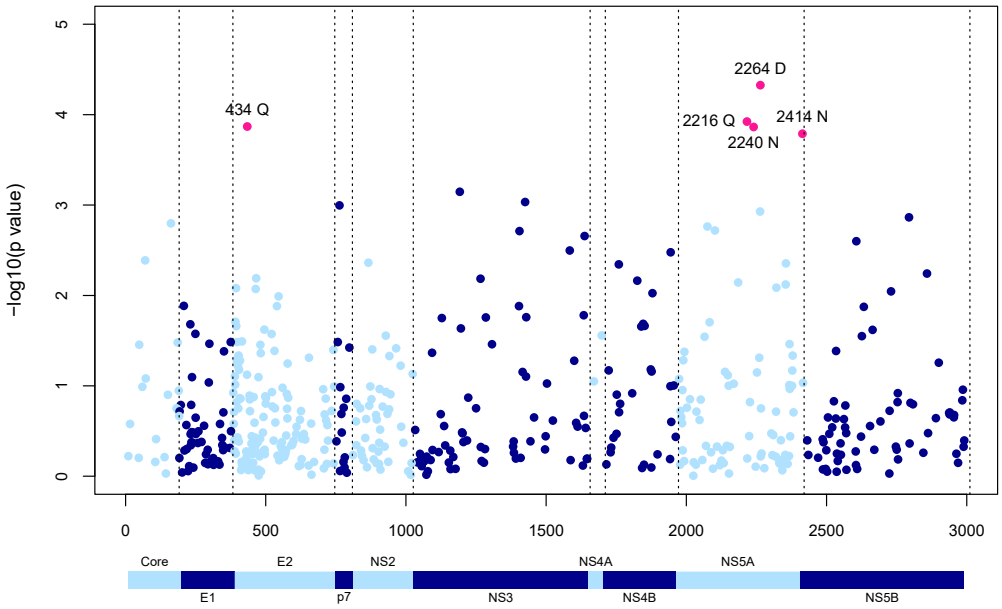
To examine the viral genetic factors influencing HCV RNA levels, a V-GWAS was performed to test for association between viral load and amino acid variants at each site across the entire HCV polyprotein. The model included age, sex, ethnicity and sequencing center as non-genetic covariates. Regarding the genetic factors, host *IFNL4* SNP rs12979860 genotypes (CC vs. non-CC) was added to control for potential confounding, as *IFNL4* genetic variation is a well-established determinant of HCV RNA levels and amino acid variation in HCV, and not include it in the model could result in spurious findings. Incorporating *IFNL4* SNP genotype data reduced the cohort size from 1347 to 1028 for gt3a patients and from 1181 to 535 for gt1a patients. Furthermore, as described in chapter 3 (Figure 3.2), the first three viral genetic PCs for gt3a and the first five for gt1a were included in the model to control the impact of viral population structure on viral load. The $\log_{10}$ viral load was the

response variable, and the presence and absence of each amino acid served as explanatory variable. To ensure statistical robustness, the study included only amino acids observed in sequences from at least 50 individuals.

HVLPs Identified in HCV Genotype 3a

For subjects with HCV gt3a infection, 1089 tests were performed across 508 unique sites. Five polymorphisms were observed to have significant associations with HCV RNA levels at a false-discovery rate (FDR) of 5% (Figure 4.1A and B, Table 4.1). Specifically, four of them – glutamine (Q) at position 2216 ($P = 1.2 \times 10^{-4}$), asparagine (N) at position 2240 ($P = 1.4 \times 10^{-4}$), aspartic acid (D) at position 2264 ($P = 4.7 \times 10^{-5}$) and asparagine (N) at position 2414 ($P = 1.6 \times 10^{-4}$) – located in the NS5A protein were associated with higher RNA levels (Figure 4.1B). Another significant association was in the E2 protein at position 434, where the presence of glutamine (Q) was associated with lower viral load compared to non-Q residue at that position ($P = 1.4 \times 10^{-4}$). Given that, the NS5A 2216Q, NS5A 2240N, NS5A 2264D, NS5A 2414N and E2 434 non-Q were associated with higher viral load and were referred to as High Viral Load associated Polymorphisms (HVLPs) in the rest of the chapter. Additionally, Figure 4.2 demonstrated the distribution of viral load across each amino acid variant at these five polymorphic sites. The corresponding forest plots showed estimated beta coefficients and p values for each variant calculated from linear regression, demonstrating the effect of alternative amino acid residues on viral load compared to the HVLP residue identified in the V-GWAS. This analysis guided subsequent in vitro mutation experiments.

A


−log10(p value)
0
1
2
3
4
5
0
500
1000
1500
2000
2500
3000
434 Q
2264 D
2216 Q
2414 N
2240 N
Core
E2
NS2
NS4A
NS5A
E1
p7
NS3
NS4B
NS5B

B

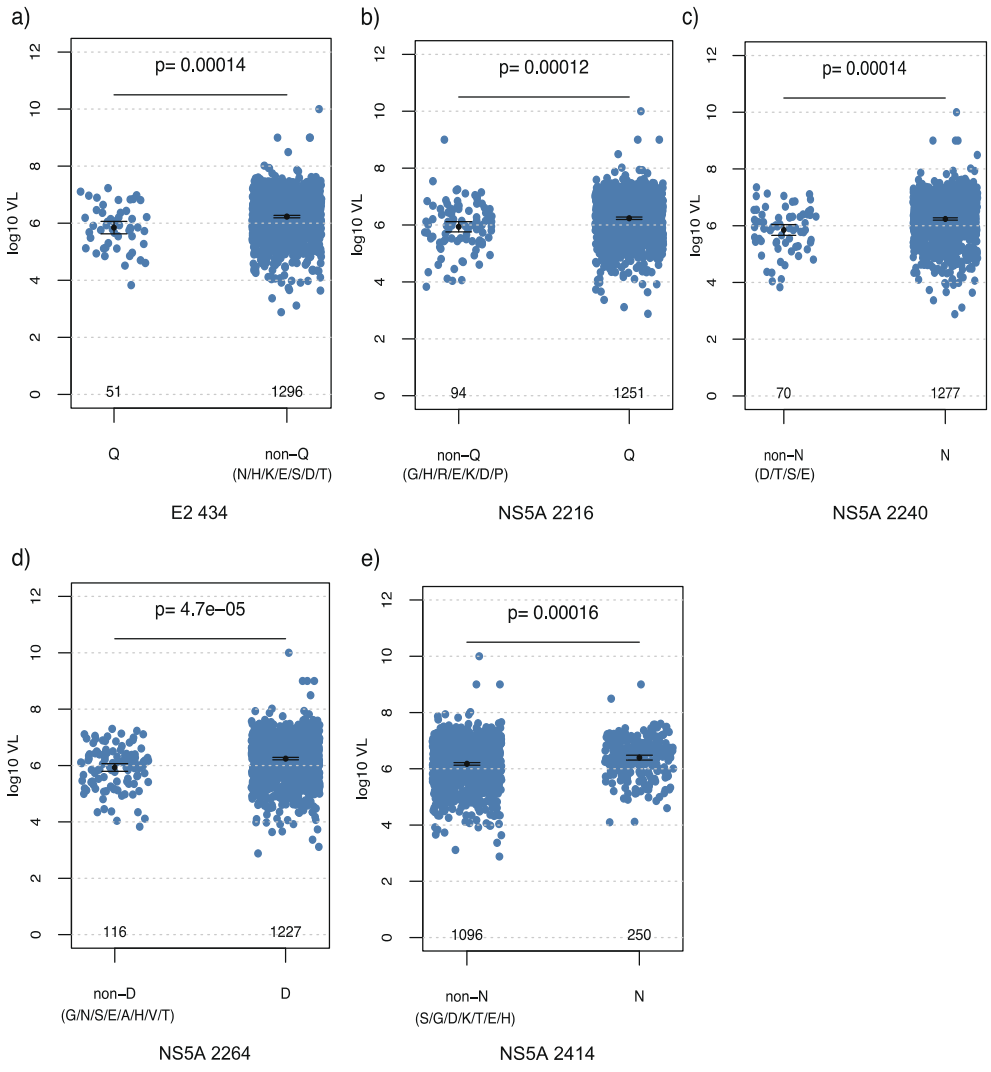

a)
p= 0.00014
log10 VL
51
1296
Q
non−Q
(N/H/K/E/S/D/T)
E2 434
b)
p= 0.00012
94
1251
non−Q
(G/H/R/E/K/D/P)
Q
NS5A 2216
c)
p= 0.00014
70
1277
non−N
(D/T/S/E)
N
NS5A 2240
d)
p= 4.7e−05
116
1227
non−D
(G/N/S/E/A/H/V/T)
D
NS5A 2264
e)
p= 0.00016
1096
250
non−N
(S/G/D/K/T/E/H)
N
NS5A 2414

Figure 4.1: Association between HCV gt3a amino acids and log_{10} viral load.
(A) Manhattan plot of association tests between gt3a HCV amino acids and viral load. For each site only the most associated p-value is plotted. The pink dots with amino acid labelled indicate the five sites significantly associated with the viral load at 5% FDR. The schematic of HCV polyprotein is under the Manhattan plot.
(B) Dot plots of the impact of identified amino acid residues on viral load. The mean change in log_{10} viral load is shown in black dot for each group of amino acids. The lines show its 95% confidence interval (CI). The p values for difference in mean calculated by linear regression. The frequency of each amino acid group was also shown in the plot.

The tested site	The most associated amino acid	P value	Effect size	Standard error	q value
2264	**d**	**4.72E-05**	**0.334**	**0.082**	**0.029**
2216	**q**	**0.00012**	**0.348**	**0.090**	**0.029**
434	**q**	**0.00014**	**-0.501**	**0.131**	**0.029**
2240	**n**	**0.00014**	**0.408**	**0.106**	**0.029**
2414	**n**	**0.00016**	**0.229**	**0.060**	**0.029**
1192	t	0.00071	-0.356	0.105	0.091
1425	i	0.00093	-0.371	0.112	0.100
763	h	0.00101	0.353	0.107	0.100
2263	p	0.00118	-0.228	0.070	0.106
2794	h	0.00137	0.340	0.106	0.111

Table 4.1: Top ten associations between gt3a HCV amino acid variations and viral load.
The table presents the first ten results from V-GWAS testing the association between baseline viral load and amino acid polymorphisms at 508 sites (in total 1089 residues were tested at these 508 sites, but only the most associated residue per site is shown in the table). P values, effect sizes, and standard errors are reported, along with q-values representing FDR-corrected p-values to account for multiple testing. Associations at 5% FDR are highlighted in bold.

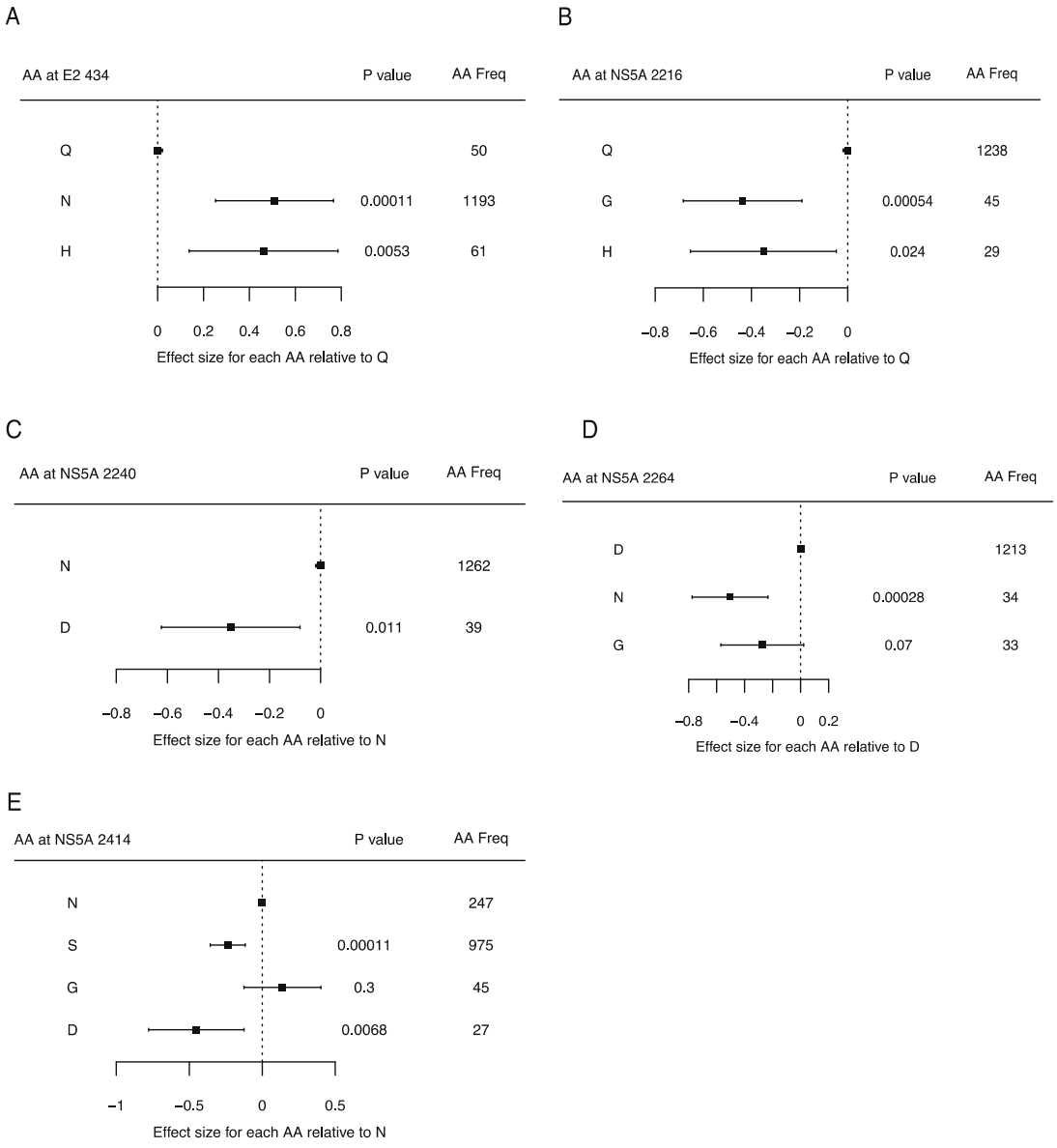


Figure 4.2: Association between amino acid residues and viral load at five polymorphic sites in HCV gt3a infection.

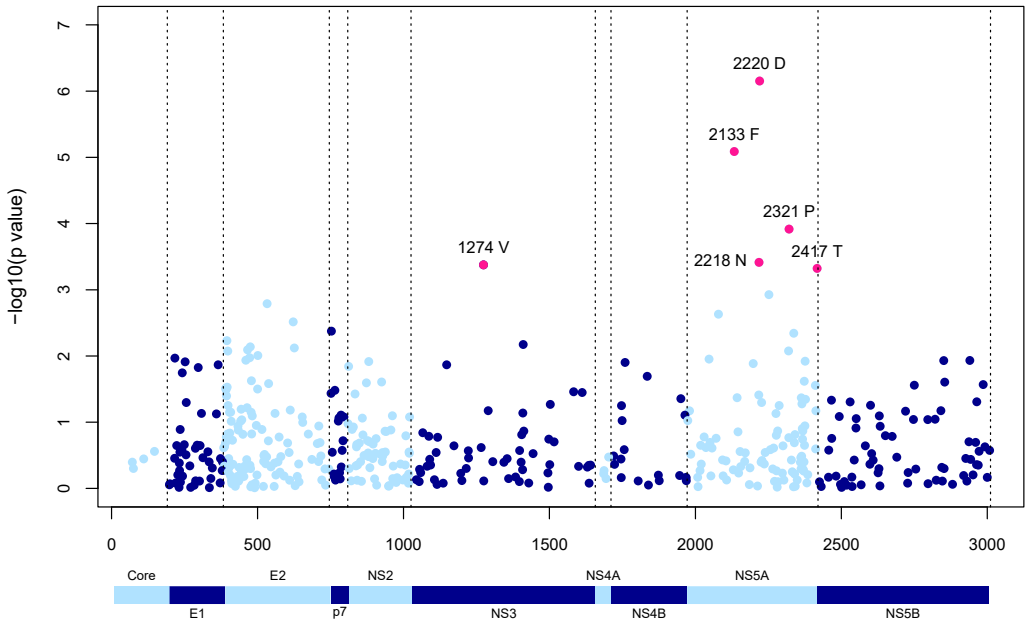
Amino acid variants at sites E2 434 (A), NS5A 2216 (B), NS5A 2240 (C), NS5A 2264 (D) and NS5A 2414 (E).

The black squares and the lines indicate the estimated effect size (β) and its 95% confidence interval (CI) estimated from logistic regression. *P* values were calculated using linear regression model. Only substitution with counts greater than twenty were shown in the figures.

HVLPs Identified in HCV Genotype 1a

For patients infected with HCV gt1a, 919 tests were performed at 458 unique sites. At an FDR of 5%, six polymorphic positions were identified as significantly associated with differences in mean $\log_{10}$ viral load within patients infected with HCV gt1a. All these positions occurred in the non-structural proteins (Figure 4.3A, Table 4.2). Five of the significant associations with higher viral load were identified in the NS5A protein (Figure 4.3B): phenylalanine (F) at position 2133 ($P = 8.2 \times 10^{-6}$), asparagine (N) at position 2218 ($P = 3.9 \times 10^{-4}$), aspartic acid (D) at position 2220 ($P = 7.2 \times 10^{-7}$), non-proline (non-P) residues at position 2321 ($P = 1.2 \times 10^{-4}$), and non-threonine (non-T) at site 2417 ($P = 4.46 \times 10^{-4}$). An additional significant association was found in the NS3 protein, where valine (V) at position 1274 was associated with higher viral load compared to non-valine residues ($P = 4.2 \times 10^{-4}$). Similarly, the following residues associated with higher viral load – NS3 1274V, NS5A 2133F, NS5A 2218N, NS5A 2220D, NS5A 2321non-P and NS5A 2417 non-T – were classified as gt1a HVLPs (also see the distribution of viral load across each amino acid variant at these positions in Figure 4.4).

A


−log10(p value)
0
1
2
3
4
5
6
7
2220 D
2133 F
2321 P
2417 T
2218 N
1274 V
0
500
1000
1500
2000
2500
3000
Core
E2
NS2
NS4A
NS5A
E1
p7
NS3
NS4B
NS5B

B

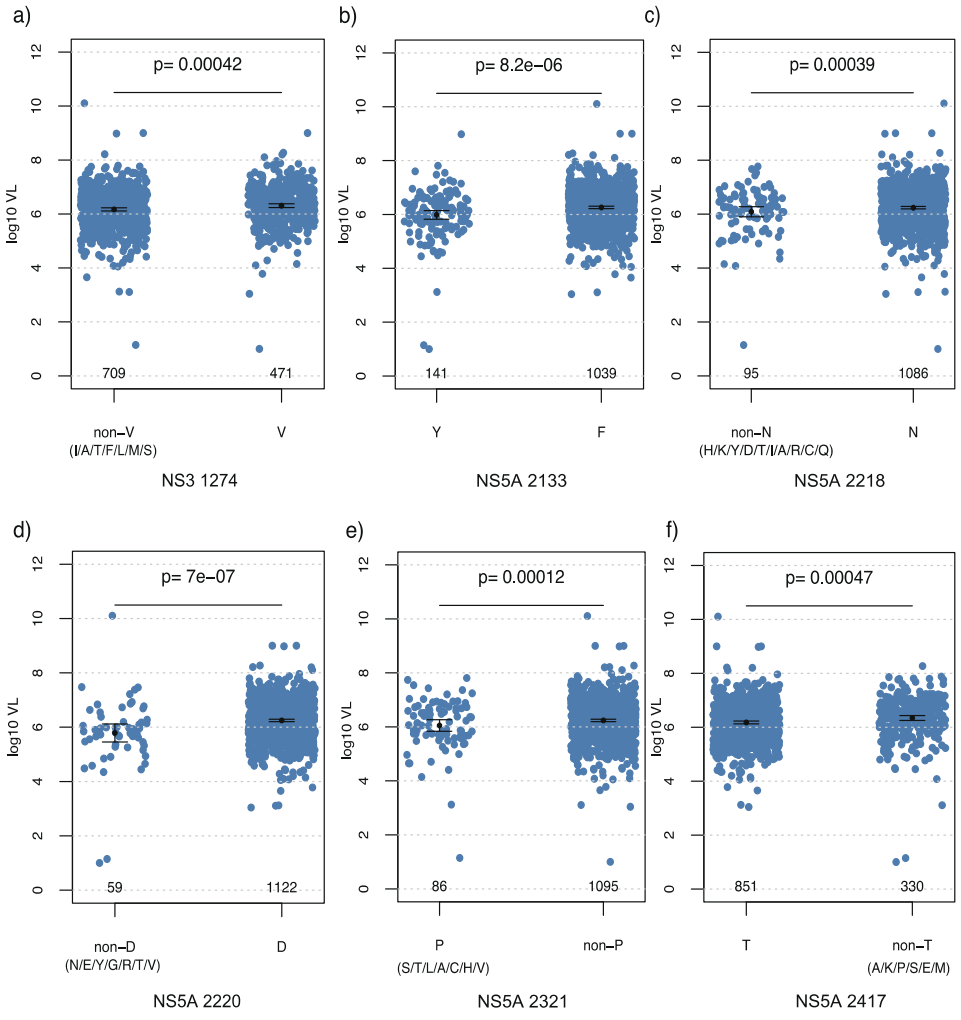

a)
p= 0.00042
log10 VL
709
471
non−V
(I/A/T/F/L/M/S)
V
NS3 1274
b)
p= 8.2e−06
141
1039
Y
F
NS5A 2133
c)
p= 0.00039
95
1086
non−N
(H/K/Y/D/T/I/A/R/C/Q)
N
NS5A 2218
d)
p= 7e−07
59
1122
non−D
(N/E/Y/G/R/T/V)
D
NS5A 2220
e)
p= 0.00012
86
1095
P
(S/T/L/A/C/H/V)
non−P
NS5A 2321
f)
p= 0.00047
851
330
T
non−T
(A/K/P/S/E/M)
NS5A 2417

Figure 4.3: Association between HCV gt1a amino acids and cirrhosis status. (A) Manhattan plot of association tests between gt1a HCV amino acids and viral load. For each site only the most associated p-value is plotted. The pink dots with amino acid labelled indicate the six sites significantly associated with the presence or absence of cirrhosis at 5% FDR. The schematic of HCV polyprotein is under the Manhattan plot. (B) Dot plots of the impact of identified amino acid residues on viral load. The mean change in $\log_{10}$ viral load is shown in black dot for each group of amino acids. The lines show its 95% confidence interval (CI). The p values for difference in mean were calculated by linear regression. The frequency of each amino acid group was also shown in the plot.

The tested site	The most associated amino acid	P value	Effect size	Standard error	q value
2220	**d**	**7.05E-07**	**0.704**	**0.140**	**5.00E-04**
2133	**f**	**8.17E-06**	**0.481**	**0.107**	**0.002**
2321	**p**	**0.00012**	**-0.504**	**0.130**	**0.021**
2218	**n**	**0.00039**	**0.439**	**0.123**	**0.045**
1274	**v**	**0.00042**	**0.292**	**0.082**	**0.045**
2417	**t**	**0.00047**	**-0.256**	**0.073**	**0.045**
2252	v	0.00118	-0.261	0.080	0.084
533	e	0.00162	-0.233	0.074	0.096
2079	k	0.00234	-0.315	0.103	0.127
622	i	0.00305	0.209	0.070	0.154

Table 4.2: Top ten associations between gt1a HCV amino acid variants and cirrhosis status.

The table presents the first ten significant results from V-GWAS testing the association between presence or absence of cirrhosis and polymorphisms at 458 sites (in total 919 residues were tested at these 458 sites, but only the most associated residue per site is shown in the table). P values, effect sizes, and standard errors are reported, along with q-values representing FDR-corrected p-values to account for multiple testing. Associations at 5% FDR are highlighted in bold.

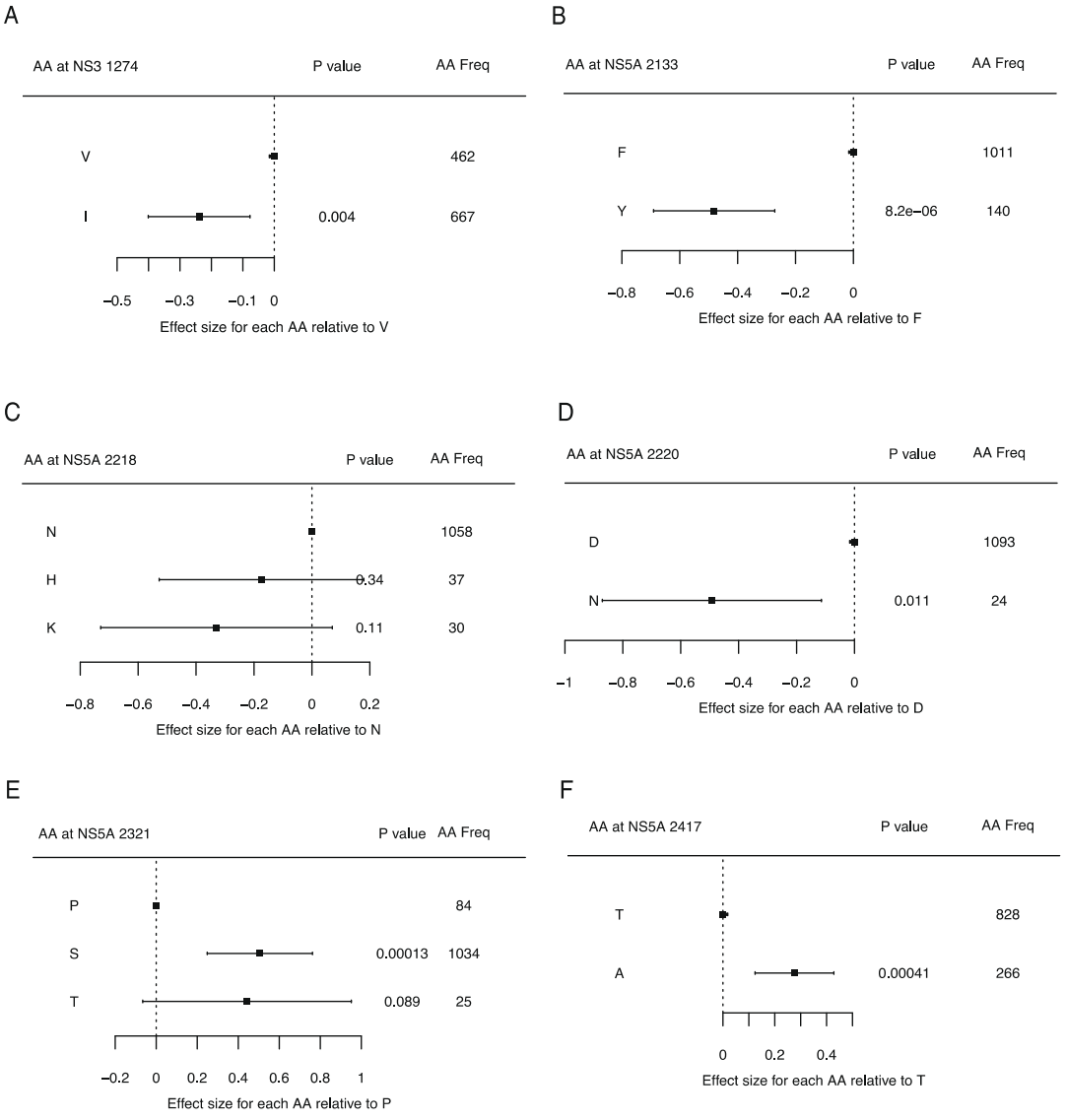


Figure 4.4: Association between each amino acid residue and viral load at each of the six sites identified within HCV gt1a infection.

(A) Amino acid variants at site 1274. (B) Amino acid variants at site 2133.

(C) Amino acid variants at site 2218. (D) Amino acid variants at site 2220.

(E) Amino acid variants at site 2321. (F) Amino acid variants at site 2417.

The black squares and the lines indicate the estimated effect size (β) and its 95% confidence interval (CI) estimated from logistic regression. *P* values were calculated using linear regression model. Only substitution with counts greater than twenty were shown in the figures.

4.4.2 Covariation of HVLPs and their impact on viral load

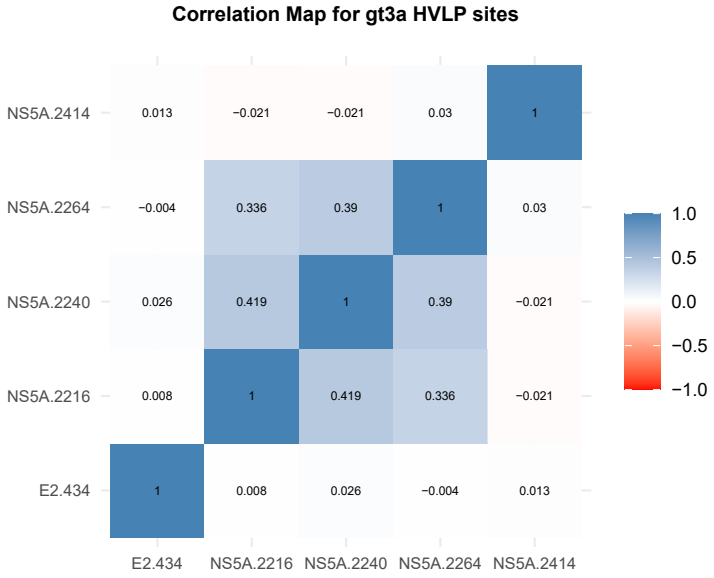
To assess the relationship between the identified HVLPs and their combined effects on viral RNA levels, linear regression was performed incorporating HVLPs combinations present in viral sequences. The same covariates were used in the V-GWAS model.

HVLPs covariation in gt3a

Before investigating the combined impact of HVLPs on viral load, correlations between each pair of HVLPs were examined first. Pearson correlation test showed moderate positive pairwise correlations among the polymorphic sites NS5A 2216, NS5A 2240 and NS5A 2264 ($r_{2216 \text{ vs. } 2240} = 0.42$, $r_{2264 \text{ vs. } 2216} = 0.34$, $r_{2264 \text{ vs. } 2240} = 0.39$; Figure 4.5A), suggesting potential linkage or co-occurrence. In contrast, NS5A 2414 and E2 434 were not correlated with other sites. Logistic regression was then performed to assess the associations among HVLPs, adjusting for the first three genetic viral PCs. Strong associations were again observed between NS5A 2216, NS5A 2240 and NS5A 2264 (Figure 4.5B). The conditional linear regression was conducted to study whether the impact of these three HVLPs on viral load was independent. For NS5A 2216, the unadjusted effect estimate (β) on viral load was 0.35 ($P = 1.2 \times 10^{-4}$); this reduced to 0.26 ($P = 0.008$) after adjusting for NS5A 2240 and to 0.26 ($P = 0.0064$) after adjusting for NS5A 2264. When adjustment was made for both sites simultaneously, the association between NS5A 2216 and $\log_{10}$ viral load remained statistically significant, but the effect size decreased further ($\beta = 0.21$, $P = 0.033$). For NS5A 2240, the β on $\log_{10}$ viral load decreased from 0.41 to 0.28 ($P = 0.015$) and to 0.3 ($P = 0.0099$) after conditioning

on sites 2216 and 2264, respectively. Adjusting for both sites together rendered the association no longer statistically significant (β = 0.21, *P* = 0.086). A similar pattern was observed for NS5A 2264, where adjustment for sites 2216 and 2240 individually slightly reduced the β estimates (β = 0.26, *P* = 0.024; β = 0.25, *P* = 0.0038, respectively). Further conditioning on both sites reduced the β from 0.33 to 0.22, but the association remained significant (*P* = 0.015).

A



B

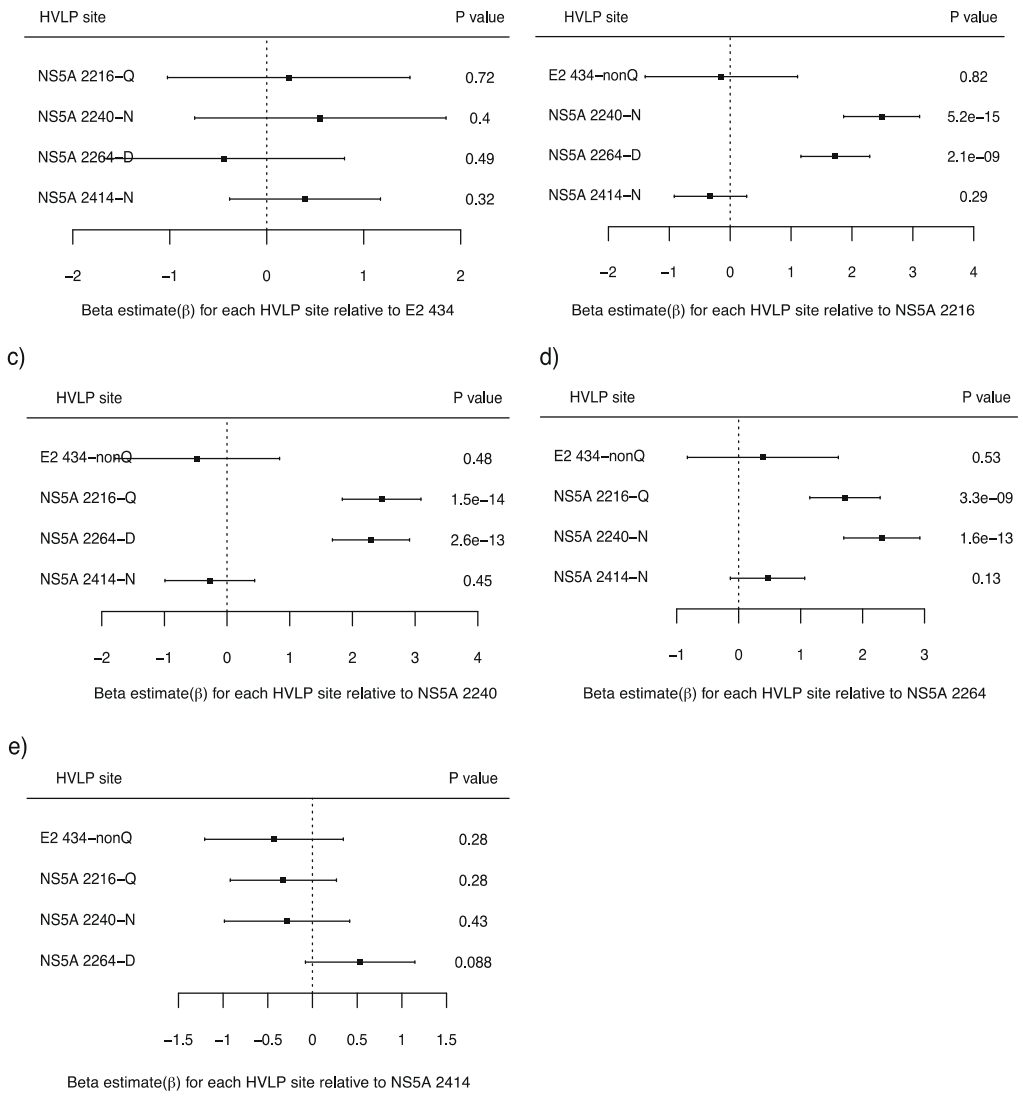
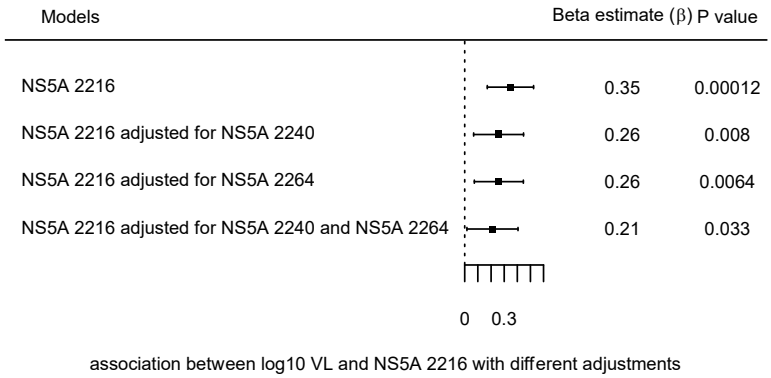


Figure 4.5: Covariation between pairs of gt3a HVLPs.

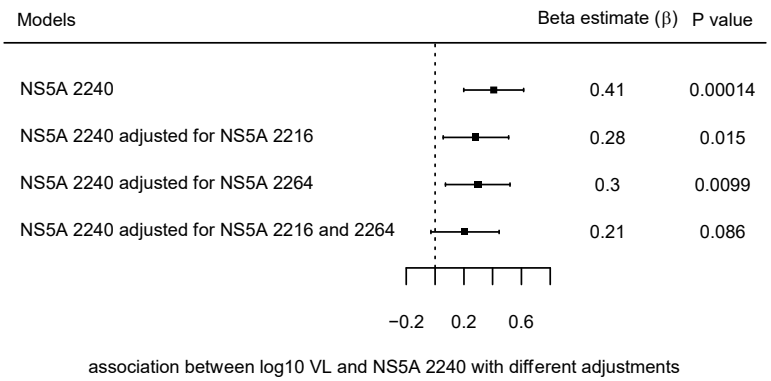
(A) Pairwise Pearson correlation matrix of binary HVLP residues. Color intensity demonstrates the power and direction of correlation, with blue indicating positive and red indicating negative correlations. Correlations were observed among amino acid variations at sites 2216, 2240 and 2264.

(B) Forest plots of association between the five reported sites using logistic regression. Each site was set as the reference group to compare against all others, presenting the resulting log odds ratios and p values. The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

A



B



C

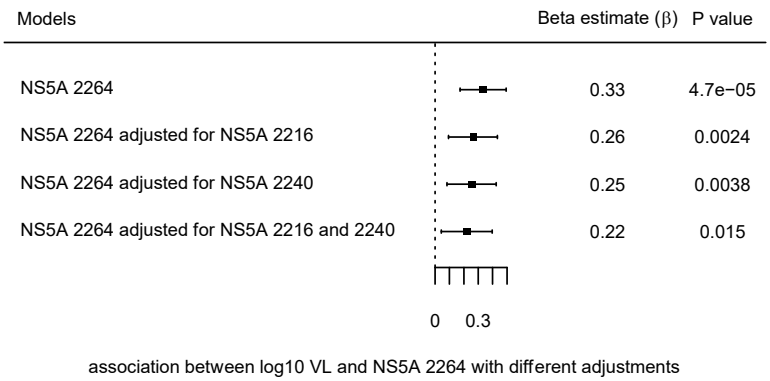


Figure 4.6: Conditional analysis of viral load after covariate adjustment.

(A) The forest plot indicated association test between viral load and amino acid variation at site 2216, with adjustment for its correlated sites 2240 and 2264 – both individually and in combination.

(B) Conditional association test between viral load and amino acid variation at site 2240, with adjustment for its correlated sites 2216 and 2264 – both individually and in combination.
(C) Conditional association test between viral load and amino acid variation at site 2264, with adjustment for its correlated sites 2240 and 2264 – both individually and in combination. The black squares and the lines indicate the estimated effect size (β) and its 95% confidence interval (CI) estimated from linear regression.

Cumulative effect of gt3a HVLPs on viral load

Individuals were stratified based on the total number of gt3a HVLPs present in their consensus viral sequences to assess the cumulative effects on HCV RNA levels. A highly significant association was observed between viral load and the number of HVLPs, with an increase in HVLPs associated with higher viral load (linear regression, $P = 5.2 \times 10^{-13}$, Figure 4.7A), indicating that each additional HVLP significantly increased the $\log_{10}$ viral load. The number of HVLPs were further classified into categorical groups to investigate how viral load changed across groups (Figure 4.7B). Compared to individuals carrying all five HVLPs (n = 207, mean for 5 HVLPs = 2,747,369 IU/ml), the viral load reduced by approximately 1.6-fold in patients carrying four HVLPs (n = 934, mean for 4 HVLPs = 1,712,782 IU/ml, $P_{4\ \mathrm{HVLPs}} = 2.2 \times 10^{-3}$), 2.9-fold in those with three HVLPs (n = 142, mean for 3 HVLPs = 959,733 IU/ml, $P_{3\ \mathrm{HVLPs}} = 7.5 \times 10^{-8}$) and 4.6-fold in subjects with two HVLPs (n = 41, mean for 2 HVLPs = 595,216 IU/ml, $P_{2\ \mathrm{HVLPs}} = 1.5 \times 10^{-9}$), suggesting a stepwise decrease as the number of HVLPs reduced. Additionally, a significant reduction in viral load was observed within patients carrying one HVLP relative to those with five HVLPs (mean for 1 HVLP = 658076 IU/ml, $P_{1\ \mathrm{HVLP}} = 1.5 \times 10^{-3}$), even though only 22 patients were tested in the analysis.

To further explore the effect of the HVLP combinations, 21 unique combinations were generated based on the presence or absence of the five reported gt3a HVLPs (E2 434 non-Q, NS5A 2216 Q, NS5A 2240 N, NS5A 2264 D and NS5A 2414) across patient viral sequences (Table 4.3). Linear regression was conducted to test for association between viral load and each unique combination of those HVLPs. The HVLPs combination with highest mean log_{10} viral load was observed in patients carrying all five HVLP residues (mean log_{10} VL = 6.44). No combination in which all HVLPs were absent was observed. The lowest log_{10} viral load (log_{10} VL = 5.12) was seen in a single patient with the combination E2 434Q, NS5A non-Q, NS5A 2240 N, NS5A 2264 non-D and NS5A 2414 non-N, and therefore does not represent a group mean. However, to improve statistical reliability, only combinations observed in more than 20 individuals were included in the forest plot (Figure 4.8). Notably, the lowest mean viral load was observed in patients with three HVLPs at NS5A 2216 (Q), NS5A 2240 (N) and NS5A 2264 (D) among groups with reasonable sample size. An overall reduction in the beta estimates for the association with log_{10} viral load was observed as the number of HVLPs in patient sequences decreased, although combinations with very low frequency (n=1) should be interpreted cautiously.

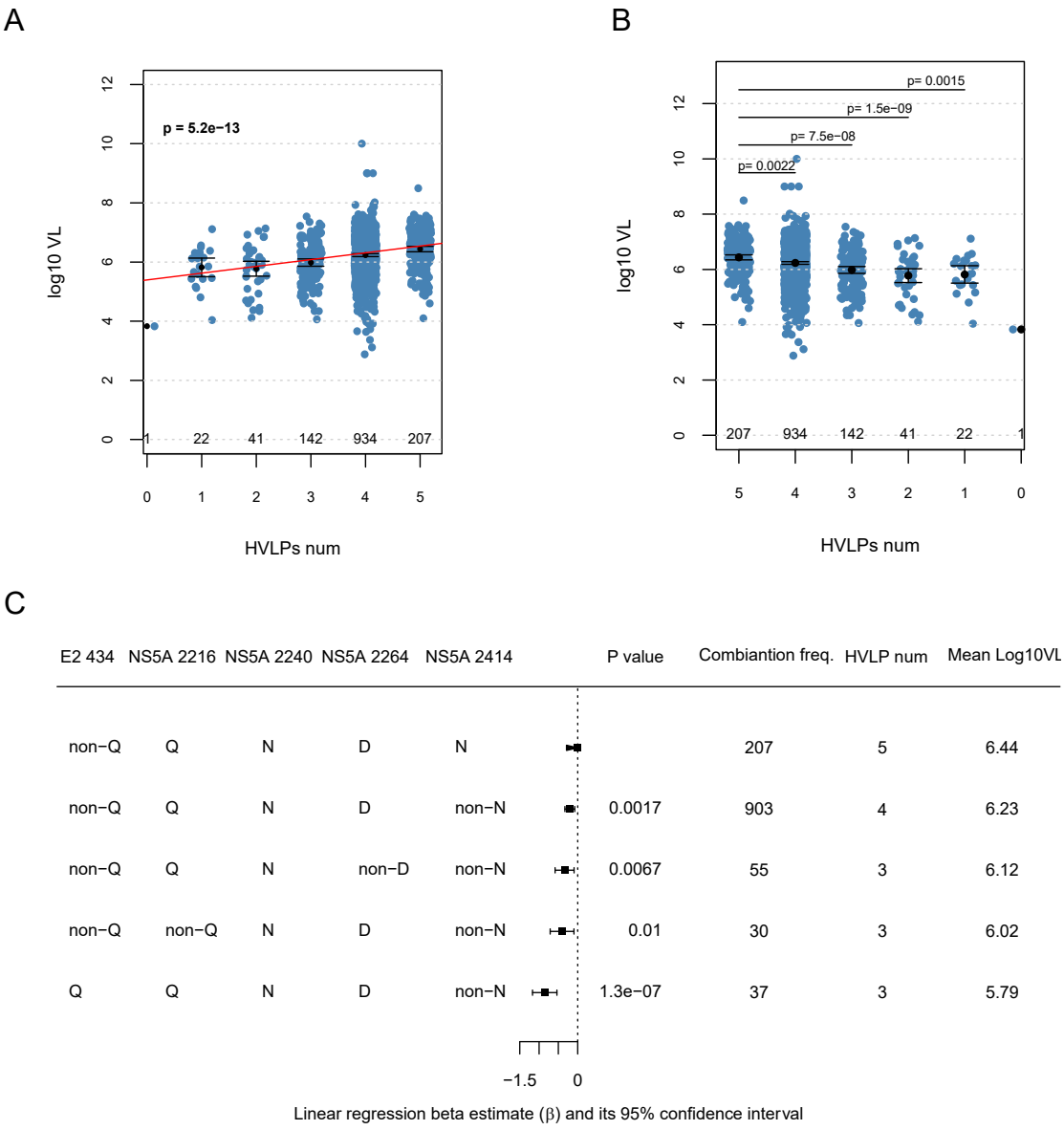


Figure 4.7: Association between log_{10} viral load and multiple gt3a HVLPs.
(A) and (B): The distribution of log_{10} viral load across patients stratified by the number of HVLPs carried. The total number of HVLPs present at each group was shown below the x-axis. The p-value in (A) was calculated from linear regression. The frequency of each group was shown above the x-axis. The red line in (A) represents the linear regression fit, showing a positive trend. The p values in (B) were calculated from linear regression comparing each HVLP count group to the reference group carrying five HVLPs. Individual data points were represented by blue dots, with group means ± standard error indicated by black bars.
(C) Covariation between the five gt3a HVLPs and their associations with log_{10} viral load. Combinations with a frequency greater than 20 involved in tests are listed in

the forest plot on the left. The p-values (linear regression), frequency of each tested combination, number of HVLPs carried by each combination and the mean value of log_{10} viral load for each combination were shown on the right.

The black squares and the lines indicate the estimated beta and its 95% confidence interval estimated from linear regression.

CAP combinations (n = 21)					Regression outcomes				Combination characteristics		
site434	site2216	site2240	site2264	site2414	β	SE	z	p value	Combination freq. *	HVLPs num	Mean log_{10}VL
non-Q	Q	N	D	N	0	0	0	NA	207	5	6.439
non-Q	non-Q	N	D	N	0.287	0.245	1.172	0.241	11	4	6.802
Q	Q	N	D	N	0.025	0.324	0.077	0.939	6	4	6.509
non-Q	Q	N	D	non-N	-0.204	0.065	-3.150	0.002	903	4	6.228
non-Q	Q	non-N	D	N	-0.229	0.298	-0.769	0.442	7	4	6.292
non-Q	Q	N	non-D	N	-0.641	0.362	-1.773	0.077	6	4	5.869
Q	Q	N	non-D	N	0.565	0.714	0.792	0.429	1	3	7.230
non-Q	Q	N	non-D	non-N	-0.340	0.125	-2.719	0.007	55	3	6.119
non-Q	Q	non-N	D	non-N	-0.365	0.259	-1.411	0.158	11	3	5.969
non-Q	non-Q	N	D	non-N	-0.405	0.157	-2.581	0.010	30	3	6.015
non-Q	non-Q	N	non-D	N	-0.746	0.418	-1.783	0.075	3	3	5.757
Q	Q	N	D	non-N	-0.854	0.161	-5.320	1.29E-07	37	3	5.791
non-Q	non-Q	non-N	D	N	-1.660	0.714	-2.325	0.020	1	3	4.945
Q	non-Q	N	D	non-N	-0.183	0.715	-0.256	0.798	1	2	6.364
non-Q	Q	non-N	non-D	non-N	-0.648	0.233	-2.785	0.005	11	2	5.865
non-Q	non-Q	non-N	D	non-N	-0.927	0.298	-3.114	0.002	9	2	5.954
non-Q	non-Q	N	non-D	non-N	-0.978	0.245	-3.995	6.94E-05	11	2	5.612
non-Q	non-Q	non-N	non-D	N	-1.201	0.417	-2.884	0.004	5	2	5.537
non-Q	non-Q	non-N	non-D	non-N	-0.524	0.206	-2.544	0.011	19	1	5.847
Q	non-Q	non-N	D	non-N	-0.808	0.716	-1.130	0.259	1	1	5.664
Q	non-Q	N	non-D	non-N	-1.526	0.714	-2.137	0.033	1	1	5.120

**** Results for combinations with low frequency (n < 20) are shown for descriptive purposes only and should be interpreted cautiously. Particularly, estimates for combinations with n < 5 are extremely unstable and not statistically meaningful.***

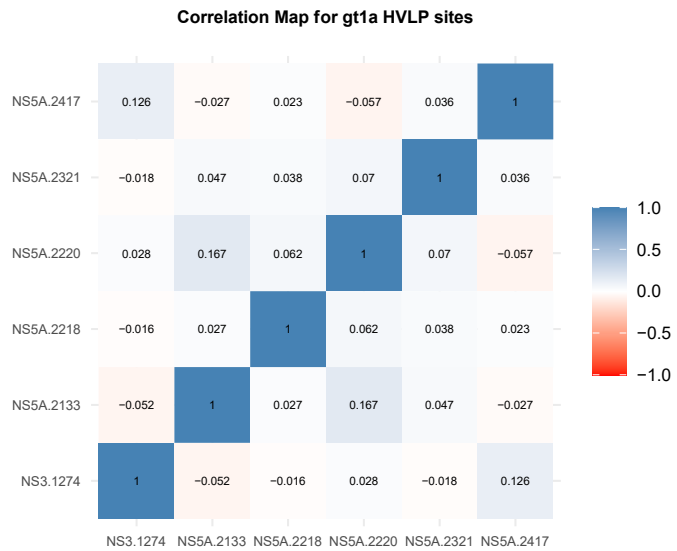
Table 4.3: Associations between gt3a HVLP combinations and log_{10} viral load.

The table presents the total 21 unique combinations based on the specific presence or absence of the five individual HVLPs in patients' sequences. The linear regression outcomes (estimated beta, standard errors, p values, etc) and characteristics of these combinations are also shown in the table.

Covariation of gt1a HVLPs

A similar analysis pattern was performed for patients infected with gt1a to test the cumulative effect of HVLPs on viral load. Pearson pairwise correlation indicated no strong associations among the gt1a HVLPs (Figure 4.8A). Logistic regression accounted for viral population structure by including the first five viral PCs as covariates was further performed to access their covariation (Figure 4.8B). In particular, a nominally association was observed between NS3 1274 and NS5A 2417 (P = 0.018). The estimated beta (β) for association between NS3 1274 and $\log_{10}$ viral load was slightly reduced from 0.29 to 0.27 after adjusting for NS5A 2417 but remained significant (P = 0.00083). NS5A 2133 was significantly associated with NS5A 2220 ($P = 1.7 \times 10^{-6}$). After adjusting the association of NS5A 2133 with viral load for NS5A 2220, the effect size (β) decreases from 0.48 to 0.37 ($P = 6.7\times10^{-4}$, Figure 4.9A). Similarly, adjusting NS5A 2220 for NS5A 2133 resulted in a reduced estimate that remained significant (β = 0.58, $P = 5.5\times10^{-5}$, Figure 4.9B), and vice versa. Collectively, conditional analyses confirmed that the effect of each variant on viral load was independent of the others but partially overlapping.

A



B

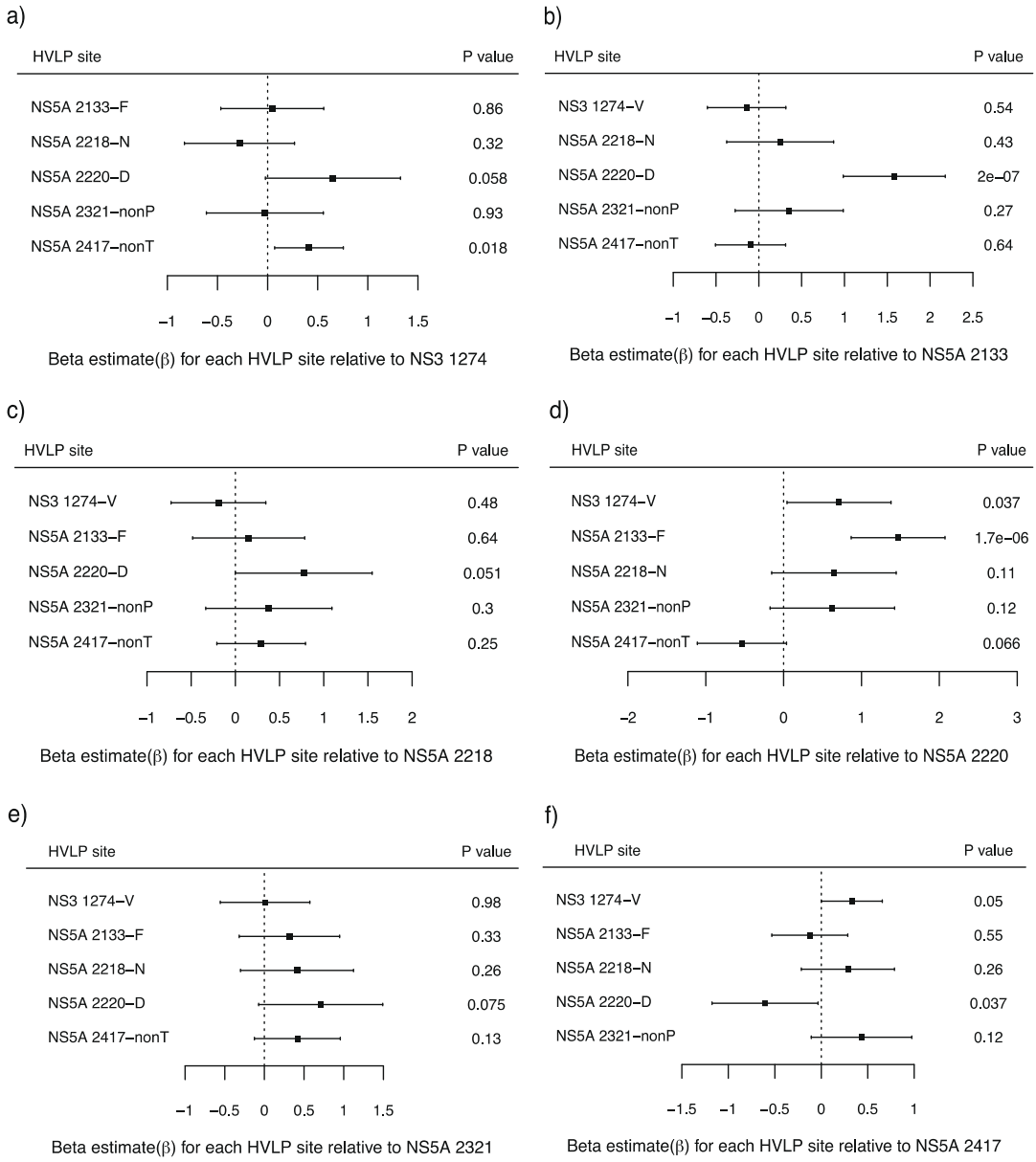
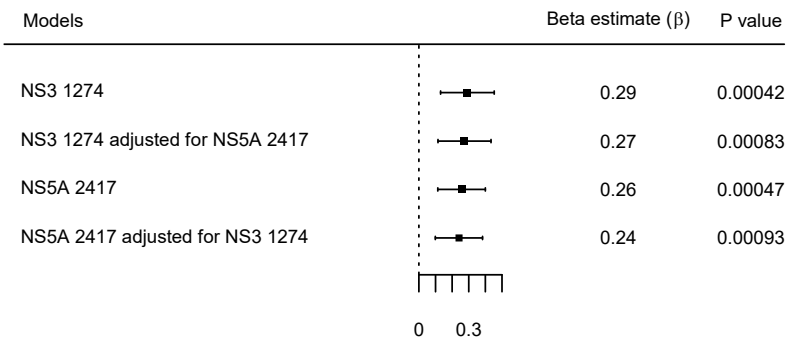


Figure 4.8: Covariation between pairs of gt1a HVLPs.

(A) Pairwise Pearson correlation matrix of binary HVLP residues. Color intensity demonstrates the power and direction of correlation, with blue indicating positive and red indicating negative correlations.

(B) Forest plots of association between the six reported sites using logistic regression. Each site was set as the reference group to compare against all others, presenting the resulting log odds ratios and p values. The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

A



B

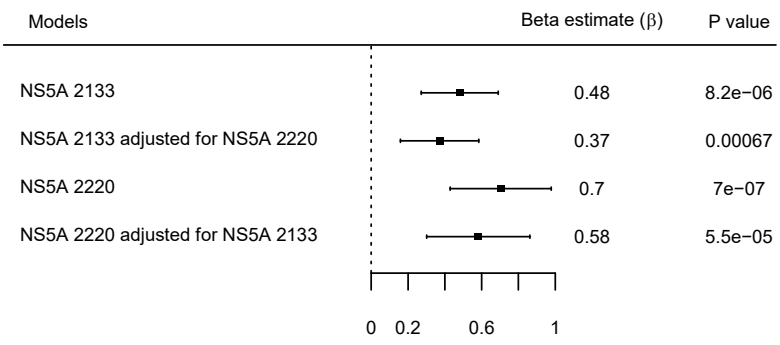


Figure 4.9: Conditional analysis of viral load after covariate adjustment.

(A) The forest plot indicated association test between $\log_{10}$ viral load and amino acid variation at site 1274 and 2417. For each site, effect sizes and p values from linear regression models with and without adjustment for its correlated site were presented.

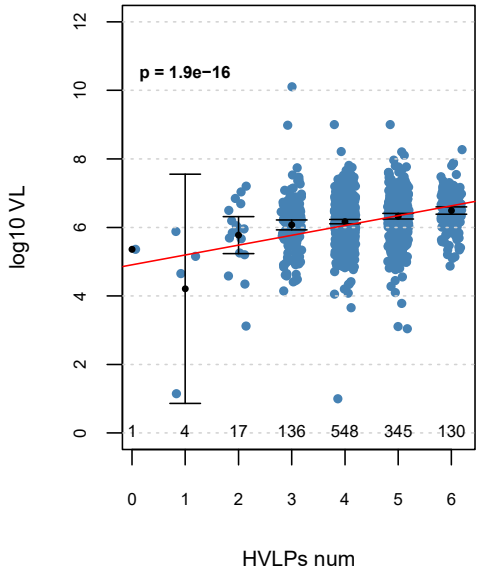
(B) The forest plot indicated association test between $\log_{10}$ viral load and amino acid variation at site 2133 and 2220. Effect sizes and p values from linear regression models with and without the adjustment for its correlated site were shown on the right.

Cumulative effect of gt1a HVLPs on viral load

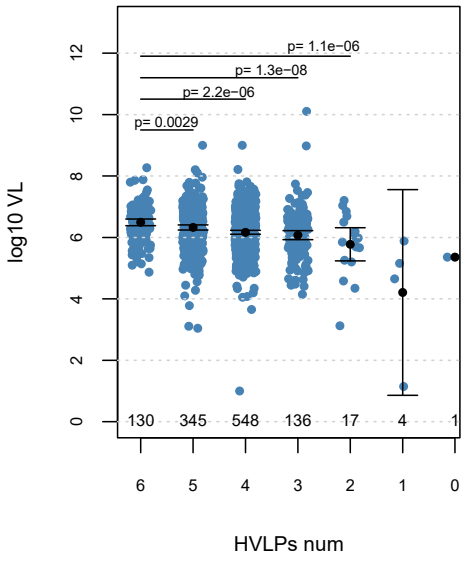
When patients were categorized into groups according to the number of gt1a HVLPs present, a strong positive association was observed between the number of HVLPs and $\log_{10}$ viral load ($P < 1.9 \times 10^{-16}$, Figure 4.10A). Patients carrying all six gt1a HVLPs had the highest viral load in mean (n = 130, mean for 6 HVLPs = 3,098,120 IU/ml, Figure 4.10B). In comparison, individuals with five (n = 345, mean for 5 HVLPs = 2,113,130 IU/ml), four (n = 548, mean for 4 HVLPs = 1,468,453 IU/ml), three (n = 136, mean for 3 HVLPs = 1,186,124 IU/ml) and two gt1a HVLPs (n = 17, mean for 2 HVLPs = 594,655 IU/ml) showed approximately 1.6-fold ($P_{5\ HVLP} = 0.0029$), 2.1-fold ($P_{4\ HVLP} = 2.2 \times 10^{-6}$), 2.6-fold ($P_{3\ HVLP} = 1.3 \times 10^{-8}$) and 5.2-fold ($P_{2\ HVLP} = 1.1 \times 10^{-6}$) lower viral loads, respectively. Subjects with only one HVLP showed a sharp reduction in HCV RNA levels compared to those with six (mean for 1 HVLP = 16,150 IU/ml, $P_{1\ HVLP} = 3.9 \times 10^{-10}$). However, because this group included only four patients, the estimate and p value were not statistically reliable and should be interpreted with caution. Additionally, only one patient was observed carrying all six non-HVLPs at those sites, which did not allow for a reliable comparison.

Following that, 40 unique combinations were produced based on the gt1a HVLPs present in the patient sequences. The impact of each unique combination on viral load was demonstrated in Figure 4.10C and Table 4.4. The extremely high viral load (mean $\log_{10}$ VL = 6.49) was observed in patients whose sequences carried all six gt1a HVLPs simultaneously (NS3 1274V, NS5A 2133F, NS5A 2218N, NS5A 2220D, NS5A 2321 non-P, and NS5A 2417 non-T). In contrast, combinations with fewer HVLPs were associated with progressively lower mean HCV RNA levels, indicating a cumulative effect of HVLPs on viral load.

A

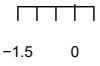


B



C

NS3 1274	NS5A 2133	NS5A 2218	NS5A 2220	NS5A 2321	NS5A 2417	P value	Combiantion freq.	HVLP num	Mean Log10VL
V	F	N	D	non−P	non−T		130	6	6.49
non−V	F	N	D	non−P	non−T	0.027	117	5	6.33
V	F	N	D	non−P	T	0.0042	202	5	6.29
V	F	N	D	P	T	0.13	20	4	6.23
V	F	non−N	D	non−P	T	0.027	24	4	6.28
non−V	F	N	D	non−P	T	1.7e−05	416	4	6.17
V	Y	N	D	non−P	T	0.0082	38	4	6.08
non−V	Y	N	D	non−P	T	0.046	38	3	6.26
non−V	F	N	D	P	T	0.00088	28	3	6.01
non−V	F	non−N	D	non−P	T	3.6e−05	32	3	6.03



Linear regression beta estimate (β) and its 95% confidence interval

Figure 4.10: Association between log_{10} viral load and multiple gt1a HVLPs.
(A) and (B): The distribution of log_{10} viral load across patients carrying different number of gt1a HVLPs. The total number of HVLPs present at each group was shown below the x-axis. The p-value in (A) was calculated from linear regression. The frequency of each group was shown above the x-axis. The red line in (A) represents the linear regression fit, showing a positive trend. The p values in (B) were calculated from linear regression comparing each HVLP count group to the reference group carrying six HVLPs. Individual data points were represented by blue dots, with group means ± standard error indicated by black bars.
(C) Covariation between the six gt1a HVLPs and their associations with log_{10} viral load. Combinations with a frequency greater than 20 involved in tests are listed in the forest plot on the left. The p-values (linear regression), frequency of each tested combination, number of HVLPs carried by each combination and the mean value of log_{10} viral load for each combination were shown on the right.
The black squares and the lines indicate the estimated beta and its 95% confidence interval estimated from linear regression.

CAP combinations (n = 40)						Regression outcomes				Combination characteristics		
site1274	site2133	site2218	site2220	site2321	site2417	β	SE	z	p value	Combination freq.*	HVLPs num	Mean log_{10}VL
V	F	N	D	non-P	non-T	0	0	0	NA	130	6	6.491
V	F	non-N	D	non-P	non-T	0.279	0.507	0.550	0.583	6	5	6.981
V	F	N	D	P	non-T	0.223	0.489	0.456	0.648	5	5	6.975
non-V	F	N	D	non-P	non-T	-0.284	0.128	-2.224	0.027	117	5	6.334
V	F	N	D	non-P	T	-0.306	0.106	-2.880	0.004	202	5	6.290
V	Y	N	D	non-P	non-T	-0.453	0.297	-1.524	0.128	12	5	6.265
V	F	N	non-D	non-P	non-T	-0.594	0.492	-1.207	0.228	3	5	6.310
non-V	F	N	non-D	non-P	non-T	-0.219	0.402	-0.545	0.586	7	4	6.304
V	Y	non-N	D	non-P	non-T	-0.308	0.691	-0.446	0.656	2	4	6.054
non-V	Y	N	D	non-P	non-T	-0.361	0.337	-1.073	0.284	13	4	6.194
V	F	N	D	P	T	-0.389	0.256	-1.515	0.130	20	4	6.232
V	F	non-N	D	non-P	T	-0.458	0.206	-2.221	0.027	24	4	6.279
non-V	F	N	D	non-P	T	-0.477	0.110	-4.350	1.67E-05	416	4	6.171
non-V	F	non-N	D	non-P	non-T	-0.503	0.684	-0.735	0.463	6	4	6.097
non-V	F	N	D	P	non-T	-0.511	0.415	-1.234	0.218	7	4	6.405
V	Y	N	D	non-P	T	-0.560	0.211	-2.655	0.008	38	4	6.075
V	Y	N	non-D	non-P	non-T	-0.595	0.690	-0.863	0.388	3	4	4.759
V	F	N	non-D	non-P	T	-0.626	0.350	-1.787	0.075	8	4	5.915
V	Y	N	D	P	non-T	-1.221	0.684	-1.786	0.075	2	4	6.594
non-V	Y	non-N	D	non-P	non-T	-0.361	0.491	-0.734	0.463	4	3	6.380
non-V	Y	N	D	non-P	T	-0.440	0.220	-1.997	0.046	38	3	6.261
non-V	F	N	non-D	non-P	T	-0.588	0.321	-1.831	0.068	11	3	6.530
non-V	F	N	D	P	T	-0.735	0.220	-3.346	0.001	28	3	6.009
non-V	F	non-N	D	non-P	T	-0.911	0.218	-4.172	3.60E-05	32	3	6.029
V	Y	N	non-D	non-P	T	-1.337	0.693	-1.928	0.054	3	3	5.307
non-V	Y	N	non-D	non-P	non-T	-2.103	0.496	-4.242	2.67E-05	4	3	5.078
V	F	non-N	D	P	T	-2.479	0.690	-3.596	3.58E-04	4	3	5.717
non-V	F	non-N	D	P	T	0.368	0.691	0.533	0.594	2	2	7.121
non-V	F	N	non-D	P	T	-0.465	0.407	-1.143	0.254	3	2	6.136
V	Y	non-N	non-D	non-P	T	-0.638	0.687	-0.929	0.353	1	2	5.983
non-V	Y	non-N	D	non-P	T	-0.736	0.493	-1.493	0.136	2	2	5.848
non-V	Y	N	non-D	non-P	T	-1.208	0.403	-2.996	0.003	4	2	5.644
non-V	Y	N	non-D	P	non-T	-1.489	0.692	-2.152	0.032	2	2	5.472
non-V	F	non-N	non-D	non-P	T	-2.323	0.687	-3.383	0.001	1	2	4.347
non-V	Y	N	D	P	T	-3.308	0.689	-4.799	2.15E-06	2	2	4.983
non-V	Y	non-N	D	P	T	-0.566	0.685	-0.827	0.409	1	1	5.881
non-V	Y	non-N	non-D	non-P	T	-1.570	0.689	-2.278	0.023	1	1	5.154
non-V	Y	N	non-D	P	T	-1.686	0.689	-2.447	0.015	1	1	4.651
non-V	Y	non-N	non-D	P	non-T	-5.502	0.686	-8.018	8.70E-15	1	1	1.146
non-V	Y	non-N	non-D	P	T	-1.437	0.688	-2.088	0.037	1	0	5.360

**** Results for combinations with low frequency (n < 20) are shown for descriptive purposes only and should be interpreted cautiously. Particularly, estimates for combinations with n < 5 are extremely unstable and not statistically meaningful.***

Table 4.4: Associations between gt1a HVLP combinations and log_{10} viral load. The table presents the total 40 unique combinations based on the specific presence or absence of the five individual HVLPs in patients' sequences. The linear regression outcomes (estimated beta, standard errors, p values, etc.) and characteristics of these combinations are also shown in the table.

4.4.3 Effects of HVLPs on liver disease

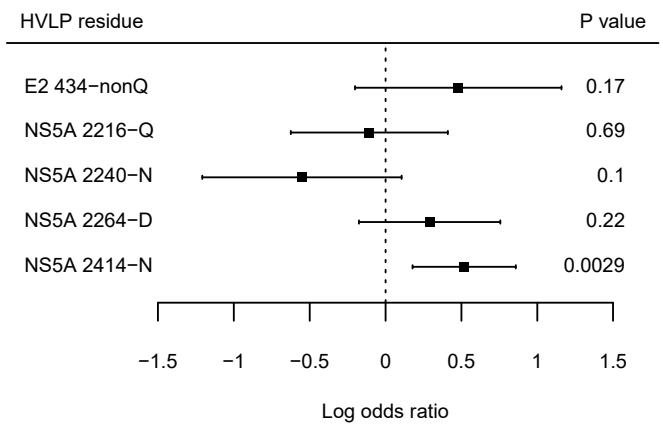
HCV RNA levels have been implicated as a potential contributor to the progression of advanced liver disease, including cirrhosis and hepatocellular carcinoma (HCC), although the evidence is mixed [220-222]. To assess the association between viral load and liver disease in chronically infected patients from our cohort, this work conducted logistic regression to test whether the identified HVLPs associated with the presence of cirrhosis and HCC, respectively.

Effect of gt3a HVLPs on Cirrhosis and HCC

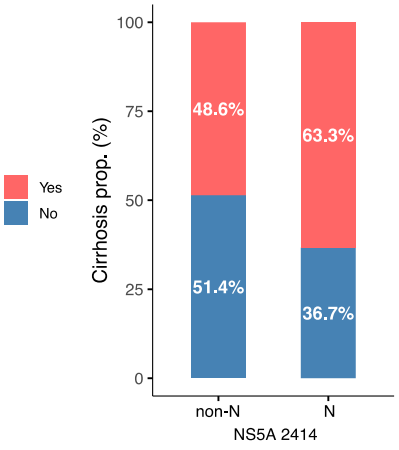
The association between individual HVLP and cirrhosis status was examined using logistic regression model that included sex, age, ethnicity, log_{10} viral load, the first three viral PCs and sequencing center. Among HCV gt3a HVLPs, only asparagine (N, HVLP residue) at NS5A 2414 showed a strong association with an increased prevalence of cirrhosis (P = 0.0029, Figure 4.11A), whereas no significant associations were found for the other HVLPs. The cirrhosis rate for patients carrying asparagine (N) at this position in their viral sequences was 63.3%, compared to 48.6% in those carrying non-asparagine residues (Figure 4.11B). This was consistent with the findings in chapter 3 (Figure 3.3B), where the presence of

a non-serine residue (CAP residue) at the same position was associated with a higher rate of cirrhosis. Next, this study tested whether the total number of HVLPs carried by a patient was associated with cirrhosis status. The association indicated a trend toward increased odds of cirrhosis with a greater number of HVLPs (logOR = 0.165, *P* = 0.071), though this did not reach statistical significance. Among patients carrying three or more HVLPs, cirrhosis prevalence increased gradually as the number of HVLP increased: 42.8% in patients with three HVLPs, 48.9% with four HVLPs and 62% with all five (Figure 4.11C). In contrast, those with 2 or with 0/1 HVLPs showed higher cirrhosis rates (69.4% and 72.7%, respectively), with no statistically significant associations, large likely due to limited sample sizes.

A



B



C

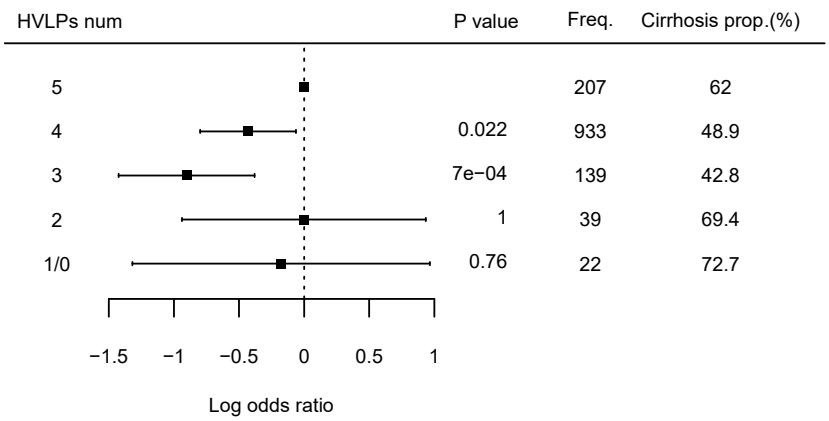


Figure 4.11: Association between cirrhosis status and gt3a HVLPs.
(A)The forest plot displayed the associations between individual HVLP residue and cirrhosis status. The positive log odds ratios represented an association with the presence of cirrhosis, while the negative indicated an association with its absence. Only the HVLP at NS5A 2414 was observed a significant association with the presence of cirrhosis. (B) The cirrhosis proportions across HVLP and non-HVLP group at position NS5A 2414. (C) The forest plot showed association between different numbers of HVLPs and cirrhosis status.
Regarding the forest plots, the black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

Following that, a similar analysis was conducted to explore the association between HVLPs and the presence or absence of HCC. In addition to the covariates previously mentioned, the cirrhosis status was also included to avoid confounding since nearly all HCC cases occur in the context of cirrhosis. The presence of HVLP at NS5A 2216 (Q) was significantly associated with the absence of HCC (P = 0.0063, Figure 4.12A). Specifically, HCC was identified in only 12.1% of patients carrying the Q residue, compared to 26.6% of patients carrying non-Q residues at this site (Figure 4.12B). No significant association was observed between HCC status and the total number of HVLPs carried by patients (logOR = -0.07, P = 0.53). Patients carrying two HVLPs reached the highest HCC prevalence (38.5%), while the lowest prevalence occurred in those with four gt3a HVLPs (9.86%) (Figure 4.12C).

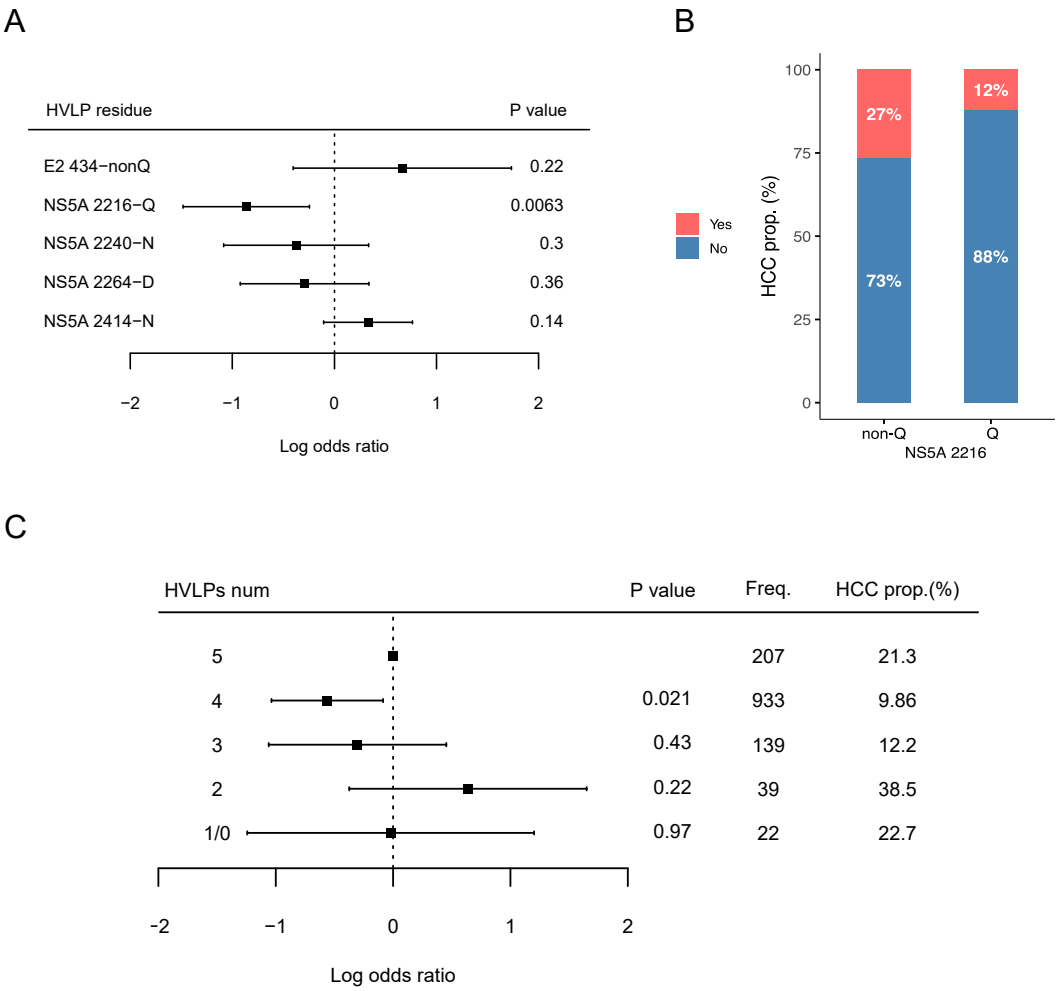


Figure 4.12: Association between HCC status and gt3a HVLPs.

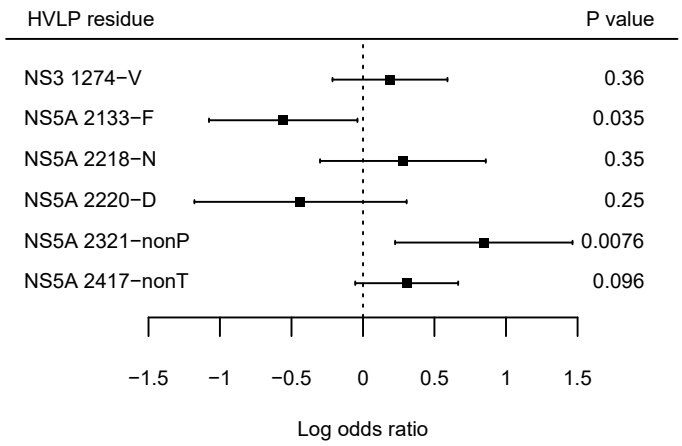
(A)The forest plot displayed the associations between individual HVLP residue and HCC status. The positive log odds ratios represented an association with the presence of HCC, while the negative indicated an association with its absence. Only the HVLP at NS5A 2216 was observed a significant association with the absence of HCC. (B) The HCC proportions across HVLP and non-HVLP group at position NS5A 2216. (C) The forest plot showed association between different numbers of HVLPs and HCC status.

Regarding the forest plots, the black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

Effect of gt1a HVLPs on Cirrhosis and HCC

Following the same approach, an analysis was conducted for cirrhosis status and gt1a HVLPs, incorporating sex, age, the first five viral PCs, $\log_{10}$ viral load, cirrhosis associated two clades and sequencing center as covariates into the logistic regression model. Among gt1a HVLPs, amino acid variants at position NS5A 2321 (P = 0.0048, Figure 4.13A) showed a significant association with cirrhosis status, with patients carrying non-proline (non-P, HVLP residue) exhibiting a higher cirrhosis rate (47.5%) compared to proline carriers (38.4%, Figure 4.13C). The HVLP residue phenylalanine (F) at NS5A 2133 showed a marginally significant negative association with cirrhosis (P = 0.035), with a cirrhosis prevalence of 46.5% compared with 49.7% among patients carrying tyrosine (Y) at this position (Figure 4.13A & 4.13B). The cumulative analysis did not indicate any association between cirrhosis status and the total number of gt1a HVLPs (logOR = 0.13, P = 0.18). Patients with three gt1a HVLPs had the lowest cirrhosis prevalence of 41.2%, whereas the highest prevalence was observed within those carrying fewer than two HVLPs (77.3%, Figure 4.13D). Additionally, no significant associations were observed between HCC status and any gt1a HVLPs (Figure 4.14A & B).

A



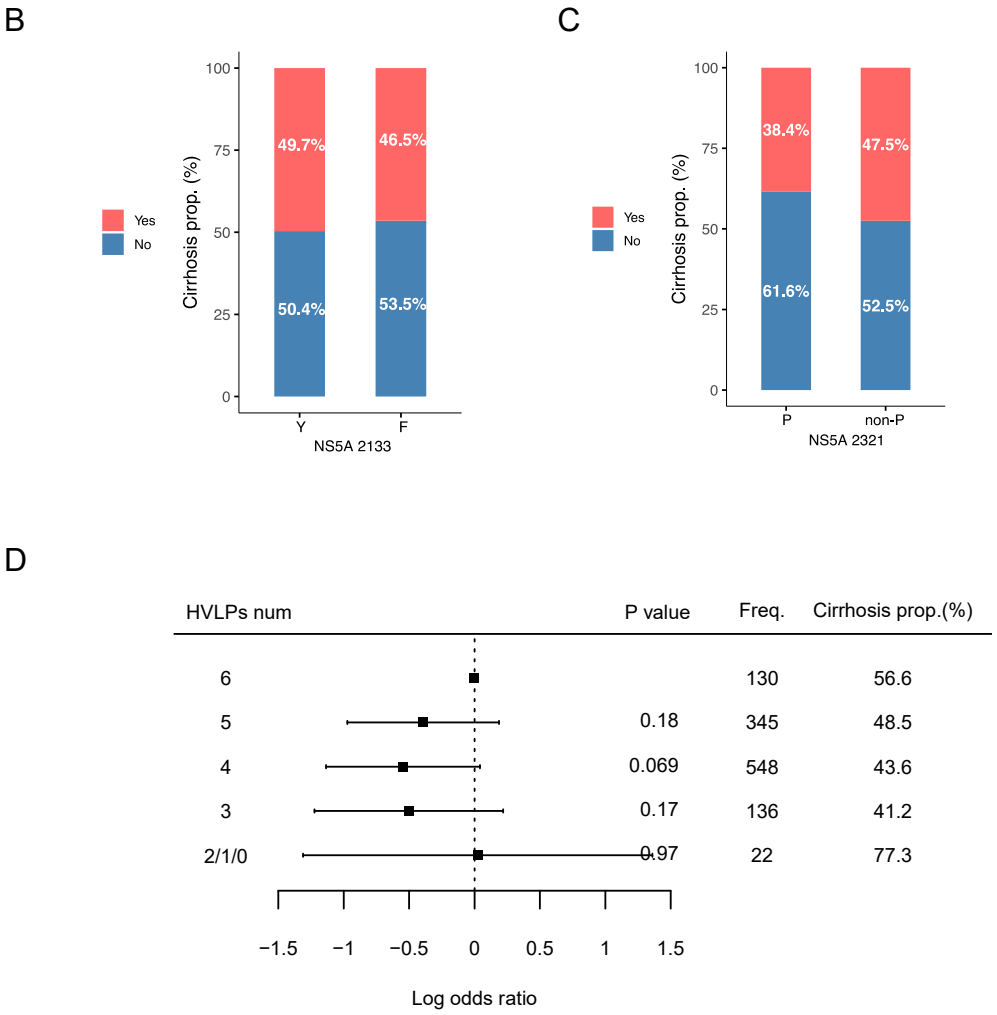
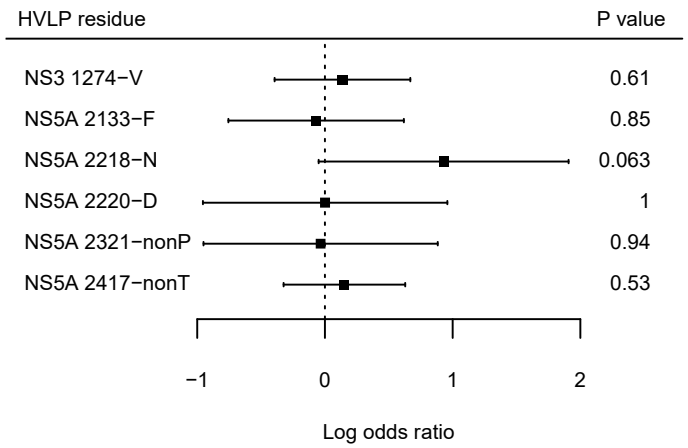


Figure 4.13: Association between cirrhosis status and gt1a HVLPs.
(A)The forest plot displayed the associations between individual HVLP residue and cirrhosis status. HVLP at NS5A 2321 was associated with the presence of cirrhosis, while the HVLP residue at position 2133 was observed association with its absence marginally. (B) and (C) The cirrhosis proportions across HVLP and non-HVLP group at position NS5A 2133 and 2321, respectively. (D) The forest plot showed association between different numbers of gt1a HVLPs and cirrhosis status. Regarding the forest plots, the black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

A



B

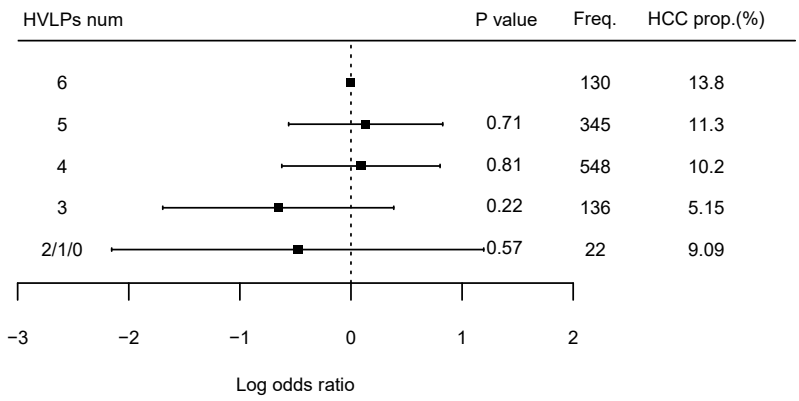


Figure 4.14: Association between HCC status and gt1a HVLPs.
(A)The forest plot displayed the associations between individual HVLP residue and HCC status. (B) The forest plot showed association between different numbers of HVLPs and HCC status. The p values, frequency of each HVLP number group and relative HCC proportion were shown on the right of the table.
The black squares and the lines indicate the estimated effect size (log(odds ratio)) and its 95% confidence interval (CI) estimated from logistic regression.

4.4.4 Impact of HLA alleles on HVLPs

Previous work has indicated that HLA genes play an important role in chronic HCV infection [223, 224]. By participating in cytotoxic T lymphocytes (CTLs/$CD8^+$ T cells)

mediated immune response, HLA genes present HCV epitopes to $CD8^+$ T cells or helper T cells ($CD4^+$ T cells), which drives the emergence of immune escape variants [225]. Because the HVLPs identified in this study represent viral residues associated with replication advantage, their presence or absence may also reflect host immunogenetic selection. Accordingly, the relationship between these HVLPs and HLA alleles was examined to explore whether host immune pressure contributes to their distribution and persistence in chronic HCV infection.

HLA alleles and gt3a HVLPs

Multivariable logistic regression was used to examine the association between gt3a HVLPs and 148 two-field HLA alleles, adjusting for the first three viral PCs, the first five host PCs, sex, age, ethnicity and sequencing center. At an FDR of 5%, only the amino acid variation at position NS5A 2216 showed a significant association with HLA-C*14:02 ($P = 3.06 \times 10^{-4}$, Table 4.5), which lies within a known epitope [146]. No associations were observed between other HVLP polymorphic sites and HLA alleles. Conditional analysis, adjusting for HLA-C*14:02 allele, showed minimal differences in the association between NS5A 2216 and $\log_{10}$ viral load ($\beta_{original} = 0.34$, $P = 1.7 \times 10^{-4}$; model adjusted for HLA-C*14:02 allele, $\beta = 0.32$, $P = 4.1 \times 10^{-4}$, Figure 4.15). No significant association was observed between HLA-C*14:02 and viral load ($P = 0.13$).

HLA	P values	Log OR	SE	FDR
HLA-C*14:02	**3.06×10^{-4}**	**-1.636**	**0.453**	**0.045**
HLA-A*01:01	0.002	0.884	0.287	0.155
HLA-DRB4*01:01	0.008	-0.808	0.304	0.304
HLA-A*29:02	0.011	-0.957	0.376	0.304
HLA-DQA1*05:05	0.012	1.295	0.518	0.304

Table 4.5: The top five associations between HLA alleles and the residue present at position NS5A 2216.

The association at a 5% FDR was highlighted in bold.

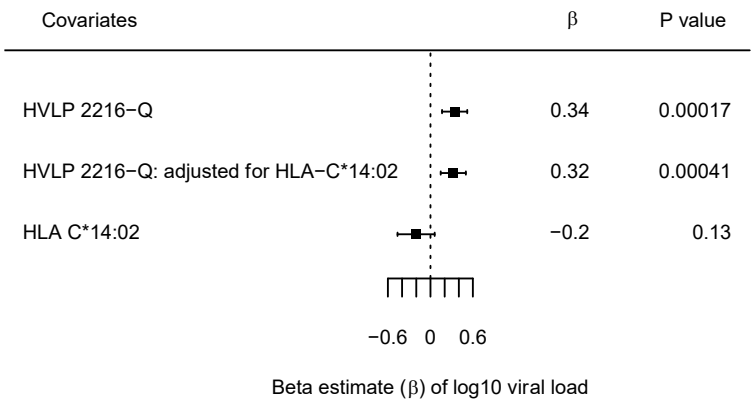


Figure 4.15: Association between log_{10} viral load and variation at NS5A 2216 with or without HLA-C*14:02.

The forest plot indicated whether the impact of HLA-C*14:02 on site1444 was independent. The black squares and the lines indicate the estimated effect size of viral load and its 95% confidence interval estimated from linear regression. The p-values calculated by linear regression are shown on the right of the table.

Interaction between HLA alleles and gt3a HVLPs

The strong association between viral site 2216 and HLA-C*14:02 allele (P = 3.06 $\times$ 10^{-4}) suggested that the impact of this site on viral load may be modulated by host HLA genotype. When stratified by HLA-C*14:02 status (C*14:02- vs. C*14:02+) and by amino acid at 2216 (glutamine vs. non-glutamine), the analysis demonstrated that among individuals lacking HLA-C*14:02, viruses carrying HVLP residue glutamine (Q) at 2216 were associated with significantly higher viral load compared with those with non-Q residues (P = 0.00023; Figure 4.16, blue-dot panel). By contrast, within HLA-C*14:02 carriers, no significant difference was observed between Q and non-Q variants (P = 0.53; Figure 4.16, pink-dot panel). This pattern is consistent with the expected dynamics of HLA-mediated selection: a residue that appears replication-fit in non-carriers can lose its apparent advantage in carriers due to increased immune recognition (see Discussion for potential immunological explanations). However, the formal interaction analysis between NS5A 2216 and HLA-C*14:02 did not reach significance ($P_{\text{site2216-Q : C*14:02}}$ = 0.33), indicating that although glutamine was linked to higher viral load in individuals carrying no HLA-C*14:02 allele, the difference in effect between absence and presence of C*14:02 could not be confirmed statistically, likely due to the small number of HLA-C*14:02 carriers.

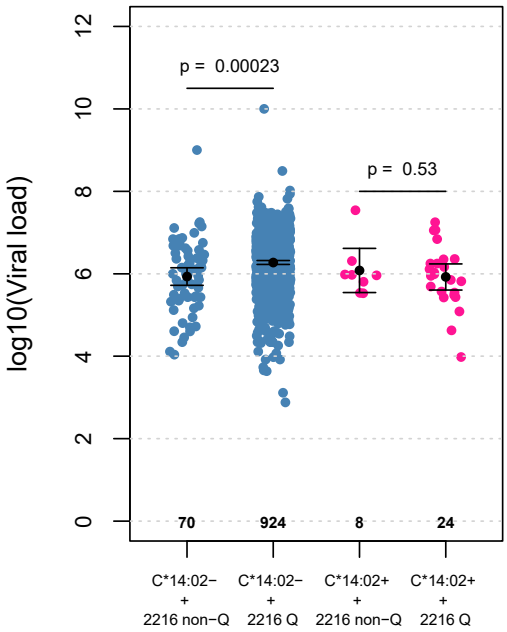


Figure 4.16: Interaction between HLA-C*14:02 and viral amino acid at site 2216.

Log_{10} viral load across the patients infected with virus carrying glutamine (Q) or non-glutamine (non-Q) at site 2216, stratified by the non-carriers (C*14:02-, blue dots) and carriers (C*14:02+, pink dots) of HLA-C*14:02. The number of patients for each interaction group was shown above the x-axis in bold. The mean change in log_{10} viral load was shown in black dots. The solid black lines indicated standard errors. P values were from linear regression.

HLA alleles and gt1a HVLPs

Next, parallel analyses were performed for gt1a HVLPs but restricted to only 535 patients with validated host genome data. At a 5% FDR, HLA-B*27:05 and HLA-B*38:01 alleles showed significant associations with the presence of NS5A 2133 phenylalanine ($P = 2.7 \times 10^{-4}$) and NS5A 2218 asparagine ($P = 5.1 \times 10^{-5}$), respectively (Table 4.6 & 4.7). HLA-B*38:01 has previously been shown to present a gt3a-derived T-cell epitope [146]. For the HLA-B*27:05 allele, the association

between HVLP NS5A 2133 phenylalanine (F) and viral load changed minimally after adjusting for this allele compared with the original analysis ($\beta_{original}$ = 0.43, P = 5.5 × 10^{-5}; adjusted model, β = 0.47, P = 2.4 × 10^{-5}, Figure 4.17A). However, HLA-B*27:05 allele showed no impact on $\log_{10}$ viral load itself (P = 0.17). The beta (β) estimate for the association between NS5A 2218 and $\log_{10}$ viral load was slightly reduced from 0.42 to 0.38 after adjusting for HLA-B*38:01 ($P_{original}$ = 6.9 × 10^{-4}, $P_{adjusted\ model}$ = 2.9 × 10^{-3}, Figure 4.17B). Likewise, HLA-B*38:01 did not indicate association with $\log_{10}$ viral load (P = 0.14).

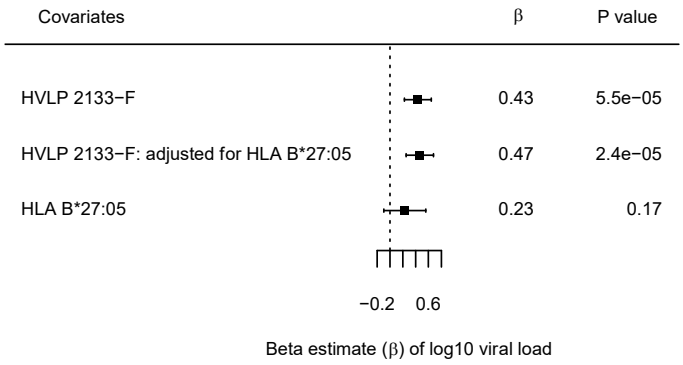
HLA	P values	Log OR	SE	FDR
HLA-B*27:05	**2.73×10^{-4}**	**-1.827**	**0.502**	**0.040**
HLA-C*02:02	0.007	-1.326	0.489	0.494
HLA-A*02:01	0.011	-0.609	0.239	0.542
HLA-C*07:01	0.037	0.693	0.332	0.819
HLA-C*01:02	0.038	-1.049	0.504	0.819

Table 4.6: The top five associations between HLA alleles and the residue present at position NS5A 2133.
The association at 5% FDR was highlighted in bold.

HLA	P values	Log OR	SE	FDR
HLA-B*38:01	**5.08×10^{-5}**	**-2.555**	**0.631**	**0.008**
HLA-C*12:03	0.003	-1.375	0.461	0.145
HLA-A*26:01	0.003	-1.814	0.610	0.145
HLA-DPB1*02:01	0.011	-0.830	0.326	0.351
HLA-DQB1*03:251	0.012	-2.783	1.106	0.351

Table 4.7: The top five associations between HLA alleles and the residue present at position NS5A 2218.
The association at 5% FDR was highlighted in bold.

A



B

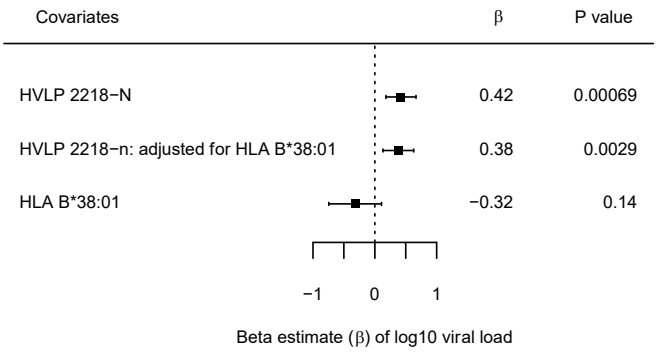


Figure 4.17: Association between log_{10} viral load and variation at site 2133 and site 2218 with different adjustments.

(A) Impact of site 2133 on log_{10} viral load with or without HLA-B*27:05.

(B) Impact of site 2218 on log_{10} viral load with or without HLA-B*38:01.

The black squares and the lines indicate the estimated effect size of viral load and its 95% confidence interval estimated from linear regression. The p-values calculated by linear regression are shown on the right of the table.

Interaction between HLA alleles and gt1a HVLPs

To investigate whether the impact of gt1a HVLPs on viral load was influenced by corresponding HLA alleles, interaction analysis was also performed using the same approach as for gt3a. When stratified by the presence or absence of the

HLA-B*27:05 allele, viruses encoding the HVLP residue phenylalanine (F) at position 2133 were significantly associated with higher $\log_{10}$ viral load compared with those encoding tyrosine (Y) in patients lacking HLA-B*27:05 (P = 1.4 $\times 10^{-5}$; Figure 4.18A, blue-dot panel). In contrast, no significant difference was observed between phenylalanine and tyrosine variants among HLA-B*27:05 carriers (P = 0.74, Figure 4.18A, pink-dot panel). Although this pattern suggested that the replication advantage of HVLP (phenylalanine) at position 2133 may be limited to individuals without HLA-B*27:05, the formal interaction test between position 2133 and HLA-B*27:05 was not statistically significant ($P_{\text{site2133-F : B*27:05}}$ = 0.51), indicating that the observed difference in association may have occurred by chance. The limited number of HLA-B*27:05 carriers (n = 22 vs. non-carriers n = 513) in the study likely reduced statistical power to detect interaction effects; therefore, these findings should be interpreted cautiously and validated in a larger cohort. This mirrors the pattern observed at NS5A 2216 (Figure 4.16), consistent with HLA-mediated modulation of viral fitness in which the replication advantage seen in non-carriers is absent in carriers.

When analysing the interplay between HLA-B*38:01 and variation at site 2218, a marginally significant interaction was observed ($P_{\text{site2218-N : B*38:01}}$ = 0.06). When individuals were grouped by HLA-B*38:01 status, the presence of the HVLP residue asparagine (N) at position 2218 was significantly associated with higher viral load compared with non-asparagine residues among non-carriers of HLA-B*38:01 (P = 0.00067, Figure 4.18B, blue-dot panel). This association disappeared within HLA-B*38:01 carriers (P = 0.11, Figure 4.18B, pink-dot panel). This indicated that asparagine at 2218 may confer a replication advantage only in the

absence of HLA-B*38:01, although this advantage may be diminished in the presence of the allele, potentially due to immune recognition. Similar to HLA-B*27:05, the relatively small number of HLA-B*38:01 carriers (n = 16 vs. non-carriers n = 519) may have introduced statistical bias in detecting interaction effects, so these findings should also be interpreted with caution and further validated in larger, independent cohorts. Taken together with the results for HLA-C*14:02 in gt3a and HLA-B*27:05 in gt1a, this pattern at NS5A 2218 reinforces a recurring trend seen in both genotypes, in which residues that appear replication-fit in non-carriers show diminished advantage in carriers.

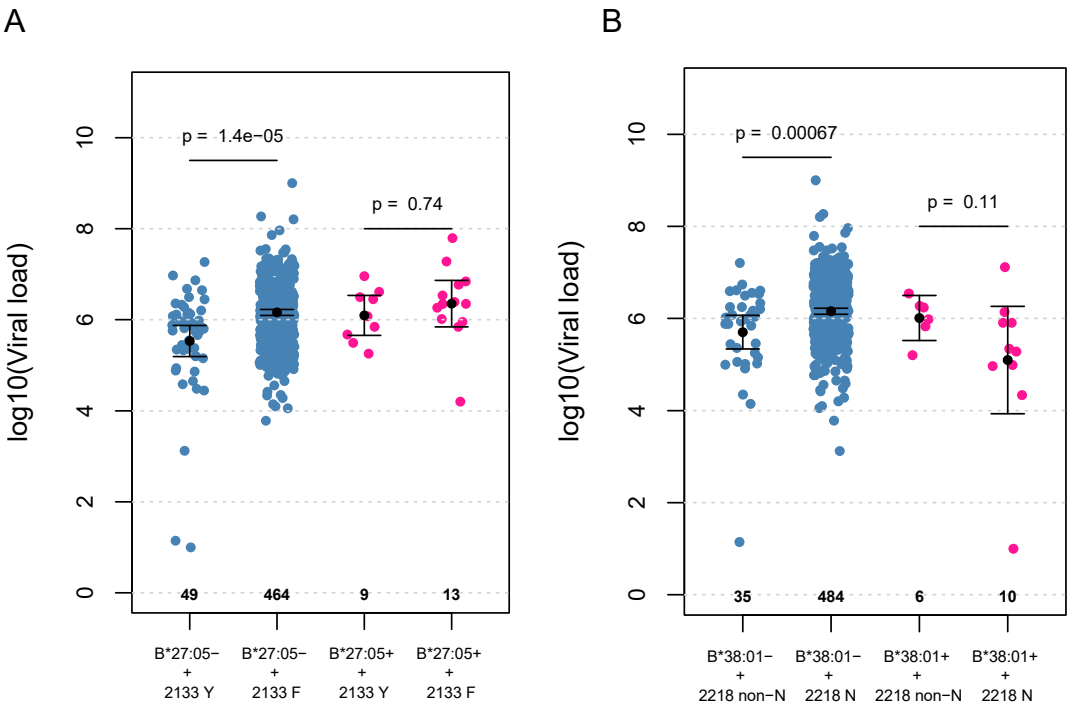


Figure 4.18: Interaction between HLA alleles and gt1a HVLPs.
(A) Association between variation at NS5A 2133 and viral load, stratified by the non-carriers (B*27:05-, blue dots) and carriers (B*27:05+, pink dots) of HLA-B*27:05. (B) Association between variation at NS5A 2218 and viral load, stratified by the non-carriers (B*38:01-, blue dots) and carriers (B*38:01+, pink dots) of HLA-

B*38:01. The number of patients for each interaction group was shown above the x-axis in bold. The mean change in log_{10} viral load was shown in black dots. The solid black lines indicated standard errors. *P* values were from linear regression.

4.5 Discussion

This chapter first comprehensively examined viral amino acid polymorphisms associated with HCV RNA level in a large cohort of patients with HCV gt3a and 1a infections. Full-length viral sequences and host genetic data were obtained from 1028 patients infected with HCV gt3a and 535 with HCV gt1a. Viral load was measured on multiple occasions prior to cohort enrollment, and the highest historical value was used to maximize the number of patients included in the analysis. V-GWAS was used to assess the effects of amino acid variation across the entire polyprotein on viral load, identifying five residues in HCV gt3a and six in gt1a that were significantly associated with viral load; these were referred to as HVLPs. Additionally, three HLA alleles were found to be associated with HVLPs. One of these alleles influenced viral load depending on whether patients carried HVLP residues. These interactions between HLA alleles and viral HVLPs highlight the role of host genetic variation in shaping viral replication dynamics.

Five viral polymorphic sites identified in individuals infected with HCV gt3a were significantly associated with plasma HCV RNA levels. Among these sites, aspartic acid (D) at NS5A 2264, asparagine (N) at NS5A 2240 and NS5A 2414, glutamine (Q) at NS5A 2216 and non-glutamine (non-Q) residues at E2 434 were significantly

associated with higher viral load. These polymorphisms were classified as HVLPs. The distribution of gt3a HVLPs showed enrichment within the NS5A protein, with notable co-occurrence of aspartic acid at NS5A 2264, asparagine at NS5A 2240 and glutamine at NS5A 2216 (Figure 4.5A). The HVLPs identified within patients infected with HCV gt1a viruses – NS5A 2220 aspartic acid (D), NS5A 2133 phenylalanine (F), NS5A 2321 non-proline (non-P), NS5A 2218 asparagine (N), NS3 1274 valine (V) and NS5A 2417 non-threonine (non-T) – were also predominantly localized within the NS5A protein. The NS5A protein is known to play a regulatory role in organizing the HCV RNA replication complex and viral assembly [10]. However, due to the lack of validated crystal structure, the precise mechanisms of how NS5A is involved in replication remains elusive. In this study, the enrichment of HVLPs within NS5A suggests that small changes in amino acids in this protein may impact viral replication. One possible pathway is that the HVLP residues enhance the stability of NS5A replication complex and avoid RNA degradation during assembly process, thereby allowing viral isolates with these residues to replicate more effectively. Likewise, the gt1a HVLP at NS3 1274 may influence the replication directly, as NS3 plays an essential role in the HCV life cycle through its enzymatic activities. Variation at this amino acid position may affect the enzyme efficiency, potentially leading to incomplete separation of newly replicated viral RNA strands from their templates or impaired unwinding of RNA secondary structures [14, 15]. Given these observations, further investigation – such as functional assays or structure modelling – is required to elucidate the mechanism of action of these HVLPs and their impact on viral replication.

Viral site NS5A 2414 is especially noteworthy. A previous study has shown that the presence of different amino acids at NS5A 2414 confers significant changes in mean $\log_{10}$ viral load [146]. This is consistent with the findings in the present work and provides further evidence of the functional relevance of this viral site. Furthermore, in studying the association between HVLPs and liver disease, the presence of asparagine at NS5A position 2414 was notably associated with a higher rate of cirrhosis (Figure 4.11A and B). This aligns with findings in chapter 3, where a non-serine residue at the same position – identified as a cirrhosis associated polymorphism (CAP) – was significantly associated with the cirrhosis prevalence. This association was largely driven by substitution between asparagine and serine at this site (Chapter 3, section 3.4.3, Figure 3.3B and 3.4E). The amino acid variation at this position has also been reported to be associated with the genetic variation in *IFNL4* gene. The impact of *IFNL4* genotype on viral load and cirrhosis seems to be modulated by the presence of specific viral residues at this site (Chapter 3, section 3.4.7, Table 3.8). Taken together, these findings suggest that NS5A 2414 may play a critical role in viral fitness and immune evasion, possibly by altering NS5A's interactions with host antiviral pathways or its role in replication complex assembly and RNA binding. The differential association with *IFNL4* genotypes further implicates this site in shaping the host immune environment, potentially affecting interferon signalling or the cellular antiviral state. Further functional studies are required to investigate how specific substitutions at this site modulate viral replication dynamics and contribute to disease progression.

All patients in this study had chronic HCV infection, indicating that their viral populations had evolved successful adaptation to evade the host immune system.

One potential mechanism for this evasion is the emergence of site-specific mutations under HLA-restricted T-cell pressure, enabling the virus to reduce antigen presentation and T cell recognition. In this chapter, significant associations were observed between HVLPs and HLA alleles, including gt3a HVLP glutamine (Q) at NS5A 2216 with HLA-C*14:02 ($P = 3.06 \times 10^{-4}$), gt1a HVLP phenylalanine (F) at NS5A 2133 with HLA-B*27:05 ($P = 2.7 \times 10^{-4}$), and asparagine (N) at NS5A 2218 with HLA-B*38:01 ($P = 5.1 \times 10^{-5}$). These associations suggest host HLA alleles may influence the replication advantage of the HVLP residues, although definitive evidence remains limited. Among HLA-C*14:02 non-carriers, viruses with glutamine at NS5A 2216 were significantly associated with higher viral load relative to those with non-glutamine residues (Figure 4.17); this association was not observed in carriers of the HLA-C*14:02 allele. Similarly, gt1a HVLP residues, phenylalanine at NS5A 2133 and asparagine at NS5A 2218, were significantly associated with higher viral load compared with alternative residues only in individuals lacking the relevant HLA allele (Figure 4.18). In carriers of the corresponding HLA allele, however, the replication advantage disappeared. Importantly, this functional pattern was paralleled by the distribution of residues across corresponding HLA alleles: in all three cases, the HVLP residue was present at very high frequency in non-carriers (e.g., 93% for 2216Q in C*14:02 non-carriers, 90% for 2133F in B*27:05 non-carriers, and 93% for 2218N in B*38:01 non-carriers), but its frequency was substantially reduced in carriers (75%, 59%, and 63%, respectively; Table 4.8). Such shifts in residue prevalence between carrier and non-carrier groups reinforce the idea that these sites are under HLA-mediated immune pressure.

A

gt3a pos 2216 / HLA-C*14:02	non-Q	Q (HVLP)
C*14:02 –	70 (7%)	924 (93%)
C*14:02 +	8 (25%)	24 (75%)

B

gt1a pos 2133 / HLA-B*27:05	Y	F (HVLP)
B*27:05 –	49 (10%)	464 (90%)
B*27:05 +	9 (41%)	13 (59%)

C

gt1a pos 2218 / HLA-B*38:01	non-N	N (HVLP)
B*38:01 –	35 (7%)	484 (93%)
B*38:01 +	6 (37.5%)	10 (62.5%)

Table 4.8: Amino acid distribution at HLA-associated viral sites.
(A) gt3a NS5A 2216 and HLA-C*14:02; (B) gt1a NS5A 2133 and HLA-B*27:05; (C) gt1a NS5A 2218 and HLA-B*38:01.

This consistent pattern across all three sites aligns with established models of HLA-mediated viral evolution. In the presence of the allele, the HVLP residues may be more efficiently presented to T cells, increasing immune recognition and thus neutralizing their replication advantage. Over time, such pressure is expected to select alternative residues (so-called escape mutations) at these positions in carriers, which reduce T-cell recognition even at a modest replication cost. These

cross-sectional data likely captured a snapshot of this co-evolutionary process, rather than its end point. The consistency of both functional (viral load) and frequency-based (residue distribution) evidence across three independent HLA-viral residue pairs strengthens the case for genuine immune pressure, even though formal interaction terms did not reach statistical significance. This lack of statistical significance probably reflected limited power due to the small number of HLA allele carriers in this cohort (HLA-C*14:02 carriers = 32; HLA-B*27:05 carriers = 22; HLA-B*38:01 carriers = 16). Given all these limitations, future longitudinal studies and larger, more diverse cohorts integrating viral sequences with host immunogenetic data will be essential. Such studies will help to better characterize the dynamics of immune-driven evolution in chronic HCV infections, confirm allele-specific selection pressures, as well as quantify replication-immune trade-offs at each site more precisely.

The immunogenetic framework described above also provides a plausible explanation for why certain HVLPs, gt3a HVLP NS5A 2216Q and gt1a NS5A 2133F, despite being associated with higher viral load, were paradoxically linked to lower prevalence of HCC and cirrhosis, respectively (Figure 4.12A and 4.13A). At first glance, these opposite patterns appear unexpected, because higher viral load levels are generally assumed to cause more severe liver injury. However, the interaction analyses with HLA alleles in this work suggest a more complex relationship. In carriers of HLA-C*14:02 and HLA-B*27:05, the HVLP residues (NS5A 2216Q and NS5A 2133F) are more efficiently presented to CD8+ T cells, thereby eliciting more effective adaptive immune responses. This enhanced recognition could potentially facilitate viral control, attenuate chronic immune activation, and lower the risk of hepatocyte damage, all of which are critical factors

in preventing progression to advanced disease. By contrast, non-HVLP residues (2216-nonQ and 2133Y) may act as escape variants under HLA-mediated immune pressure (Figure 4.16; Figure 4.18A; Table 4.8 A and B), potentially reducing HLA binding affinity and subsequent T-cell recognition. Over the long term, such immune selection in these carriers may partially explain why these specific HVLPs are associated with a lower prevalence of liver disease at the population level. However, as the non-carriers constitute the majority of the patients, this specific allele-restricted pathway is absent among most subjects. Thus, disease progression in these non-carriers is more likely driven by other immune pathways, such as sustained innate immune activation and cytokine-mediated inflammation, which are not directly tied to variation at these specific viral sites. Overall, these findings indicate that the relationship between viral replication advantage and liver disease risk is more complex than a simple linear link between higher viral load and worse outcomes. Replication advantage should therefore be interpreted cautiously as a predictor of clinical outcome, since it could be influenced by a broader network of host-virus interactions.

Chapter 5

Concluding Remarks

5.1 Summary of main findings

Taken together, this thesis systematically explored how viral genetic determinants shape chronic HCV infection outcomes, including antibody responses, disease progression and viral replication. By integrating host immunogenetic information, the work provides potential mechanistic insights into how viral variation may influence clinical outcomes. These findings indicate that viral polymorphisms play a pivotal role in HCV pathogenesis and highlight their potential value for improving prognostic stratification and informing vaccine strategies.

Chapter 2 assessed the non-viral and viral factors influencing antibody responses in 54 patients chronically infected with HCV gt3a. Variation within the E1/E2 envelope proteins, including specific amino acid residues, N-glycosylation motifs, and intra-patient diversity within HVR1, was shown to significantly influence antibody binding and neutralization. These findings are consistent with existing models of immune evasion in HCV, in which envelope variability and glycan shielding modulate exposure of neutralizing epitopes[166, 167]. Importantly, the impact of host factors such as sex, cirrhosis status, treatment history, and *IFNL4* genotype on antibody responses indicates that humoral immunity in chronic infection is shaped by multiple contributing factors rather than by viral or host factors alone.

Beyond antibody responses, the work in chapters 3 and 4 provides insights into how viral genetic determinants influence disease progression and viral replication at a genome-wide level. Cirrhosis associated polymorphisms (CAPs) were

identified across structural and non-structural proteins in both gt3a and gt1a infection, with several mapping to NS5A. The additive association between CAPs and viral load suggests that viral replication levels may contribute to the observed relationship between viral polymorphisms and liver disease progression. The covariation of CAPs also suggests that chronic HCV disease outcomes may reflect the cumulative effects of multiple viral polymorphisms rather than single pathogenic mutations.

The identification of higher viral load associated polymorphisms (HVLPs) further strengthens this link between viral sequence variation and replication capacity. HVLPs were distributed across the viral polyprotein, again with enrichment in NS5A, highlighting the central role of this protein in regulating viral replication and persistence [4, 131]. The overlap between HVLPs and CAPs at specific NS5A residues suggests that viral variants associated with higher replication may also be linked to liver disease risk. Conversely, the inverse associations observed for certain HVLPs with cirrhosis or HCC prevalence indicate that the relationship between replication capacity and clinical outcome is complex and may be modulated by cohort composition, immune pressure, or competing viral fitness costs.

Host immune selection was observed across multiple chapters. Associations between identified viral polymorphisms and *IFNL4* genotype or specific HLA alleles indicate that host immune responses shape viral evolution in ways that can influence viral replication and disease progression. Notably, *IFNL4* non-CC genotype was associated with reduced risk of progression to advanced liver

disease in a viral residue-dependent manner. In addition, the association between CAP residues and specific HLA alleles, together with the observation that several HVLPs conferred replication advantages only in individuals lacking the corresponding HLA alleles, suggest that HLA background may modulate HCV outcomes by shaping patterns of viral genetic variation.

5.2 Future directions

Future work should focus on validating and functionally characterizing the viral polymorphisms identified in this thesis. Replication of CAPs and HVLPs in independent, ethnically and geographically diverse cohorts will be important to confirm the robustness and generalizability of these associations, including their reported links with host HLA alleles. Such validation would help establish whether the impact of those viral variants on HCV outcomes, as well as associated host-virus patterns, is consistently observed across different host and viral populations.

In addition to population-level validation, functional virology experiments will be critical to determine whether these identified CAPs or HVLPs directly modulate viral replication and immune evasion. To this end, targeted site-directed mutagenesis within genotype-matched HCV replicon or infectious clone systems would enable systematic evaluation of the impact of individual and combined polymorphisms on viral fitness in vitro, helping to determine which viral variants directly affect viral behaviour and which represent linked passenger changes

Integrative analyses combining viral genetic variation with host transcriptional responses represent a further important direction. The analysis of liver

transcriptomic data in relation to viral polymorphisms could provide insight into host pathways associated with immune activation, interferon response, or fibrogenesis, thereby contributing to a more mechanistic understanding of disease progression and highlighting candidate biomarkers for improved patient stratification.

Finally, computational protein modelling offers a complementary approach to explore the potential structural and functional consequences of key viral polymorphisms, particularly within proteins such as NS5A for which experimental structural data remain limited. When integrated with functional and immunological data, predictive modelling of wild-type and variant forms of viral proteins may help prioritize residues with the greatest potential relevance to replication or therapeutic targeting.

5.3 Overall contributions and implications

In conclusion, this thesis addresses a key gap in current understanding of how naturally occurring viral genetic variation contributes to immune recognition, viral replication, and liver disease progression across large patient cohorts by applying a genome-wide approach. By moving beyond approaches focused primarily on host factors or restricted to individual viral regions, this work demonstrates that viral genetic variation is an important contributor to clinically relevant outcomes and brings these findings together within a coherent conceptual framework. The identification of viral polymorphisms associated with chronic HCV infection outcomes advances understanding of how viral genetic determinants shape clinically relevant phenotypes and underscores the importance of considering viral determinants alongside host immunogenetic background. Importantly, insights into

envelope variation, glycosylation, and antibody accessibility provide valuable context for future cross-neutralization studies and the rational development of vaccine immunogens. Furthermore, in the era of DAA therapy, where viral clearance does not fully eliminate established cirrhosis or HCC risk – these findings highlight the potential value of viral genetic markers for refining risk stratification. This is especially relevant for informing long-term surveillance after SVR and prioritising treatment in resource-limited settings. Together, these findings contribute to the broader field of viral immunology, offering a framework to improve prognostic accuracy and strengthen the long-term control of chronic HCV infection.

Bibliography

1. Schmidt, A.J., et al., *Prevalence of hepatitis C in a Swiss sample of men who have sex with men: whom to screen for HCV infection?* BMC Public Health, 2014. **14**: p. 3.
2. Hiebert-Suwondo, L., et al., *A 2024 global report on national policies, programmes, and progress towards hepatitis C elimination: findings from 33 hepatitis elimination profiles.* Lancet Gastroenterol Hepatol, 2025. **10**(7): p. 685-700.
3. Lindenbach, B.D. and C.M. Rice, *Unravelling hepatitis C virus replication from genome to function.* Nature, 2005. **436**(7053): p. 933-8.
4. Bode, J.G., E.D. Brenndorfer, and D. Haussinger, *Hepatitis C virus (HCV) employs multiple strategies to subvert the host innate antiviral response.* Biol Chem, 2008. **389**(10): p. 1283-98.
5. Andre, P., et al., *Characterization of low- and very-low-density hepatitis C virus RNA-containing particles.* J Virol, 2002. **76**(14): p. 6919-28.
6. Dustin, L.B., et al., *Hepatitis C virus: life cycle in cells, infection and host response, and analysis of molecular markers influencing the outcome of infection and response to therapy.* Clin Microbiol Infect, 2016. **22**(10): p. 826-832.
7. Zeisel, M.B., D.J. Felmlee, and T.F. Baumert, *Hepatitis C virus entry.* Curr Top Microbiol Immunol, 2013. **369**: p. 87-112.
8. Timpe, J.M., et al., *Hepatitis C virus cell-cell transmission in hepatoma cells in the presence of neutralizing antibodies.* Hepatology, 2008. **47**(1): p. 17-24.
9. Popescu, C.I., et al., *Hepatitis C virus life cycle and lipid metabolism.* Biology (Basel), 2014. **3**(4): p. 892-921.
10. Manns, M.P., et al., *Hepatitis C virus infection.* Nat Rev Dis Primers, 2017. **3**: p. 17006.
11. Honda, M., et al., *A phylogenetically conserved stem-loop structure at the 5' border of the internal ribosome entry site of hepatitis C virus is required for cap-independent viral translation.* J Virol, 1999. **73**(2): p. 1165-74.
12. Niepmann, M., *Hepatitis C virus RNA translation.* Curr Top Microbiol Immunol, 2013. **369**: p. 143-66.
13. Moradpour, D. and F. Penin, *Hepatitis C virus proteins: from structure to function.* Curr Top Microbiol Immunol, 2013. **369**: p. 113-42.
14. Lohmann, V., *Hepatitis C virus RNA replication.* Curr Top Microbiol Immunol, 2013. **369**: p. 167-98.

15. Scheel, T.K. and C.M. Rice, *Understanding the hepatitis C virus life cycle paves the way for highly effective therapies.* Nat Med, 2013. **19**(7): p. 837-49.
16. Gotte, M. and J.J. Feld, *Direct-acting antiviral agents for hepatitis C: structural and mechanistic insights.* Nat Rev Gastroenterol Hepatol, 2016. **13**(6): p. 338-51.
17. Rheault, M., et al., *Elucidating the distinct contributions of miR-122 in the HCV life cycle reveals insights into virion assembly.* Nucleic Acids Res, 2023. **51**(5): p. 2447-2463.
18. Lindenbach, B.D., *Virion assembly and release.* Curr Top Microbiol Immunol, 2013. **369**: p. 199-218.
19. Mbisa, J.L., et al., *Identification of 2 Novel Subtypes of Hepatitis C Virus Genotype 8 and a Potential New Genotype Successfully Treated With Direct Acting Antivirals.* J Infect Dis, 2024. **230**(6): p. e1254-e1262.
20. Galli, A. and J. Bukh, *Mechanisms and Consequences of Genetic Variation in Hepatitis C Virus (HCV).* Curr Top Microbiol Immunol, 2023. **439**: p. 237-264.
21. Petruzziello, A., et al., *Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes.* World J Gastroenterol, 2016. **22**(34): p. 7824-40.
22. Lin, S.K., et al., *Using host genetics to infer the global spread and evolutionary history of HCV subtype 3a.* Virus Evol, 2021. **7**(2): p. veab065.
23. Smith, D.B., et al., *Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource.* Hepatology, 2014. **59**(1): p. 318-27.
24. Murphy, D.G., et al., *Hepatitis C virus genotype 7, a new genotype originating from central Africa.* J Clin Microbiol, 2015. **53**(3): p. 967-72.
25. Borgia, S.M., et al., *Identification of a Novel Hepatitis C Virus Genotype From Punjab, India: Expanding Classification of Hepatitis C Virus Into 8 Genotypes.* J Infect Dis, 2018. **218**(11): p. 1722-1729.
26. Hedskog, C., et al., *Identification of 19 Novel Hepatitis C Virus Subtypes-Further Expanding HCV Classification.* Open Forum Infect Dis, 2019. **6**(3): p. ofz076.
27. *International Committee on Taxonomy of Viruses (ICTV): Hepacivirus C Classification*. 2022; Available from: https://ictv.global/sg_wiki/flaviviridae/hepacivirus.
28. Ogata, N., et al., *Nucleotide sequence and mutation rate of the H strain of hepatitis C virus.* Proc Natl Acad Sci U S A, 1991. **88**(8): p. 3392-6.
29. Okamoto, H., et al., *Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability.* Virology, 1992. **190**(2): p. 894-9.
30. Biebricher, C.K. and M. Eigen, *The error threshold.* Virus Res, 2005. **107**(2): p. 117-27.
31. Chambers, T.J., et al., *Quasispecies heterogeneity within the E1/E2 region as a pretreatment variable during pegylated interferon therapy of chronic hepatitis C virus infection.* J Virol, 2005. **79**(5): p. 3071-83.
32. Laskus, T., et al., *Analysis of hepatitis C virus quasispecies transmission and evolution in patients infected through blood transfusion.* Gastroenterology, 2004. **127**(3): p. 764-76.

33. Smith, D.B., *Evolution of the hypervariable region of hepatitis C virus.* J Viral Hepat, 1999. **6 Suppl 1**: p. 41-6.
34. Labonte, P., et al., *Modulation of hepatitis C virus RNA-dependent RNA polymerase activity by structure-based site-directed mutagenesis.* J Biol Chem, 2002. **277**(41): p. 38838-46.
35. Merani, S., et al., *Effect of Immune Pressure on Hepatitis C Virus Evolution: Insights From a Single-Source Outbreak.* Hepatology, 2011. **53**(2): p. 396-405.
36. Bengsch, B., R. Thimme, and H.E. Blum, *Role of host genetic factors in the outcome of hepatitis C virus infection.* Viruses, 2009. **1**(2): p. 104-125.
37. Dowd, K.A., et al., *Selection pressure from neutralizing antibodies drives sequence evolution during acute infection with hepatitis C virus.* Gastroenterology, 2009. **136**(7): p. 2377-86.
38. Ball, J.K., A.W. Tarr, and J.A. McKeating, *The past, present and future of neutralizing antibodies for hepatitis C virus.* Antiviral Res, 2014. **105**(100): p. 100-11.
39. Echeverria, N., et al., *Hepatitis C virus genetic variability and evolution.* World J Hepatol, 2015. **7**(6): p. 831-45.
40. Simmonds, P., et al., *Impact of virus subtype and host IFNL4 genotype on large-scale RNA structure formation in the genome of hepatitis C virus.* RNA, 2020. **26**(11): p. 1541-1556.
41. Grebely, J., et al., *Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: towards a vaccine.* Lancet Infect Dis, 2012. **12**(5): p. 408-14.
42. Chigbu, D.I., et al., *Hepatitis C Virus Infection: Host(-)Virus Interaction and Mechanisms of Viral Persistence.* Cells, 2019. **8**(4).
43. Kaplan, D.E., *Immunopathogenesis of Hepatitis C Virus Infection.* Gastroenterol Clin North Am, 2015. **44**(4): p. 735-60.
44. Stetson, D.B. and R. Medzhitov, *Type I interferons in host defense.* Immunity, 2006. **25**(3): p. 373-81.
45. Israelow, B., et al., *HepG2 cells mount an effective antiviral interferon-lambda based innate immune response to hepatitis C virus infection.* Hepatology, 2014. **60**(4): p. 1170-9.
46. Randall, G., et al., *Silencing of USP18 potentiates the antiviral activity of interferon against hepatitis C virus infection.* Gastroenterology, 2006. **131**(5): p. 1584-91.
47. Ketscher, L., et al., *Selective inactivation of USP18 isopeptidase activity in vivo enhances ISG15 conjugation and viral resistance.* Proc Natl Acad Sci U S A, 2015. **112**(5): p. 1577-82.
48. Brown, M.G., et al., *Vital involvement of a natural killer cell activation receptor in resistance to viral infection.* Science, 2001. **292**(5518): p. 934-7.
49. Guidotti, L.G. and F.V. Chisari, *Noncytolytic control of viral infections by the innate and adaptive immune response.* Annu Rev Immunol, 2001. **19**: p. 65-91.
50. Xu, Y. and J. Zhong, *Innate immunity against hepatitis C virus.* Curr Opin Immunol, 2016. **42**: p. 98-104.
51. Klenerman, P. and R. Thimme, *T cell responses in hepatitis C: the good, the bad and the unconventional.* Gut, 2012. **61**(8): p. 1226-34.

52. Logvinoff, C., et al., *Neutralizing antibody response during acute and chronic hepatitis C virus infection.* Proc Natl Acad Sci U S A, 2004. **101**(27): p. 10149-54.
53. Yu, M.Y., et al., *Neutralizing antibodies to hepatitis C virus (HCV) in immune globulins derived from anti-HCV-positive plasma.* Proc Natl Acad Sci U S A, 2004. **101**(20): p. 7705-10.
54. Law, M., *Antibody Responses in Hepatitis C Infection.* Cold Spring Harb Perspect Med, 2021. **11**(3).
55. Bhandari, R., et al., *High rate of hepatitis C reinfection following antiviral treatment in the North East England Prisons.* J Viral Hepat, 2020. **27**(4): p. 449-452.
56. Islam, N., et al., *Hepatitis C cross-genotype immunity and implications for vaccine development.* Sci Rep, 2017. **7**(1): p. 12326.
57. Walker, M.R., et al., *Clearance of hepatitis C virus is associated with early and potent but narrowly-directed, Envelope-specific antibodies.* Sci Rep, 2019. **9**(1): p. 13300.
58. Underwood, A.P., et al., *Understanding the Determinants of BnAb Induction in Acute HCV Infection.* Viruses, 2018. **10**(11).
59. Tarr, A.W., et al., *Hepatitis C patient-derived glycoproteins exhibit marked differences in susceptibility to serum neutralizing antibodies: genetic subtype defines antigenic but not neutralization serotype.* J Virol, 2011. **85**(9): p. 4246-57.
60. Kinchen, V.J. and J.R. Bailey, *Defining Breadth of Hepatitis C Virus Neutralization.* Front Immunol, 2018. **9**: p. 1703.
61. Keck, Z.Y., et al., *Mutations in hepatitis C virus E2 located outside the CD81 binding sites lead to escape from broadly neutralizing antibodies but compromise virus infectivity.* J Virol, 2009. **83**(12): p. 6149-60.
62. Kinchen, V.J., et al., *Broadly Neutralizing Antibody Mediated Clearance of Human Hepatitis C Virus Infection.* Cell Host Microbe, 2018. **24**(5): p. 717-730 e5.
63. Swann, R.E., et al., *Broad Anti-Hepatitis C Virus (HCV) Antibody Responses Are Associated with Improved Clinical Disease Parameters in Chronic HCV Infection.* J Virol, 2016. **90**(9): p. 4530-4543.
64. Boisvert, M., et al., *Novel E2 Glycoprotein Tetramer Detects Hepatitis C Virus-Specific Memory B Cells.* J Immunol, 2016. **197**(12): p. 4848-4858.
65. Thimme, R., et al., *Viral and immunological determinants of hepatitis C virus clearance, persistence, and disease.* Proc Natl Acad Sci U S A, 2002. **99**(24): p. 15661-8.
66. Shin, E.C., et al., *Delayed induction, not impaired recruitment, of specific CD8(+) T cells causes the late onset of acute hepatitis C.* Gastroenterology, 2011. **141**(2): p. 686-95, 695 e1.
67. Lechner, F., et al., *Analysis of successful immune responses in persons infected with hepatitis C virus.* J Exp Med, 2000. **191**(9): p. 1499-512.
68. Luxenburger, H., et al., *HCV-Specific T Cell Responses During and After Chronic HCV Infection.* Viruses, 2018. **10**(11).
69. Chen, D.Y., et al., *Hepatitis C virus-specific CD4+ T cell phenotype and function in different infection outcomes.* J Clin Invest, 2020. **130**(2): p. 768-773.

70. Shin, E.C., et al., *The frequency of CD127(+) hepatitis C virus (HCV)-specific T cells but not the expression of exhaustion markers predicts the outcome of acute HCV infection.* J Virol, 2013. **87**(8): p. 4772-7.
71. Golden-Mason, L., et al., *Loss of IL-7 receptor alpha-chain (CD127) expression in acute HCV infection associated with viral persistence.* Hepatology, 2006. **44**(5): p. 1098-109.
72. Dustin, L.B., *Innate and Adaptive Immune Responses in Chronic HCV Infection.* Curr Drug Targets, 2017. **18**(7): p. 826-843.
73. Kurktschiev, P.D., et al., *Dysfunctional CD8+ T cells in hepatitis B and C are characterized by a lack of antigen-specific T-bet induction.* J Exp Med, 2014. **211**(10): p. 2047-59.
74. Schulze zur Wiesch, J., et al., *Broad repertoire of the CD4+ Th cell response in spontaneously controlled hepatitis C virus infection includes dominant and highly promiscuous epitopes.* J Immunol, 2005. **175**(6): p. 3603-13.
75. Eckels, D.D., et al., *Immunobiology of hepatitis C virus (HCV) infection: the role of CD4 T cells in HCV infection.* Immunol Rev, 2000. **174**: p. 90-7.
76. Gerlach, J.T., et al., *Recurrence of hepatitis C virus after loss of virus-specific CD4(+) T-cell response in acute hepatitis C.* Gastroenterology, 1999. **117**(4): p. 933-41.
77. Thimme, R., *T cell immunity to hepatitis C virus: Lessons for a prophylactic vaccine.* J Hepatol, 2021. **74**(1): p. 220-229.
78. Blank, C.U., et al., *Defining 'T cell exhaustion'.* Nat Rev Immunol, 2019. **19**(11): p. 665-674.
79. Wieland, D., et al., *TCF1(+) hepatitis C virus-specific CD8(+) T cells are maintained after cessation of chronic antigen stimulation.* Nat Commun, 2017. **8**: p. 15050.
80. Utzschneider, D.T., et al., *T Cell Factor 1-Expressing Memory-like CD8(+) T Cells Sustain the Immune Response to Chronic Viral Infections.* Immunity, 2016. **45**(2): p. 415-27.
81. Khan, O., et al., *TOX transcriptionally and epigenetically programs CD8(+) T cell exhaustion.* Nature, 2019. **571**(7764): p. 211-218.
82. Alfei, F., et al., *TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection.* Nature, 2019. **571**(7764): p. 265-269.
83. Erickson, A.L., et al., *The outcome of hepatitis C virus infection is predicted by escape mutations in epitopes targeted by cytotoxic T lymphocytes.* Immunity, 2001. **15**(6): p. 883-95.
84. Timm, J., et al., *CD8 epitope escape and reversion in acute HCV infection.* J Exp Med, 2004. **200**(12): p. 1593-604.
85. Nakamoto, N., et al., *Synergistic reversal of intrahepatic HCV-specific CD8 T cell exhaustion by combined PD-1/CTLA-4 blockade.* PLoS Pathog, 2009. **5**(2): p. e1000313.
86. Ackermann, C., et al., *HCV-specific CD4+ T cells of patients with acute and chronic HCV infection display high expression of TIGIT and other co-inhibitory molecules.* Sci Rep, 2019. **9**(1): p. 10624.
87. Cox, A.L., *MEDICINE. Global control of hepatitis C virus.* Science, 2015. **349**(6250): p. 790-1.
88. Midgard, H., et al., *Hepatitis C reinfection after sustained virological response.* J Hepatol, 2016. **64**(5): p. 1020-1026.
89. Al-Khazraji, A., et al., *Identifying Barriers to the Treatment of Chronic Hepatitis C Infection.* Dig Dis, 2020. **38**(1): p. 46-52.

90. Shoukry, N.H., *Hepatitis C Vaccines, Antibodies, and T Cells.* Front Immunol, 2018. **9**: p. 1480.
91. Barnes, E., et al., *Implementation of a controlled human infection model for evaluation of HCV vaccine candidates.* Hepatology, 2023. **77**(5): p. 1757-1772.
92. Li, P., et al., *Engineering of N-glycosylation of hepatitis C virus envelope protein E2 enhances T cell responses for DNA immunization.* Vaccine, 2007. **25**(8): p. 1544-51.
93. Folgori, A., et al., *A T-cell HCV vaccine eliciting effective immunity against heterologous virus challenge in chimpanzees.* Nat Med, 2006. **12**(2): p. 190-7.
94. Kelly, C., et al., *Cross-reactivity of hepatitis C virus specific vaccine-induced T cells at immunodominant epitopes.* Eur J Immunol, 2015. **45**(1): p. 309-16.
95. Barnes, E., et al., *Novel adenovirus-based vaccines induce broad and sustained T cell responses to HCV in man.* Sci Transl Med, 2012. **4**(115): p. 115ra1.
96. Choo, Q.L., et al., *Vaccination of chimpanzees against infection by the hepatitis C virus.* Proc Natl Acad Sci U S A, 1994. **91**(4): p. 1294-8.
97. Meunier, J.C., et al., *Vaccine-induced cross-genotype reactive neutralizing antibodies against hepatitis C virus.* J Infect Dis, 2011. **204**(8): p. 1186-90.
98. Stamataki, Z., et al., *Hepatitis C virus envelope glycoprotein immunization of rodents elicits cross-reactive neutralizing antibodies.* Vaccine, 2007. **25**(45): p. 7773-84.
99. Logan, M., et al., *Native Folding of a Recombinant gpE1/gpE2 Heterodimer Vaccine Antigen from a Precursor Protein Fused with Fc IgG.* J Virol, 2017. **91**(1).
100. Meunier, J.C., et al., *Isolation and characterization of broadly neutralizing human monoclonal antibodies to the e1 glycoprotein of hepatitis C virus.* J Virol, 2008. **82**(2): p. 966-73.
101. Law, J.L., et al., *A hepatitis C virus (HCV) vaccine comprising envelope glycoproteins gpE1/gpE2 derived from a single isolate elicits broad cross-genotype neutralizing antibodies in humans.* PLoS One, 2013. **8**(3): p. e59776.
102. Drane, D., et al., *Priming of CD4+ and CD8+ T cell responses using a HCV core ISCOMATRIX vaccine: a phase I study in healthy volunteers.* Hum Vaccin, 2009. **5**(3): p. 151-7.
103. Swadling, L., et al., *A human vaccine strategy based on chimpanzee adenoviral and MVA vectors that primes, boosts, and sustains functional HCV-specific T cell memory.* Sci Transl Med, 2014. **6**(261): p. 261ra153.
104. Page, K., et al., *Randomized Trial of a Vaccine Regimen to Prevent Chronic HCV Infection.* N Engl J Med, 2021. **384**(6): p. 541-549.
105. Swadling, L., et al., *Highly-Immunogenic Virally-Vectored T-cell Vaccines Cannot Overcome Subversion of the T-cell Response by HCV during Chronic Infection.* Vaccines (Basel), 2016. **4**(3).
106. Cox A.L., P.K., Melia M, Veenhuis R, Massaccesi G, Osburn W, Wagner K, Giudice L, Stein E, Asher AK, Vassilev V, Lin L, Nicosia A, Capone S, Scarselli E, Folgori A, Gorman R, Chang S, Wolff P, Liang T.J, Ghany M, Wierzbicki M, Lum P. , *LB10. A Randomized, Double-Blind, Placebo-Controlled Efficacy Trial of a Vaccine to Prevent Chronic Hepatitis C Virus*

Infection in an at-Risk Population. Open Forum Infect Dis, 2019. **6**(Suppl 2): p. S997.
107. Center, R.J., et al., *Enhancing the antigenicity and immunogenicity of monomeric forms of hepatitis C virus E2 for use as a preventive vaccine.* J Biol Chem, 2020. **295**(21): p. 7179-7192.
108. Donnison, T., et al., *Viral vectored hepatitis C virus vaccines generate pan-genotypic T cell responses to conserved subdominant epitopes.* Vaccine, 2020. **38**(32): p. 5036-5048.
109. Devarbhavi, H., et al., *Global burden of liver disease: 2023 update.* J Hepatol, 2023. **79**(2): p. 516-537.
110. Iredale, J.P., *Models of liver fibrosis: exploring the dynamic nature of inflammation and repair in a solid organ.* J Clin Invest, 2007. **117**(3): p. 539-48.
111. Elpek, G.O., *Cellular and molecular mechanisms in the pathogenesis of liver fibrosis: An update.* World J Gastroenterol, 2014. **20**(23): p. 7260-76.
112. Zhou, W.C., Q.B. Zhang, and L. Qiao, *Pathogenesis of liver cirrhosis.* World J Gastroenterol, 2014. **20**(23): p. 7312-24.
113. Kisseleva, T., et al., *Myofibroblasts revert to an inactive phenotype during regression of liver fibrosis.* Proc Natl Acad Sci U S A, 2012. **109**(24): p. 9448-53.
114. Mederacke, I., et al., *Fate tracing reveals hepatic stellate cells as dominant contributors to liver fibrosis independent of its aetiology.* Nat Commun, 2013. **4**: p. 2823.
115. Roehlen, N., E. Crouchet, and T.F. Baumert, *Liver Fibrosis: Mechanistic Concepts and Therapeutic Perspectives.* Cells, 2020. **9**(4).
116. Ringelhan, M., et al., *The immunology of hepatocellular carcinoma.* Nat Immunol, 2018. **19**(3): p. 222-232.
117. Lawal, G., et al., *The Immunology of Hepatocellular Carcinoma.* Vaccines (Basel), 2021. **9**(10).
118. Gines, P., et al., *Liver cirrhosis.* Lancet, 2021. **398**(10308): p. 1359-1376.
119. Li, J., et al., *HMGB1-induced autophagy facilitates hepatic stellate cells activation: a new pathway in liver fibrosis.* Clin Sci (Lond), 2018. **132**(15): p. 1645-1667.
120. Zhan, S.S., et al., *Phagocytosis of apoptotic bodies by hepatic stellate cells induces NADPH oxidase and is associated with liver fibrosis in vivo.* Hepatology, 2006. **43**(3): p. 435-43.
121. Luedde, T., N. Kaplowitz, and R.F. Schwabe, *Cell death and cell death responses in liver disease: mechanisms and clinical relevance.* Gastroenterology, 2014. **147**(4): p. 765-783 e4.
122. Jacobson Brown, P.M. and M.G. Neuman, *Immunopathogenesis of hepatitis C viral infection: Th1/Th2 responses and the role of cytokines.* Clin Biochem, 2001. **34**(3): p. 167-71.
123. Boltjes, A., et al., *The role of Kupffer cells in hepatitis B and hepatitis C virus infections.* J Hepatol, 2014. **61**(3): p. 660-71.
124. Baskic, D., et al., *Chronic Hepatitis C: Conspectus of immunological events in the course of fibrosis evolution.* PLoS One, 2019. **14**(7): p. e0219508.
125. Ioannou, G.N., et al., *Increased Risk for Hepatocellular Carcinoma Persists Up to 10 Years After HCV Eradication in Patients With Baseline Cirrhosis or High FIB-4 Scores.* Gastroenterology, 2019. **157**(5): p. 1264-1278 e4.

126. Hamdane, N., et al., *HCV-Induced Epigenetic Changes Associated With Liver Cancer Risk Persist After Sustained Virologic Response.* Gastroenterology, 2019. **156**(8): p. 2313-2329 e7.
127. Perez, S., et al., *Hepatitis C virus leaves an epigenetic signature post cure of infection by direct-acting antivirals.* PLoS Genet, 2019. **15**(6): p. e1008181.
128. Yang, T.H., et al., *Genetic Susceptibility to Hepatocellular Carcinoma in Patients with Chronic Hepatitis Virus Infection.* Viruses, 2023. **15**(2).
129. Falleti, E., et al., *PNPLA3 rs738409C/G polymorphism in cirrhosis: relationship with the aetiology of liver disease and hepatocellular carcinoma occurrence.* Liver Int, 2011. **31**(8): p. 1137-43.
130. Fabris, C., et al., *IL-28B rs12979860 C/T allele distribution in patients with liver cirrhosis: role in the course of chronic viral hepatitis and the development of HCC.* J Hepatol, 2011. **54**(4): p. 716-22.
131. El-Shamy, A., et al., *Polymorphisms of the core, NS3, and NS5A proteins of hepatitis C virus genotype 1b associate with development of hepatocellular carcinoma.* Hepatology, 2013. **58**(2): p. 555-63.
132. Chen, X., et al., *Polymorphisms of the NS3 Proteins of Hepatitis C Virus Genotype 1b Are Associated with Liver Cirrhosis.* AIDS Res Hum Retroviruses, 2016. **32**(9): p. 892-4.
133. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.
134. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
135. Cordell, H.J., *Detecting gene-gene interactions that underlie human diseases.* Nat Rev Genet, 2009. **10**(6): p. 392-404.
136. Thomas, D., *Gene--environment-wide association studies: emerging approaches.* Nat Rev Genet, 2010. **11**(4): p. 259-72.
137. Davila, S. and M.L. Hibberd, *Genome-wide association studies are coming for human infectious diseases.* Genome Med, 2009. **1**(2): p. 19.
138. Power, R.A., J. Parkhill, and T. de Oliveira, *Microbial genome-wide association studies: lessons from human GWAS.* Nat Rev Genet, 2017. **18**(1): p. 41-50.
139. Young, B.C., et al., *Panton-Valentine leucocidin is the key determinant of pyomyositis in a bacterial GWAS.* Elife, 2019. **8**.
140. Lees, J.A., et al., *Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis.* Nat Commun, 2019. **10**(1): p. 2176.
141. Davies, M.R., et al., *Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics.* Nat Genet, 2019. **51**(6): p. 1035-1043.
142. Power, R.A., J. Parkhill, and T. de Oliveira, *Microbial genome-wide association studies: lessons from human GWAS.* Nature Reviews Genetics, 2017. **18**(1): p. 41-50.
143. San, J.E., et al., *Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls.* Frontiers in Microbiology, 2020. **10**.

144. Bartha, I., et al., *A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control.* Elife, 2013. **2**.
145. Power, R.A., et al., *Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance.* PLoS One, 2016. **11**(9): p. e0163746.
146. Ansari, M.A., et al., *Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus.* Nat Genet, 2017. **49**(5): p. 666-673.
147. Smith, D.A., et al., *Viral genome wide association study identifies novel hepatitis C virus polymorphisms associated with sofosbuvir treatment failure.* Nat Commun, 2021. **12**(1): p. 6105.
148. Cornberg, M. and M.P. Manns, *The curing regimens of HCV: A SWOT analysis.* Antivir Ther, 2022. **27**(1-2): p. 13596535211072672.
149. Duncan, J.D., et al., *Hepatitis C Virus Vaccine: Challenges and Prospects.* Vaccines (Basel), 2020. **8**(1).
150. Stamataki, Z., et al., *Hepatitis C virus entry and neutralization.* Clin Liver Dis, 2008. **12**(3): p. 693-712, x.
151. Dreux, M., et al., *Receptor complementation and mutagenesis reveal SR-BI as an essential HCV entry factor and functionally imply its intra- and extra-cellular domains.* PLoS Pathog, 2009. **5**(2): p. e1000310.
152. Pileri, P., et al., *Binding of hepatitis C virus to CD81.* Science, 1998. **282**(5390): p. 938-41.
153. Flint, M. and J.A. McKeating, *The role of the hepatitis C virus glycoproteins in infection.* Rev Med Virol, 2000. **10**(2): p. 101-17.
154. Liu, S., et al., *Tight junction proteins claudin-1 and occludin control hepatitis C virus entry and are downregulated during infection to prevent superinfection.* J Virol, 2009. **83**(4): p. 2011-4.
155. Douam, F., et al., *Critical interaction between E1 and E2 glycoproteins determines binding and fusion properties of hepatitis C virus during cell entry.* Hepatology, 2014. **59**(3): p. 776-88.
156. Torrents de la Pena, A., et al., *Structure of the hepatitis C virus E1E2 glycoprotein complex.* Science, 2022. **378**(6617): p. 263-269.
157. Banda, D.H., et al., *A central hydrophobic E1 region controls the pH range of hepatitis C virus membrane fusion and susceptibility to fusion inhibitors.* J Hepatol, 2019. **70**(6): p. 1082-1092.
158. Augestad, E.H., et al., *The hepatitis C virus envelope protein complex is a dimer of heterodimers.* Nature, 2024. **633**(8030): p. 704-709.
159. Kong, L., et al., *Hepatitis C virus E2 envelope glycoprotein core structure.* Science, 2013. **342**(6162): p. 1090-4.
160. Bartosch, B., et al., *Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor.* J Biol Chem, 2003. **278**(43): p. 41624-30.
161. Tarr, A.W., et al., *Immunization with a synthetic consensus hepatitis C virus E2 glycoprotein ectodomain elicits virus-neutralizing antibodies.* Antiviral Res, 2018. **160**: p. 25-37.
162. Alzua, G.P., et al., *Inactivated genotype 1a, 2a and 3a HCV vaccine candidates induced broadly neutralising antibodies in mice.* Gut, 2023. **72**(3): p. 560-572.

163. Sliepen, K., et al., *Induction of cross-neutralizing antibodies by a permuted hepatitis C virus glycoprotein nanoparticle vaccine candidate.* Nat Commun, 2022. **13**(1): p. 7271.
164. Spiro, R.G., *Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds.* Glycobiology, 2002. **12**(4): p. 43R-56R.
165. Goffard, A., et al., *Role of N-linked glycans in the functions of hepatitis C virus envelope glycoproteins.* J Virol, 2005. **79**(13): p. 8400-9.
166. Lavie, M., X. Hanoulle, and J. Dubuisson, *Glycan Shielding and Modulation of Hepatitis C Virus Neutralizing Antibodies.* Frontiers in Immunology, 2018. **9**.
167. Alzua, G.P., et al., *Identification of novel neutralizing determinants for protection against HCV.* Hepatology, 2023. **77**(3): p. 982-996.
168. Ge, D., et al., *Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance.* Nature, 2009. **461**(7262): p. 399-401.
169. Thomas, D.L., et al., *Genetic variation in IL28B and spontaneous clearance of hepatitis C virus.* Nature, 2009. **461**(7265): p. 798-801.
170. Terczynska-Dyla, E., et al., *Reduced IFNlambda4 activity is associated with improved HCV clearance and reduced expression of interferon-stimulated genes.* Nat Commun, 2014. **5**: p. 5699.
171. Eslam, M., et al., *IFN-lambda3, not IFN-lambda4, likely mediates IFNL3-IFNL4 haplotype-dependent hepatic inflammation and fibrosis.* Nat Genet, 2017. **49**(5): p. 795-800.
172. Huang, J., et al., *The Associations of HLA-A*02:01 and DRB1*11:01 with Hepatitis C Virus Spontaneous Clearance Are Independent of IL28B in the Chinese Population.* Sci Rep, 2016. **6**: p. 31485.
173. Foster, G.R., et al., *Efficacy of sofosbuvir plus ribavirin with or without peginterferon-alfa in patients with hepatitis C virus genotype 3 infection and treatment-experienced patients with cirrhosis and hepatitis C virus genotype 2 infection.* Gastroenterology, 2015. **149**(6): p. 1462-70.
174. Owsianka, A., et al., *Monoclonal antibody AP33 defines a broadly neutralizing epitope on the hepatitis C virus E2 envelope glycoprotein.* J Virol, 2005. **79**(17): p. 11095-104.
175. Urbanowicz, R.A., J.K. Ball, and A.W. Tarr, *Cloning and Analysis of Authentic Patient-Derived HCV E1/E2 Glycoproteins.* Methods Mol Biol, 2019. **1911**: p. 275-294.
176. Bailey, J.R., et al., *Standardized Method for the Study of Antibody Neutralization of HCV Pseudoparticles (HCVpp).* Methods Mol Biol, 2019. **1911**: p. 441-450.
177. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
178. Paradis, E. and K. Schliep, *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.* Bioinformatics, 2019. **35**(3): p. 526-528.
179. Revell, L.J., *phytools: an R package for phylogenetic comparative biology (and other things).* Methods in ecology and evolution, 2012(2): p. 217-223.
180. Yariv, B., et al., *Using evolutionary data to make sense of macromolecules with a "face-lifted" ConSurf.* Protein Sci, 2023. **32**(3): p. e4582.

181. Edwards, V.C., et al., *The role of neutralizing antibodies in hepatitis C virus infection.* J Gen Virol, 2012. **93**(Pt 1): p. 1-19.
182. Kumar, A., et al., *Regions of hepatitis C virus E2 required for membrane association.* Nat Commun, 2023. **14**(1): p. 433.
183. Clayton, R.F., et al., *Analysis of antigenicity and topology of E2 glycoprotein present on recombinant hepatitis C virus-like particles.* J Virol, 2002. **76**(15): p. 7672-82.
184. Tzarum, N., I.A. Wilson, and M. Law, *The Neutralizing Face of Hepatitis C Virus E2 Envelope Glycoprotein.* Front Immunol, 2018. **9**: p. 1315.
185. Owsianka, A.M., et al., *Identification of conserved residues in the E2 envelope glycoprotein of the hepatitis C virus that are critical for CD81 binding.* J Virol, 2006. **80**(17): p. 8695-704.
186. Iacob, R.E., et al., *Structural elucidation of critical residues involved in binding of human monoclonal antibodies to hepatitis C virus E2 envelope glycoprotein.* Biochim Biophys Acta, 2008. **1784**(3): p. 530-42.
187. Pfaff-Kilgore, J.M., et al., *Sites of vulnerability in HCV E1E2 identified by comprehensive functional screening.* Cell Rep, 2022. **39**(8): p. 110859.
188. Helle, F., et al., *The neutralizing activity of anti-hepatitis C virus antibodies is modulated by specific glycans on the E2 envelope protein.* J Virol, 2007. **81**(15): p. 8101-11.
189. Breitling, J. and M. Aebi, *N-linked protein glycosylation in the endoplasmic reticulum.* Cold Spring Harb Perspect Biol, 2013. **5**(8): p. a013359.
190. Shakin-Eshleman, S.H., S.L. Spitalnik, and L. Kasturi, *The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency.* J Biol Chem, 1996. **271**(11): p. 6363-6.
191. Kasturi, L., H. Chen, and S.H. Shakin-Eshleman, *Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors.* Biochem J, 1997. **323 (Pt 2)**(Pt 2): p. 415-9.
192. Fearns, R. and J. Deval, *New antiviral approaches for respiratory syncytial virus and other mononegaviruses: Inhibiting the RNA polymerase.* Antiviral Res, 2016. **134**: p. 63-76.
193. Ansari, M.A., et al., *Interferon lambda 4 impacts the genetic diversity of hepatitis C virus.* Elife, 2019. **8**.
194. Chen, Q., et al., *Interferon lambda 4 impairs hepatitis C viral antigen presentation and attenuates T cell responses.* Nat Commun, 2021. **12**(1): p. 4882.
195. D'Amico, G., et al., *New concepts on the clinical course and stratification of compensated and decompensated cirrhosis.* Hepatol Int, 2018. **12**(Suppl 1): p. 34-43.
196. Llovet, J.M., et al., *Hepatocellular carcinoma.* Nat Rev Dis Primers, 2016. **2**: p. 16018.
197. Baumert, T.F., et al., *Status of Direct-Acting Antiviral Therapy for Hepatitis C Virus Infection and Remaining Challenges.* Gastroenterology, 2019. **156**(2): p. 431-445.
198. van der Meer, A.J., et al., *Association between sustained virological response and all-cause mortality among patients with chronic hepatitis C and advanced hepatic fibrosis.* JAMA, 2012. **308**(24): p. 2584-93.

199. Aleman, S., et al., *A risk for hepatocellular carcinoma persists long-term after sustained virologic response in patients with hepatitis C-associated liver cirrhosis.* Clin Infect Dis, 2013. **57**(2): p. 230-6.
200. Umemura, T., et al., *Investigation of the Effect of KIR-HLA Pairs on Hepatocellular Carcinoma in Hepatitis C Virus Cirrhotic Patients.* Cancers (Basel), 2021. **13**(13).
201. Kobayashi, M., et al., *Influence of amino-acid polymorphism in the core protein on progression of liver disease in patients infected with hepatitis C virus genotype 1b.* J Med Virol, 2010. **82**(1): p. 41-8.
202. McLauchlan, J., et al., *Cohort Profile: The Hepatitis C Virus (HCV) Research UK Clinical Database and Biobank.* Int J Epidemiol, 2017. **46**(5): p. 1391-1391h.
203. Cooke, G.S., et al., *Strategic treatment optimization for HCV (STOPHCV1): a randomised controlled trial of ultrashort duration therapy for chronic hepatitis C.* Wellcome Open Res, 2021. **6**: p. 93.
204. Gogarten, S.M., et al., *Genetic association testing using the GENESIS R/Bioconductor package.* Bioinformatics, 2019. **35**(24): p. 5346-5348.
205. Zheng, X., et al., *A high-performance computing toolset for relatedness and principal component analysis of SNP data.* Bioinformatics, 2012. **28**(24): p. 3326-8.
206. Minh, B.Q., et al., *IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.* Mol Biol Evol, 2020. **37**(5): p. 1530-1534.
207. Ansari, M.A. and X. Didelot, *Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree.* Genetics, 2016. **204**(1): p. 89-98.
208. Alric, L., et al., *Study of host- and virus-related factors associated with spontaneous hepatitis C virus clearance.* Tissue Antigens, 2000. **56**(2): p. 154-8.
209. Isaguliants, M.G. and N.N. Ozeretskovskaya, *Host background factors contributing to hepatitis C virus clearance.* Curr Pharm Biotechnol, 2003. **4**(3): p. 185-93.
210. Singh, R., et al., *A comparative review of HLA associations with hepatitis B and C viral infections across global populations.* World J Gastroenterol, 2007. **13**(12): p. 1770-87.
211. Khakoo, S.I., et al., *HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection.* Science, 2004. **305**(5685): p. 872-4.
212. Yindom, L.M., et al., *Influence of HLA class I and HLA-KIR compound genotypes on HIV-2 infection and markers of disease progression in a Manjako community in West Africa.* J Virol, 2010. **84**(16): p. 8202-8.
213. Mohlenberg, M., T.R. O'Brien, and R. Hartmann, *The role of IFNL4 in liver inflammation and progression of fibrosis.* Genes Immun, 2022. **23**(3-4): p. 111-117.
214. Neumann-Haefelin, C., et al., *Analysis of the evolutionary forces in an immunodominant CD8 epitope in hepatitis C virus at a population level.* Journal of Virology, 2008. **82**(7): p. 3438-3451.
215. Timm, J., et al., *Human leukocyte antigen-associated sequence polymorphisms in hepatitis C virus reveal reproducible immune responses and constraints on viral evolution.* Hepatology, 2007. **46**(2): p. 339-349.
216. in *WHO Guidelines on Hepatitis B and C Testing*. 2017: Geneva.

217. Peeling, R.W., et al., *The future of viral hepatitis testing: innovations in testing technologies and approaches.* BMC Infect Dis, 2017. **17**(Suppl 1): p. 699.
218. Mangia, A., et al., *HLA class II favors clearance of HCV infection and progression of the chronic liver damage.* J Hepatol, 1999. **30**(6): p. 984-9.
219. McKiernan, S.M., et al., *Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source.* Hepatology, 2004. **40**(1): p. 108-14.
220. Persico, M., et al., *Hepatitis C virus carriers with persistently normal ALT levels: biological peculiarities and update of the natural history of liver disease at 10 years.* J Viral Hepat, 2006. **13**(5): p. 290-6.
221. Noh, R., et al., *Clinical Impact of Viral Load on the Development of Hepatocellular Carcinoma and Liver-Related Mortality in Patients with Hepatitis C Virus Infection.* Gastroenterol Res Pract, 2016. **2016**: p. 7476231.
222. Freeman, A.J., et al., *Predicting progression to cirrhosis in chronic hepatitis C virus infection.* J Viral Hepat, 2003. **10**(4): p. 285-93.
223. Zavaglia, C., et al., *Association between HLA class II alleles and protection from or susceptibility to chronic hepatitis C.* J Hepatol, 1998. **28**(1): p. 1-7.
224. Thio, C.L., et al., *HLA-Cw*04 and hepatitis C virus persistence.* J Virol, 2002. **76**(10): p. 4792-7.
225. Heim, M.H. and R. Thimme, *Innate and adaptive immune responses in HCV infections.* J Hepatol, 2014. **61**(1 Suppl): p. S14-25.

Appendix A

Location	H77 num	CAP/HVLP	Associated Phenotype	AA seq (DBN3acc)	DBN3acc mutants
E1	375	L	Cirrhosis	QGNWAKVAII**M**IMFSGVDATT	M to L
E2	401	S	Cirrhosis	VVARGVRGLS**S**LLSPGPKQKL	S to G
NS3	1444	Y	Cirrhosis	VCATDALMTG**F**TGDFDSVIDC	F to Y
NS5A	2057	non-Y (H)	Cirrhosis	AGPRTCANMW**Y**GTFPINEYTT	Y to H
NS5A	2414	N/N	Cirrhosis/ VL	SCDSWSTVSD**S**EEQRVVCCSM	S to N
E2	434	non-Q (N)	VL	STALNCNDSL**N**TGFIAGLIYL	N to Q
NS5A	2216	Q	VL	LSAPSLKATC**Q**THRPHPDAEL	Q to G
NS5A	2240	N	VL	WRQEMGS**N**ITRVESETKVV	N to D
NS5A	2264	D	VL	SFEPLRAEI**D**DAELSVAAE	D to N

Table A.1: Engineered mutants corresponding to CAPs and HVLPs in HCV genotype 3a.

Location	H77 num	CAP/HVLP	Associated Phenotype	AA seq (TNcc_JX993348.1)	TNcc mutants
NS2	938	I	Cirrhosis	MAGGHYVQMA**I**IKLGALTGTY	I to V
NS5A	2036	A	Cirrhosis	HTRCHCGAEI**T**GHVKNGTMRI	T to A
NS5A	2321	S/non-P(S)	Cirrhosis/ VL	VHGCPLPPPQ**S**PPVPPPRKKR	S to P
NS3	1274	V	VL	FGAYMSKAHG**V**DPNIRTGVRT	V to I
NS5A	2133	F	VL	TELDGVRLHR**F**APPCKPLLRE	F to Y
NS5A	2218	N	VL	APSLRATCTT**N**HDSPDAELIE	N to K
NS5A	2220	D	VL	APSLRATCTTNH**D**SPDAELIE	D to N
NS5A	2417	non-T(A)	VL	SWSTVSSEAD**K**EDVVCCSMSY	K to T

Table A.2: Engineered mutants corresponding to CAPs and HVLPs in HCV genotype 1a

Appendix B

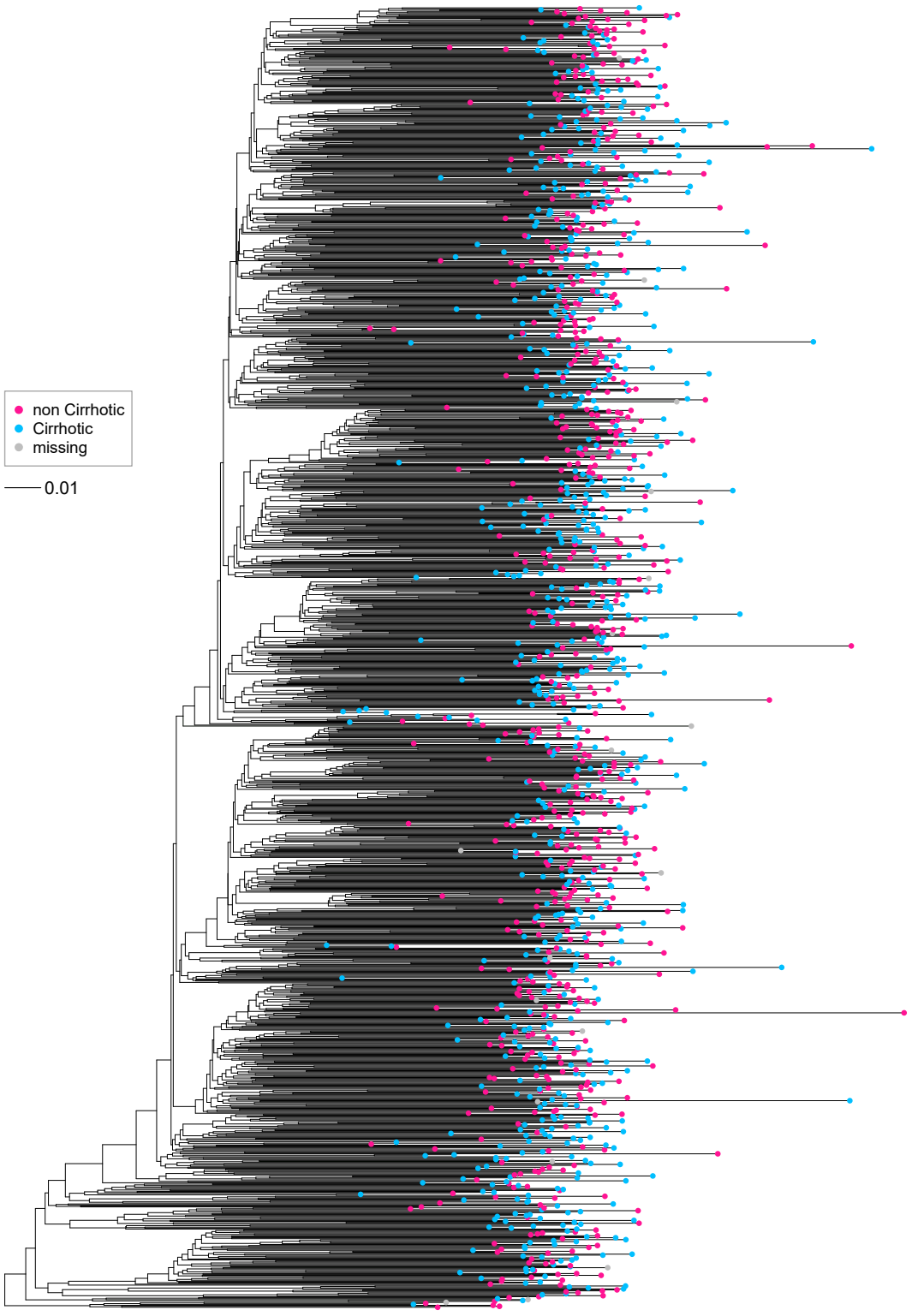


Figure B.1: Distribution of Cirrhosis status (presence vs. absence) across the HCV gt3a phylogeny.

The maximum likelihood phylogenetic tree was constructed using whole HCV gt3a genome sequences using IQ-TREE. The colour of the tips shows if the patient diagnosed with cirrhosis (blue) or not (pink). Line thickness and colour are proportional to the posterior probability of a change point on the branch. Green branches indicate an extreme high confidence change point with a high Bayes factor calculated, identifying clades where the distribution of cirrhosis significantly diverges from the rest of the tree. This tree did not show any lineages associated with cirrhosis distribution. The change point analysis conducted using treeBreaker. The scale bar represents 0.01 nucleotide substitutions per site.

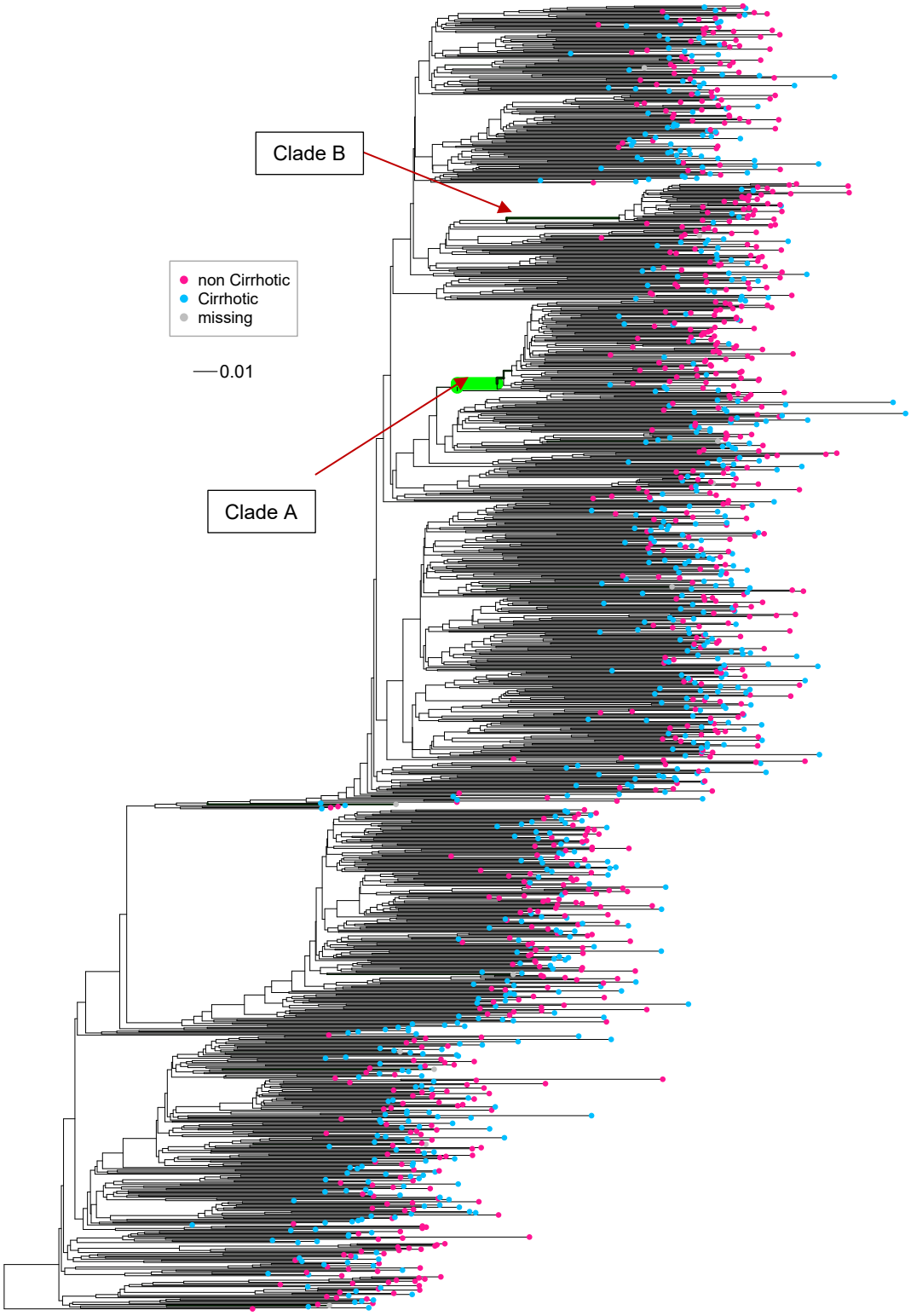

Clade B
non Cirrhotic
Cirrhotic
missing
0.01
Clade A

Figure B.2: Distribution of Cirrhosis status (presence vs. absence) across the HCV gt1a phylogeny.

The maximum likelihood phylogenetic tree was constructed using whole HCV gt1a genome sequences using IQ-TREE. The colour of the tips shows if the patient diagnosed with cirrhosis (blue) or not (pink). Line thickness and colour are proportional to the posterior probability of a change point on the branch. Green branches indicate an extreme high confidence change point with a high Bayes factor calculated, identifying clades where the distribution of cirrhosis significantly diverges from the rest of the tree. The change point analysis conducted using treeBreaker. The scale bar represents 0.01 nucleotide substitutions per site. Clade A and Clade B were lineages associated with lower cirrhosis rate. Expanded views of the two clades are shown in Figure 3.1.

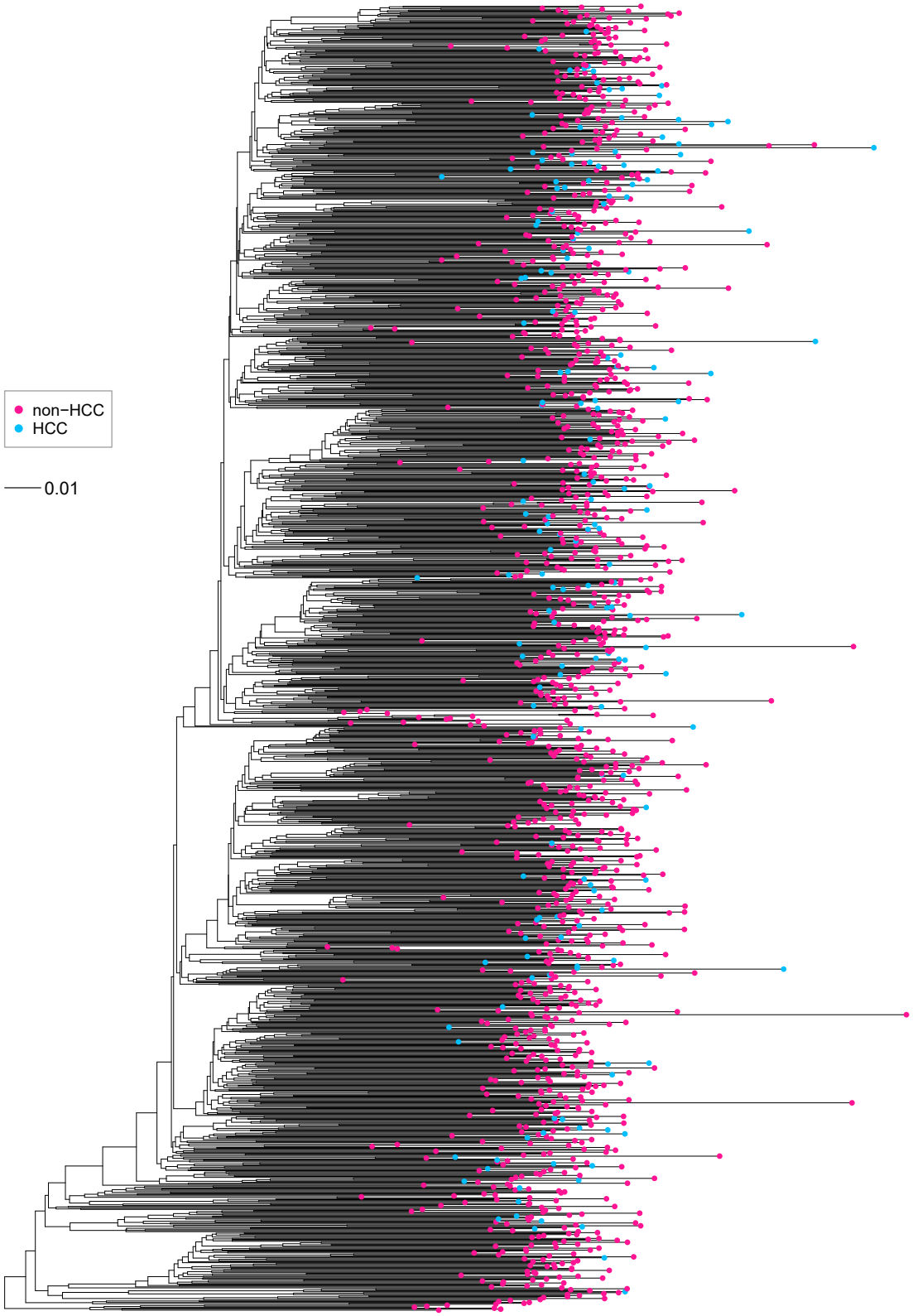

non−HCC
HCC
0.01

Figure B.3: Distribution of HCC status (presence vs. absence) across the HCV gt3a phylogeny.

The maximum likelihood phylogenetic tree was constructed using whole HCV gt3a genome sequences using IQ-TREE. The colour of the tips shows if the patient diagnosed with HCC (blue) or not (pink). Line thickness and colour are proportional to the posterior probability of a change point on the branch. Green branches indicate an extreme high confidence change point with a high Bayes factor calculated, identifying clades where the distribution of HCC significantly diverges from the rest of the tree. This tree did not identify any lineages associated with HCC distribution. The change point analysis conducted using treeBreaker. The scale bar represents 0.01 nucleotide substitutions per site.

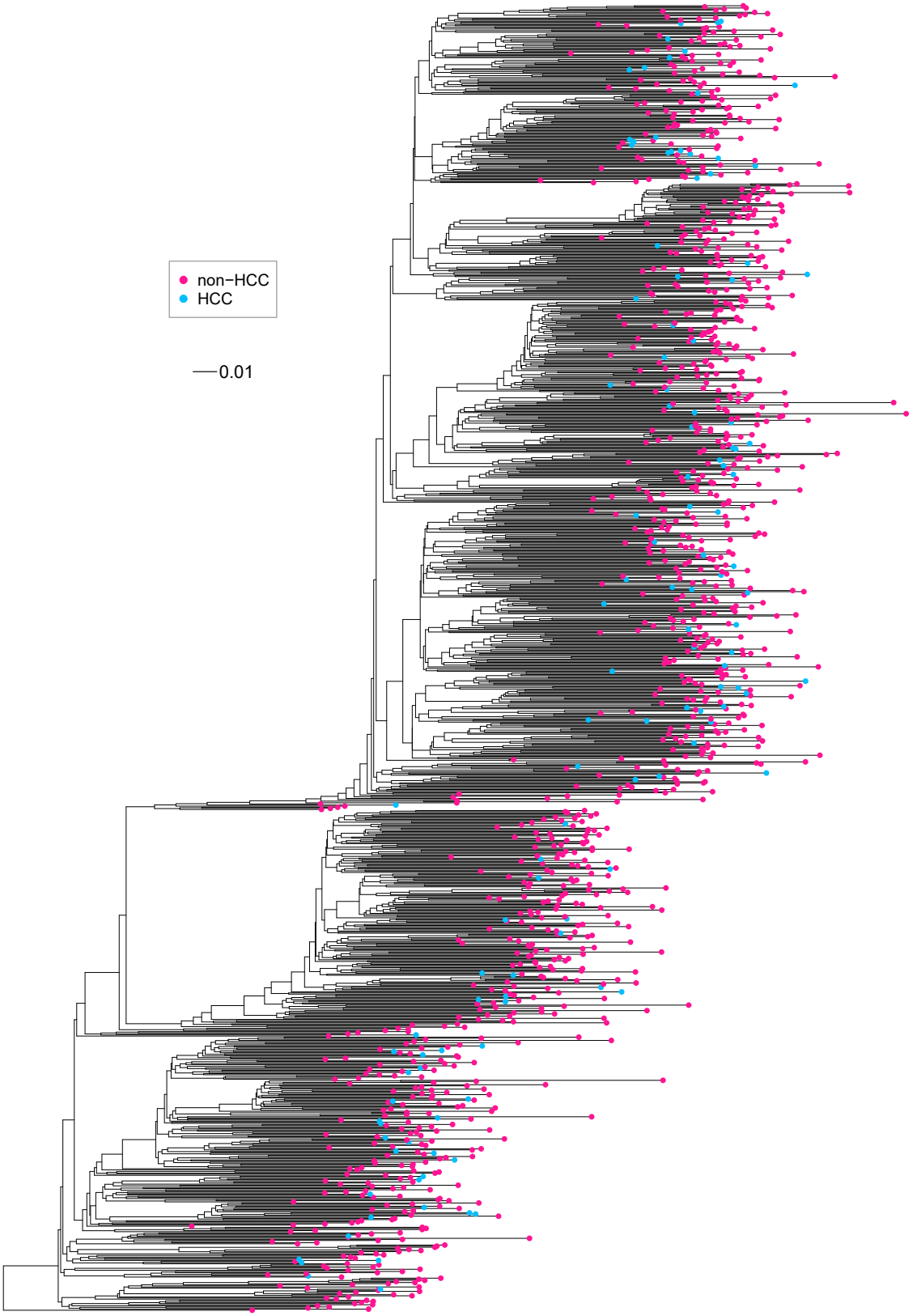

non−HCC
HCC
0.01

Figure B.4: Distribution of HCC status (presence vs. absence) across the HCV gt1a phylogeny.

The maximum likelihood phylogenetic tree was constructed using whole HCV gt3a genome sequences using IQ-TREE. The colour of the tips shows if the patient diagnosed with HCC (blue) or not (pink). Line thickness and colour are proportional to the posterior probability of a change point on the branch. Green branches indicate an extreme high confidence change point with a high Bayes factor calculated, identifying clades where the distribution of HCC significantly diverges from the rest of the tree. This tree did not identify any lineages associated with HCC distribution. The change point analysis conducted using treeBreaker. The scale bar represents 0.01 nucleotide substitutions per site.