

**Genome-to-genome analysis reveals the impact of the human innate and
adaptive immune systems on the hepatitis C virus**

M. Azim Ansari^{1,2#}, Vincent Pedergrana^{1#}, Camilla Ip¹, Andrea Magri², Annette Von
Delft³, David Bonsall³, Nimisha Chaturvedi⁴, Istvan Bartha⁴, David Smith³, George
Nicholson⁵, Gilean McVean^{1,6}, Amy Trebes¹, Paolo Piazza¹, Jacques Fellay⁴,
Graham Cooke⁷, Graham R Foster⁸, STOP-HCV Consortium, Emma Hudson³, John
McLauchlan⁹, Peter Simmonds³, Rory Bowden¹, Paul Klenerman³, Eleanor Barnes³
& Chris C. A. Spencer^{1*}

¹ Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive,
Oxford, OX3 7BN, UK

² Oxford Martin School, University of Oxford, 34 Broad Street, Oxford, OX1 3BD, UK

³ Nuffield Department of Medicine and the Oxford NHIR BRC, University of Oxford,
Oxford, OX1 3SY, UK

⁴ School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015
Lausanne, Switzerland

⁵ Department of Statistics, University of Oxford, Oxford, OX1 3LB, UK

⁶ Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery,
University of Oxford, OX3 7BN, UK

⁷ Wright-Fleming Institute, Imperial College, London, UK

⁸ Queen Mary University of London, 4 Newark Street, London, E1 4AT, UK

⁹ Centre for Virus Research, Sir Michael Stoker Building, 464, Bearsden Road,
Glasgow, G61 1QH, UK

Equal contribution

* Corresponding author

Abstract

Outcomes of hepatitis C virus (HCV) infection and treatment depend on viral and host genetic factors. We use human genome-wide genotyping arrays and new whole-genome HCV viral sequencing technologies to perform a systematic genome-to-genome study of 542 individuals chronically infected with HCV, predominately genotype 3. We show that both HLA alleles and interferon lambda innate immune system genes drive viral genome polymorphism, and that *IFNL4* genotypes determine HCV viral load through a mechanism that is dependent on a specific polymorphism in the HCV polyprotein. We highlight the interplay between innate immune responses and the viral genome in HCV control.

Introduction

Hepatitis C virus (HCV) infection presents a major health burden, infecting more than 185 million people worldwide¹ and leading to liver failure and hepatocellular cancer. Both host and virus genetic variations are associated with important clinical outcomes. Host genetic polymorphisms, most notably in the interferon lambda 3 and 4 locus, are associated with spontaneous clearance of the virus, response to treatment, viral load and progression of liver disease²⁻⁶. Similarly, viral genotypes, and distinct viral genetic motifs have been associated with the response to interferon based therapies^{7,8} whilst resistance-associated substitutions (RASs) have been identified for most of the new oral direct-acting antiviral (DAA) drugs⁹⁻¹². HCV can be divided into seven major genotypes, and most of the genetic data acquired to date focuses on HCV genotype 1 with a paucity of data addressing other genotypes. HCV genotype 3 is of particular interest as this genotype is known to infect 53 million people globally¹³, has particularly high prevalence in the UK, and is associated with a higher failure rate to DAA therapies^{14,15}. Despite these key observations, there has been few previous attempts to generate large full-length HCV genome datasets and to broadly assess the impact of host genes on the HCV viral genome.

Previous work has shown that within-host virus diversity evolves in response to the adaptive immune system, including candidate genes studies of the association between the human leukocyte antigen (HLA) type I proteins with the HCV genome^{16,17}. HLA molecules are expressed on most cell types and present viral peptides (epitopes) to cytotoxic T lymphocytes (CTLs) that kill infected cells. CTL-mediated killing of virus-infected cells drives the selection of viral polymorphisms

65 (“escape” mutants) that abrogate T cell recognition¹⁸. Understanding how host HLA
66 molecules impact on viral selection has important implications for the development of
67 HCV T cell vaccines that aim to prevent infection^{19,20}. Since both HLA molecules and
68 the viral epitopes they present are highly variable, a comprehensive host genome to
69 viral genome analysis at scale is needed to assess the relative contribution of host
70 HLA molecules in driving HCV viral genetic change. An analysis of this kind may also
71 reveal other host genes that play a key role in shaping the HCV viral genome.

72
73 Whole genome viral sequencing for HCV is particularly challenging since HCV is
74 genetically highly diverse both within and between infected individuals^{21,22}. To
75 overcome this we have recently developed new technologies that use next
76 generation sequencing with targeted enrichment^{23,24} to obtain a consensus genome
77 in each individual at scale. The falling cost of host genotyping, in combination with
78 large reference panels of human genomes²⁵ and advances in statistical
79 methodologies, mean that it is possible to obtain accurate inference of millions of
80 host genetic polymorphisms, including classical HLA alleles^{26–28}. For the first time,
81 these developments allow an unbiased genome-to-genome analysis²⁹, in large
82 cohorts of HCV infected individuals, to better understand how host genetic variation
83 drives changes within the virus during persistent infection at a population level.

84
85 We generated data from a cohort of 601 HCV infected patients (recruited to a clinical
86 study called BOSON³⁰) to systematically look for associations between host and virus
87 genomes, exploiting the fact that while host genetics remain fixed, the virus mutates,
88 allowing it to evolve during infection. We provide evidence that polymorphisms
89 relevant to the innate (*IFNL4*) and adaptive immune systems (HLA genes) are
90 associated with HCV sequence polymorphisms (so-called “footprints” in the viral
91 genome). We show that an interaction between host *IFNL4* genotype determines
92 HCV viral load through a mechanism that is dependent on a specific polymorphism in
93 the HCV polyprotein. By assessing rates of viral evolution in individuals with different
94 *IFNL4* genotypes, we highlight systematic differences in the innate immune response
95 and discuss how these might relate to previous associations with spontaneous
96 clearance and clinical treatment. We discuss the potential for a joint analysis of host
97 and viral genome data to provide information on underlying molecular interactions
98 and their importance in treating and preventing HCV, and other viral infections, in the
99 era of genomic analysis.

Results

Sample description and genetic structure

DNA samples from 567 patients within the BOSON clinical study³⁰ (out of 601 patients) were genotyped using the Affymetrix UK Biobank array. This array genotypes over 800,000 bi-allelic single nucleotide polymorphisms (SNPs) across the human genome, including a set of markers specifically chosen to capture common HLA alleles. Pre-treatment plasma samples from 583 patients in the same study were analysed to obtain HCV whole genome consensus sequences using a high-throughput HCV targeted sequence capture approach coupled with Illumina sequencing²³.

Both full-length HCV genome sequences and human genome-wide SNP data were obtained on a total of 542 patients of mainly White and Asian self-reported ancestry infected with HCV genotypes 2 or 3 (see **Supplementary Table S1** for patient demographics). After quality control and filtering of the human genotype data, approximately 330,000 common SNPs with minor allele frequency greater than 5% were available for analysis along with inferred alleles at both class I and II HLA genes. The full-length HCV genome is approximately 9.5 kb, corresponding to over 3000 encoded amino acids. In our dataset, 1226 sites of the HCV proteome were defined to be variable (where at least 10 isolates have an amino acid which differs from the consensus amino acid) to have minimal statistical power for analysis.

We characterised human genetic diversity in the cohort via principal component analysis (PCA). The first two principal components (PCs) corresponded to the sample's 83% White and 14% Asian self-reported ethnicity, (**Supplementary Figure S1**), which differed significantly in some of the inferred HLA alleles frequencies (**Supplementary Table S2**) consistent with previous observations³¹. The third PC separates individuals with Black self-reported ethnicity from the rest of the cohort. We summarised virus diversity by constructing a maximum-likelihood tree of the consensus sequences from each patient (**Supplementary Figure S2**). Major clades in the tree separated HCV genotypes 2 (8% of the sample) and 3 (which in turn comprised clades representing subtype 3a and non-subtype 3a samples (90% and 2% of the total respectively)). A PCA on virus nucleotides sequence data reflected the structure of the tree, specifically at the level of virus subtypes (**Supplementary**

Figure S3). We used the PCs as covariates to control for viral population structure in the primary genome-to-genome analysis.

Labelling the inferred tree of viral diversity with either genetic ancestry (measured by host PCs) or self-reported ethnicity revealed a group within genotype 3 that was strongly associated with Asian ancestry (**Supplementary Figure S2**). The viral group associated with hosts of Asian ancestry is basal on the phylogenetic tree relative to the viral clade associated with White ancestry. The observed relative reduction in viral diversity within hosts of White ancestry is likely to be explained by the introduction and spread of HCV genotype 3a from South Asia where it is endemic¹³.

Systematic host genome to virus genome analysis

We used the genotyped autosomal SNPs in the host genome to undertake genome-wide association studies, where the traits of interest were the presence or absence of each amino acid at the variable sites of the virus proteome, resulting in nearly one billion association tests. The analyses were performed on a compute cluster using logistic regression as implemented in PLINK2³², adjusting for sex, population structure, and assuming an additive model. To control for population structure, we included the first three PCs of the host and the first 10 PCs of the virus as covariates. We note that failure to control for either covariates leads to a significant inflation in the association test statistics (**Supplementary Figure S4**), as would be expected given the observed correlation in population structure of the virus and the host (**Supplementary Figures S1, S2 and S3**). Assuming a human genome-wide significance threshold of 5×10^{-8} ³³, and that amino acid variants in the viral genome are approximately uncorrelated once the population structure is accounted for, then a Bonferroni correction³⁴ results in a significance threshold of approximately 2×10^{-11} .

Figure 1 shows the result of the genome-to-genome association tests and highlights the strongest signals of association. Across the human genome, the most significant associations were observed between multiple SNPs in the major histocompatibility complex (MHC; chromosome 6) locus and a virus amino acid variant in non-structural protein 3 (NS3). Three other associations were observed between multiple SNPs in the host MHC and virus amino acids in NS3 and NS4B proteins ($P < 4 \times 10^{-9}$). Outside the MHC, the strongest association between host and virus was detected between the SNP rs12979860 in the *IFNL4* gene (chromosome 19) and HCV amino acids at position 2570 in NS5B protein ($P = 1.98 \times 10^{-9}$). Observed variability in the

density of nominally significant associations (**Figure 1**) is largely explained by variability in host and virus sequences, for example in the hyper-variable region (HVR) of HCV in E2 protein.

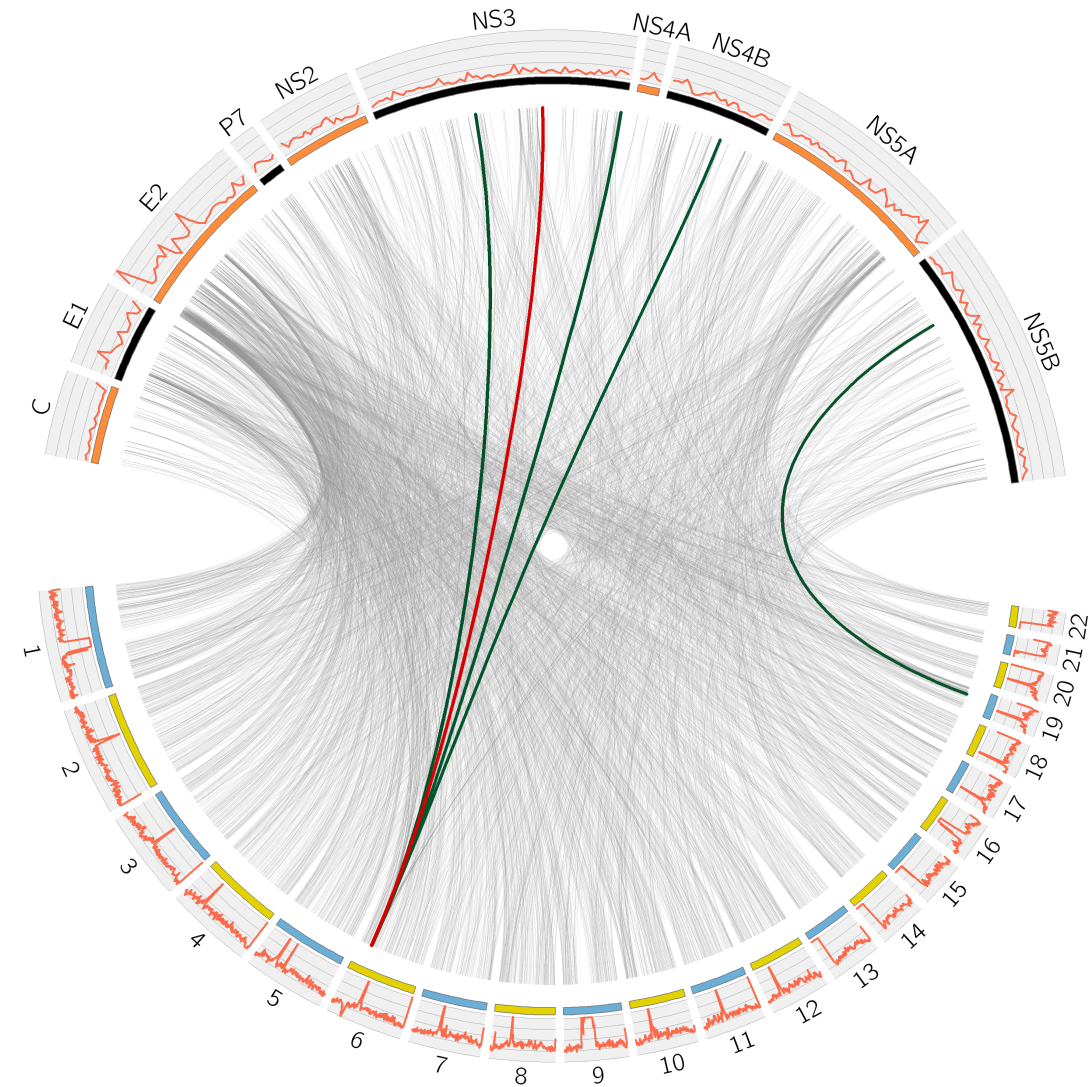


Figure 1. Human to hepatitis C virus genome-wide association study in 542 patients. The lower arc shows the human autosomes from chromosome 1 to 22, and the upper arc shows the HCV proteome from core (C) to NS5B. The red line represents the most significant association, $P < 2 \times 10^{-11}$. The four green lines represent suggestive associations, $P < 4 \times 10^{-9}$. The thin grey lines represent associations with $P < 10^{-5}$. The outer mini-panels represent, on the upper arc, the viral diversity as measured by Shannon entropy and, on the lower arc, the density of human SNPs in bins of 1Mb, with higher values away from the centre for both upper and lower arcs.

We observed 169 associations between human SNPs mapping to two loci and five HCV amino acid sites ($P < 4 \times 10^{-9}$) (**Supplementary Table S3**). Since these associations represent places where host genetic diversity has impacted on virus sequence diversity we refer to them as “footprints”. We interpret the signals of

association in the MHC region to indicate the effect of the adaptive immune system on genetic diversity in the virus genome. Whilst the effect of MHC was anticipated, the strong signal of association of the interferon lambda region with viral diversity indicates additional effects of the innate immune response.

HLA alleles to virus genome

SNPs in the MHC which show strong association with viral amino acids are likely to be correlated with alleles at HLA genes due to extensive linkage disequilibrium across the region^{35,36}. The HLA repertoire of a patient defines which viral peptides will be presented to T cells as part of the adaptive immune response, and can lead to the selection of viral mutations away from these peptides (“escape mutations”)^{16,37–40}. Upon transmission to another host, with a different HLA repertoire, reversion to the wild type may occur (“reversion mutations”).

Using the tree inferred from whole-genome viral sequences we estimated the ancestral amino acids at internal nodes of the tree⁴¹. Inferring the changes along the terminal branches of the tree aims to control for the confounding between host and virus population structures⁴² by looking at viral mutations after infection. To test for association between specific HLA alleles and viral amino acids, we constructed a 2x4 contingency table for each amino acid present at variable sites, and counted the number of changes and non-changes from ancestral amino acids in carriers, and non-carriers, for each imputed HLA allele (see **Figure 3a** for an example). We used a Fisher’s exact test to assess significance of the association along the viral genome (**Figure 2a**), and a permutation approach to estimate the false discovery rate (FDR)^{43,44}.

In the whole cohort, at a 5% FDR, 24 combinations of HLA alleles and HCV sites were significant (**Table 1**) and this increased to 153 associations at a 20% FDR (**Supplementary Table S4**). Out of 21 viral amino acid positions showing signals of association with one or more HLA alleles, 12 were located in previously reported HCV genotype 3 epitopes²⁰ (**Table 1**), which represents a strong enrichment (odds ratio, OR=5.2, $P=2.8 \times 10^{-4}$). We also observed that the NS3 protein was strongly enriched for association signals with HLA alleles (OR=5.5, $P=2.2 \times 10^{-4}$). The strongest HLA footprinting signals are found with common alleles at *HLA-A* and *HLA-B* genes, although signals are also found in *HLA-C* and the class II gene *HLA-DQA1*. At position 1646 in the HCV genome (in NS3), the footprint is seen with multiple HLA

alleles (*B*08:01*, *C*07:01* and *DQA1*05:01*), although this is potentially a result of linkage disequilibrium between these HLA alleles (r^2 between *B*08:01* and *C*07:01* = 0.49, and r^2 between *B*08:01* and *DQA1*05:01* = 0.22; **Supplementary Figure S5**). As the majority of the cohort are patients infected with genotype 3a virus and self-reporting as White (411 out of 542), we repeated the HLA analysis on these patients specifically (**Table 1**) as an additional check that the associations are not driven by correlations between viral genotypes and host ancestry. Broadly the association evidence was consistent with analysis of the whole cohort; some associations reduced in significance potentially due to loss of power because of reduced sample size (**Table 1** and **Supplementary Figures S6**); and other associations increased in significance, potentially due to effects that are specific to HCV genotype 3a. The associations reported here (in **Figure 2** and **Table 1**), and the extent to which they are genotype specific, will require replication in independent cohorts to be confirmed.

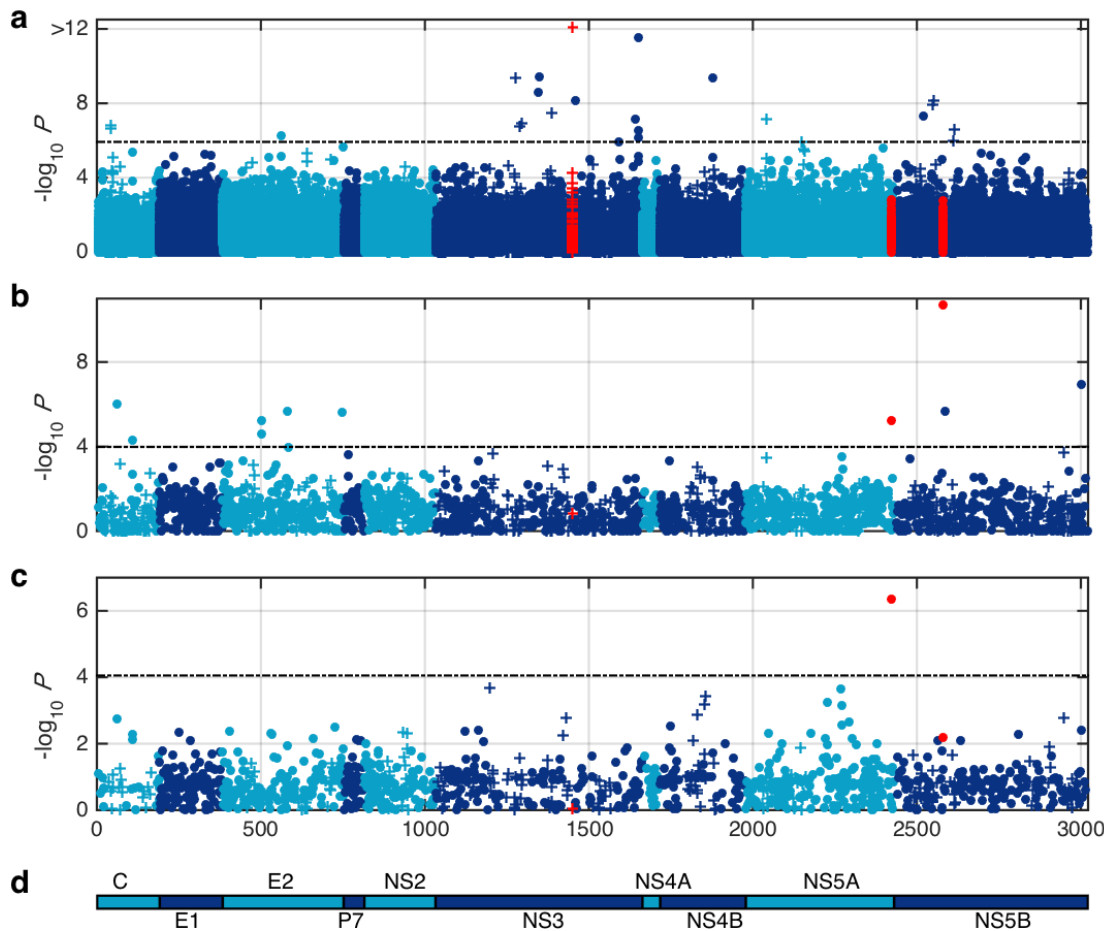


Figure 2. Association between HCV amino acids and (a) HLA alleles, (b) *IFNL4* genotypes using Fisher's exact test, and (c) pre-treatment viral load (\log_{10} PTVL) using linear regression. Sites in experimentally validated epitopes in HCV genotype

3²⁰ are indicated by a plus sign. Viral sites 1444, 2414 and 2570 are coloured as red. Dashed lines represent a 5% false discovery rate. (d) HCV polyprotein.

The most significant association was found for *HLA-A*01:01* and the viral amino acid at position 1444 in the NS3 protein. Having identified a viral amino acid subject to footprint by a host HLA allele, we can look for evidence of escape and/or reversion mutations at this viral site (**Figure 3a**). The *HLA-A*01:01* allele was associated with a tyrosine-to-phenylalanine change in the NS3 protein (Y1444F). Carriers of *HLA-A*01:01* allele very rarely lose an existing F amino acid, and there is a strong relative increase in the inferred number of changes from Y to F in individuals that carry *HLA-A*01:01* allele ($P=1 \times 10^{-32}$), consistent with known T cell selection at this epitope⁴⁵. The signal of association is at the ninth amino acid of the experimentally validated epitope (**Supplementary Figure S7**) which is known to be prone to escape mutations as it alters the contact between the epitope and the HLA groove which presents the antigen⁴⁶.

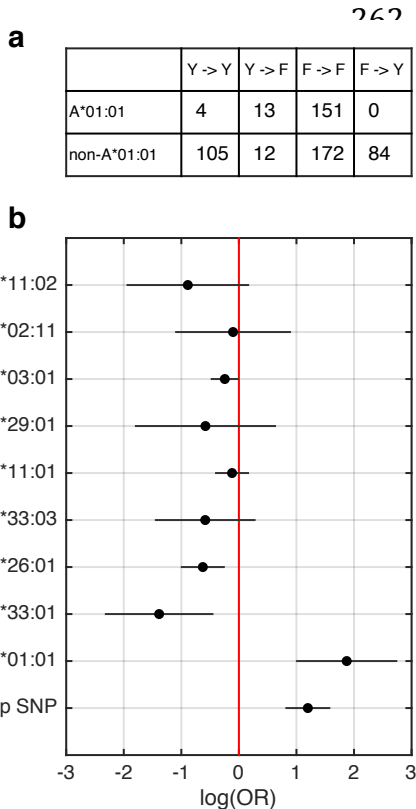


Figure 3. Association between viral position 1444 and *HLA-A* alleles. (a) Table of changes to and from the ancestral amino acids in individuals with and without the *HLA-A*01:01* allele. (b) Estimated effect size (log odds ratio) of common *HLA-A* alleles and the top associated SNP (rs3129073) on the possibility of mutation from F to Y at viral position 1444. The black dots represent the effect size estimate and the bar represent the 95% confidence intervals.

To quantify the influence of the host's HLA at position 1444, we estimated the odds ratio separately for the most common *HLA-A* alleles, and the most associated SNP in

the region (rs3129073), on the amino acid changes from F to Y. **Figure 3b** shows that our data support a very strong signal for individuals carrying *HLA-A*01:01* allele to carry a virus with the non-antigenic F amino acid which is consistent with the size of effect of the rs3129073. A weaker signal for *HLA-A*33:01* allele in the opposite direction is likely explained by these alleles acting as surrogates for non-*HLA-A*01:01* allele carriers. Association analysis across the HLA region which was conditioned on *HLA-A*01:01* allele (**Supplementary Figure S8**) removed all association signals ($P>10^{-4}$), suggesting that the primary *HLA-A*01:01* association explains other associations in the region with amino acids at position 1444. Using the 27 HLA and viral amino acid associations that surpass the 20% FDR threshold, and have sufficient observations to estimate the ORs, we observed a negative correlation between the ORs of escape and reversion ($R=-0.65$, $P<0.01$; **Supplementary Figure S9**). These observations are consistent with HLA alleles driving patterns of both escape and reversion at viral amino acids. Our analysis provides a map of their influence across the HCV genome.

IFNL4 variants to virus genome

Variants in the interferon lambda region have been associated with multiple HCV outcomes including spontaneous clearance, treatment response, viral load and liver disease progression^{2,3,47}. In our genome-wide analysis, variants around the interferon lambda region showed the strongest association with HCV amino acids outside the MHC region (the top associated SNP is rs12979860, $P=1.98\times10^{-9}$). For that SNP, the CC genotype is associated with higher rates of spontaneous clearance and interferon-based treatment response, putatively due to the fact that the C allele tags a dinucleotide insertion polymorphism (rs368234815-TT) which prevents *IFNL4* expression⁴⁸. In our cohort, only two individuals had discordant genotypes for rs12979860 and rs368234815 (which was imputed), resulting in a strong linkage disequilibrium ($r^2=0.992$) between these two SNPs. Therefore, individuals with non-CC genotypes express *IFNL4* and have increased expression of interferon stimulated genes (ISGs)⁴⁹.

To interrogate these associations further, we compared viral amino acid changes between hosts with CC and non-CC genotypes at the *IFNL4* SNP rs12979860 using the same Fisher's exact test as described above. The most significant association was with changes to and away from valine (V) at position 2570 in the viral NS5B

protein ($P=1.94 \times 10^{-11}$) (**Figure 2b**). We replicated this association using an equivalent analysis in an independent study of 360 patients of European ancestry chronically infected with HCV genotypes 2 or 3^{50,51} (one-sided $P=0.005$). In addition, a candidate gene association study in a genotype 1b single source infection cohort has shown an association between *IFNL4* genotypes (rs12979860) and an amino acid in the same region of NS5B (position 2609)³⁸ reinforcing the potential role of the locus in interactions with the host innate immune system.

Overall, using a 5% FDR, 11 significant associations were observed between *IFNL4* genotypes and amino acid positions located in core, E2, NS5A and NS5B proteins (**Supplementary Table S5**). These amino acids are not located in any known HCV genotype 3 HLA-associated epitopes, however one of the sites (position 109 in the core protein) is also associated with *HLA-B*41:02* in our dataset ($P=4.3 \times 10^{-6}$, **Table S4**). The top signals of associations between *IFNL4* genotypes and viral variants are located in viral proteins known to be HCV immune response targets³⁸, and those which disrupt human proteins in order to escape mechanisms of innate immunity⁵². We used a permutation approach to determine if any of the viral proteins are enriched in the number of associations with *IFNL4* genotypes. Only the core protein is nominally- ($P<0.05$) enriched, while the NS5A protein is nominally depleted in the number of associations with the *IFNL4* genotypes.

In the 487 genotype 3a infected patients, we defined the population consensus to be the most common amino acid and assessed whether regions of the viral genome with either high diversity (HVR1, HVR2 and HVR3) or previously reported to be correlated with interferon based therapy outcome (the IFN sensitivity-determining region (ISDR), protein kinase phosphorylation homology domain (PePHD) and the protein kinase binding domain (PKR-BD))^{7,53,54} showed difference in the number of changes from the population consensus, between patients with the CC and non-CC genotypes (**Supplementary Table S6**). Patients with the CC genotype had an increased number of mutations away from the population consensus in the ISDR ($P=3.5 \times 10^{-4}$), HVR2 ($P=2.5 \times 10^{-6}$) and PKR-BD ($P=0.012$) regions. However, we observed the same pattern across the genome ($P=5.06 \times 10^{-6}$) suggesting that the *IFNL4* effect is not necessarily specific to these regions.

To further investigate the selective pressures on the virus in patients with *IFNL4* CC and non-CC genotypes, we estimated the rates of synonymous (dS) and non-synonymous substitutions (dN) in genotype 3a infected patients (**Figure 4**). Whilst

there was no difference in dS in CC and non-CC patients (**Figure 4a**, $P=0.68$), dN was significantly higher in CC patients (**Figure 4b**, $P=1.6 \times 10^{-8}$). The lower dN/dS ratio in non-CC patients ($P=1.3 \times 10^{-10}$) is potentially indicative that the virus is under a stronger purifying selection (**Figure 4c**). This hypothesis is supported by the observation that for the same rate of synonymous substitutions dS (a surrogate for the time and amount of divergence), the HCV genome is under a larger purifying selection in patients with *IFNL4* non-CC genotypes than CC genotype (**Figure 4d**).

Estimating dN/dS ratio per viral gene showed that this ratio was significantly higher ($P<0.05$) in CC patients compared to non-CC patients in E1, E2, NS3 and NS5B (**Supplementary Figure S10**). A sliding window analysis across the HCV genome showed that in the full cohort E1 and E2 genes had a much higher dN/dS ratio compared to the rest of the genome. These envelope genes include hyper-variable regions (HVRs), and are thought to be the primary targets of the antibody-based immune response (**Supplementary Figure S11**).

We also tested whether *IFNL4* genotypes had an impact on HLA presentation as measured by our foot printing analysis. We did not find any consistent evidence for interaction between individual sites presentation by HLA alleles and *IFNL4* genotypes (**Supplementary Figure S12**), or in the mean number of escape mutations in CC and non-CC patients when comparing HLA allele carriers and non-HLA allele carriers.

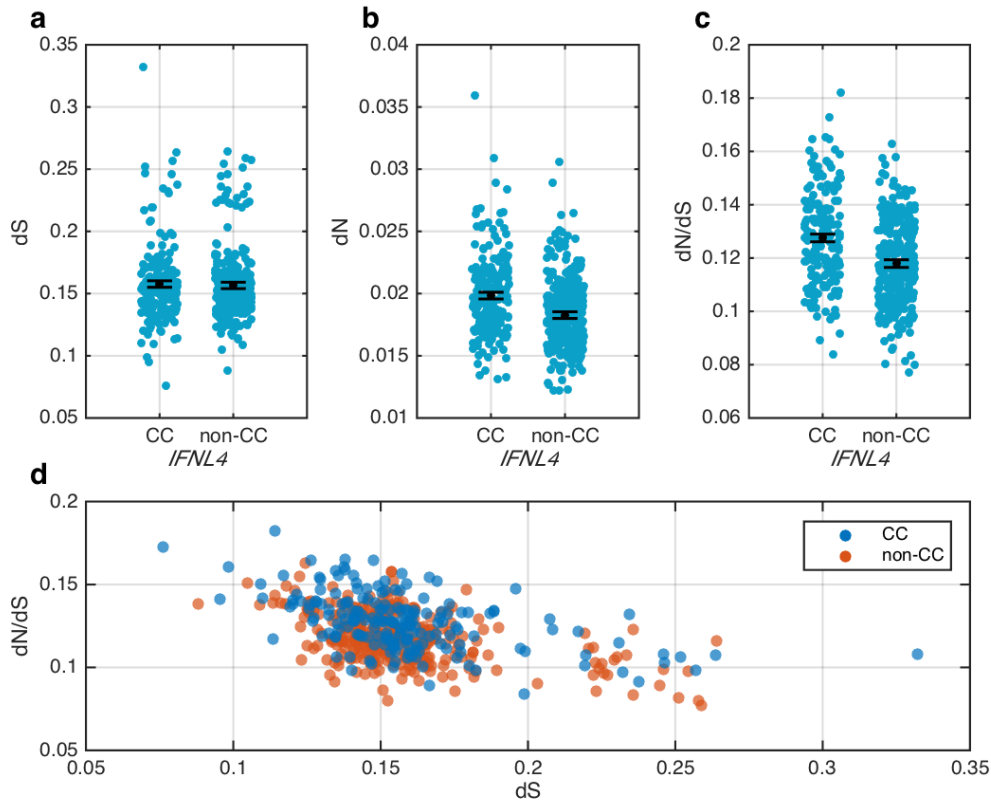


Figure 4. Association between *IFNL4* genotypes and rates of synonymous (dS) and non-synonymous (dN) substitutions in HCV genome in genotype 3a infected patients. Each blue dot represents the mean dS, dN or dN/dS ratio per patient. The mean and 95% confidence intervals are shown as black dots and bars. (a) Rate of synonymous substitution ($P=0.68$). (b) Rate of non-synonymous substitution ($P=1.6 \times 10^{-8}$). (c) dN/dS ratio ($P=1.28 \times 10^{-10}$). (d) The joint distribution of dS and dN/dS in individuals with the *IFNL4* non-CC genotypes (red dots) and with the CC genotype (blue dots).

Host and virus genetic determinants of viral load

Pre-treatment viral load (PTVL, as measured in IU/ml) was available for all patients (See **Supplementary Table S1** for details of the cohort). We performed a genome-wide association study in patients infected with HCV genotype 3a using an additive linear regression model adjusted for sex and the three first host PCs for \log_{10} transformed PTVL (\log_{10} PTVL) (**Figure 5a**). We replicated the known association between *IFNL4* variants on chromosome 19 and viral load³ (rs12979860, $P=5.9 \times 10^{-10}$) with the non-CC genotypes conferring an approximately 0.45-fold decrease in viral load (mean for non-CC = 3.4760×10^6 IU/mL and for CC = 6.3447×10^6 IU/mL).

We also performed a genome-wide association study to detect associations between viral amino acids and viral load (**Figure 2c**). The only amino acid significantly associated with \log_{10} PTVL at a 5% FDR was a change from a serine (S) to an

asparagine (N) at position 2414 in NS5A protein ($P=9.21 \times 10^{-7}$) (**Figure 5b**). This site is one of the 11 sites significantly associated (5% FDR) with *IFNL4* genotypes (**Figures 2b and 2c**). In patients with a serine at position 2414, the association between *IFNL4* genotypes and \log_{10} PTVL is highly significant ($P=9.37 \times 10^{-9}$, **Figure 5c**). However, we observed no association ($P=0.9$) between *IFNL4* genotypes and \log_{10} PTVL in patients infected with a virus that has a different amino acid (**Figure 5d**). In other words, the host's *IFNL4* genotypes determine viral load only if they are infected by a virus with the serine amino acid at position 2414 in NS5A protein (**Figures 5c and 5d**). The interaction is statistically significant when analysing either the whole cohort ($P=0.0173$) or just patients with genotype 3a infections who self-report as being White ($P=0.01692$). Together the combinations of non-CC genotypes and a serine viral amino acid at position 2414 are inferred to result in a 0.57-fold decrease in viral load compared to all other combinations (mean viral load for non-CC and serine at position 2414 $=2.8083 \times 10^6$ IU/mL and all other combinations $=6.4735 \times 10^6$ IU/mL).

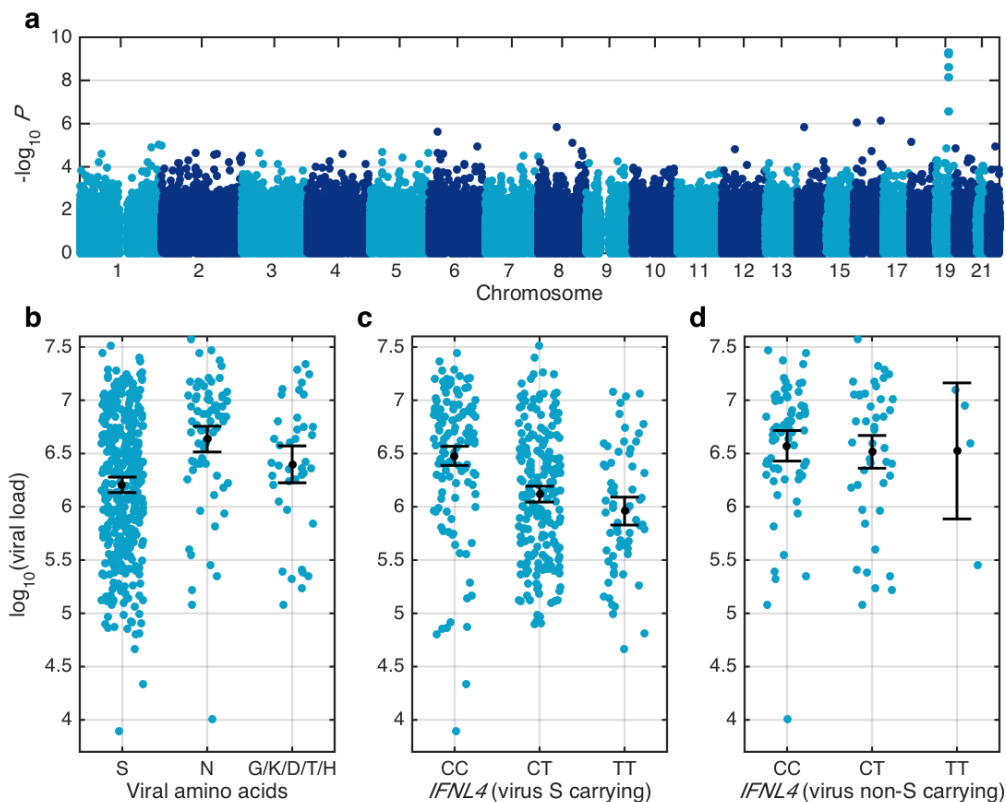


Figure 5. Association between \log_{10} transformed pre-treatment viral load (\log_{10} PTVL) and human and virus genetic variants in genotype 3a infected patients. (a) Association between human SNPs and \log_{10} PTVL. (b) Distribution (blue dots), estimated mean and 95% confidence interval (shown as black dots and bars) of \log_{10} PTVL stratified by amino acids present at viral position 2414 ($P=9.21 \times 10^{-7}$), (c) by *IFNL4* (CC, CT and TT) genotypes in patients whose virus carries a serine at

position 2414 ($P=9.37 \times 10^{-9}$) (d) or in patients whose virus does not carry a serine amino acid at position 2414 ($P=0.9$).

To investigate whether the mutations associated with higher viral loads and/or *IFNL4* genotypes influence replication ability, each was introduced into the modified S52 replicon of genotype 3a⁵⁵ and replication assessed in short term Huh7.5 cell cultures. As a positive control, introduction of an alanine to threonine change at position 1192 led to a 10-fold increase in *in vitro* replication, as previously reported⁵⁶ (**Supplementary Figure S13**). Introduction of the change from a serine to an asparagine at position 2414 resulted in approximately a 10-fold increase in replication (**Supplementary Figure S13**). However, introducing a range of other mutations nominally associated with higher viral load produced more variable outcomes. The mutant I1851V replicons show a 10-fold reduced replication compared to the wild type S52 replicon while others showed 2-fold to 3-fold increase. (**Supplementary Figure S13**). No mutations associated with *IFNL4* genotypes other than N2414S led to altered replication phenotypes *in vitro*.

We assessed the relationship between *IFNL4* genotypes, viral amino acids and viral load in genotype 3a isolates by estimating the effect of the CC genotype on the chances of observing a change from the consensus (to non-consensus) using a 2x2 Fisher's exact test (only using data in which the consensus is inferred to be ancestral). These odds ratios were compared against the estimated effect of non-consensus amino acids on viral load (whilst accounting for *IFNL4* genotypes) at sites associated with the *IFNL4* genotypes at a 20% FDR (**Supplementary Figure S14**). We observe a positive relationship ($R=0.42$, $P=0.005$) whereby non-consensus amino acids which are increased in frequency in individuals with a CC genotype tend to also associate with increased viral load. The same positive relationship was observed when estimating the effect on viral load in individuals with a CC genotype only ($R=0.35$, $P=0.02$) or non-CC genotypes only ($R=0.4$ and $P=0.007$). A nominally significant trend was observed when the analysis was done for all variant positions in the viral genome ($R=0.099$ and $P=0.04$) (**Supplementary Figure S15**).

Discussion

Here we report the first systematic analysis of associations between variation in human and HCV genomes in a large patient cohort. Advances in DNA and RNA

sequencing technology and new bioinformatics tools have allowed full-length viral consensus sequences to be obtained in large number of patients for reasonable cost (approximately £100 per sample), as well as host genetic data at millions of directly assayed and imputed polymorphisms (approximately £75 per sample). We apply a fast and simple approach to test for association between host and pathogen variants, using a logistic regression analysis corrected for both human and viral population structures by using the principal components of the genome-wide data as covariates (60 hours for approximately 2500 association studies of 330,000 SNPs). We also applied a contingency table analysis based on the inferred viral amino acid changes since infection. We anticipate that with the reduction in the cost of sequencing and genotyping, and the increasing interest in studying large patient cohorts, analyses of this kind will become a powerful approach to understanding infectious diseases.

We found strong evidence for the adaptive immune system exerting selective pressures on the HCV genome, presumably by preferentially selecting viral mutations that avoid antigenic presentation by the host's HLA proteins. Some of the observed associations are located in experimentally validated viral epitopes²⁰, however others have not been described experimentally and most likely represent sites of novel T cell escape mutations. Assuming that our analysis removes biases associated with population structure and incorrect ancestral inference, 5% of the viral amino acids (153/3021) are associated with HLA alleles (at 20% FDR). These data highlight the importance of the adaptive immune system in driving viral evolution, and serves as a map of the targets of T-cell based immunity along the HCV genome which can aid vaccine design and development²⁰.

In addition to the HLA, we now show that *IFNL4* activity may significantly shape the HCV viral genome. Previous studies have shown the “favourable” CC *IFNL4* genotype increases the chances of spontaneous resolution and interferon-based treatment success²⁻⁶. Prokunina-Olsson et al. have shown that the *IFNL4* TT/TT genotype (rs368234815), which is strongly linked to the “favourable” CC *IFNL4* genotype (rs12979860), abolishes the expression of *IFNL4*⁴⁶, whereas in individuals with the *IFNL4* Δ G/TT or Δ G/Δ G genotypes (linked to “unfavourable” non-CC *IFNL4* genotypes), *IFNL4* is expressed, leading to the downstream up-regulation of hepatic ISGs expression via the JAK-STAT pathway⁴⁹. The expression of ISGs has been shown to render the host less susceptible to exogenous INF α /γ stimulation and is associated with more infected cells in the liver⁵⁸. To date it has been

presumed that this fully explains why patients with specific *IFNL4* genotypes have differential outcomes during primary infection or with drug therapy^{58,59}. Our analysis adds to this hypothesis, with the observation that *IFNL4* variants also impact significantly on the HCV viral genome at multiple amino acid sites. Indeed, these were the strongest footprinting signals in our systematic analysis outside HLA region. The most significant association was at position 2570 in the viral NS5B protein; an association that we replicated in an independent cohort infected with HCV genotype 2 or 3. The additional signals in NS5B protein associated with HLA alleles or *IFNL4* genotypes at a 5% FDR did not replicate ($P > 0.05$) potentially due to a lack of power resulting from the smaller sample size of the replication cohort and/or the fact that the phylogenetically corrected Fisher's tests could not be performed in an equivalent way. An association between *IFNL4* and amino-acid variability in the same region of NS5B protein, has previously been reported in a candidate gene study³⁸ in a single source infection cohort. *IFNL4* has also been associated with a viral SNP associated with DAA resistance in HCV genotype 1 infection⁶⁰ although this association has not been replicated in our HCV genotype 3 cohort⁶¹. However, the broader impact of interferon lambda host genes that are associated with, and potentially select for, specific viral variants has not been previously recognised.

As far as we are aware, we also report the first association between a single virus amino acid variant (serine vs. non-serine at position 2414 in NS5A) and HCV viral load. HCV viral load is an important and clinically relevant parameter since patients with higher HCV viral loads have lower response rates to IFN and DAA based therapy⁶² (independent of *IFNL4* status). Paradoxically, the "favourable" *IFNL4* variants have also been associated with both an increase in disease progression⁶³ and high viral load^{3,64}; our data shows that site 2414 is one of the 11 sites that are putatively associated with *IFNL4* genotypes. Further, our data shows that a decrease in viral load was only observed in those patients with non-CC genotypes whose virus carried the serine amino acid at site 2414 in NS5A protein. Since the 2414 S variant is found in 85% of non-CC patients (compared to 67% of CC patients), this interplay between host and viral genes helps explain the previous observation that non-CC patients have a lower HCV viral load (**Figure 6**). Our in vitro data from a genotype 3 replicon assay shows that a change from a serine to asparagine is associated with an increase in RNA replication and perhaps hyper-phosphorylation⁶⁵, which is a negative regulator of virus replication.

Our interpretation of the data (**Figure 6**) is that the expression of *IFNL4* by $\Delta G/TT$ or $\Delta G/\Delta G$ genotypes (tagged in our cohort by the “unfavourable” non-CC genotypes) leads to the activation of additional components of the immune response, likely driven by ISGs, which interact directly with specific amino acids in the viral genome (most notably amino acid 2414 in NS5A which has a significant impact on viral load (**Figures 2b, 5 and S14**)). Our data suggest that this also leads to an overall increase in the strength of purifying selection (decrease in dN/dS, **Figure 4**), and together this leads to lower viral load. However, viruses that establish chronic infections in non-CC patients have evolved to survive in the more hostile environment (for example mutating the serine at amino acid 2414 of NS5A), which makes them less likely to respond to interferon-based therapy. In contrast, our analysis suggests that the “favourable” CC genotype and the inactivation of *IFNL4* gene (by the *IFNL4* TT genotype) disables components of the immune response (therefore removing the effect of amino acid 2414 in NS5A on viral load), which leads to a reduced level of purifying selection (**Figure 4**). It is possible that this then permits a range of mutations which confer higher replicative fitness and therefore higher viral load (**Figure S14**), but these viruses are more susceptible to interferon-based treatments. At the population level, we would expect a balance in the relative contribution of these mechanisms as viruses move between individuals with CC and non-CC genotypes. Our results make the prediction that the outcome of a new infection will be dependent on both the HLA alleles and the *IFNL4* genotype of the patient who is the source of the new infection. Further analysis is required to fully understand the impact of “favourable” CC and “unfavourable” non-CC genotypes on the different components of the immune system and to establish their clinical relevance before, during and after infection.

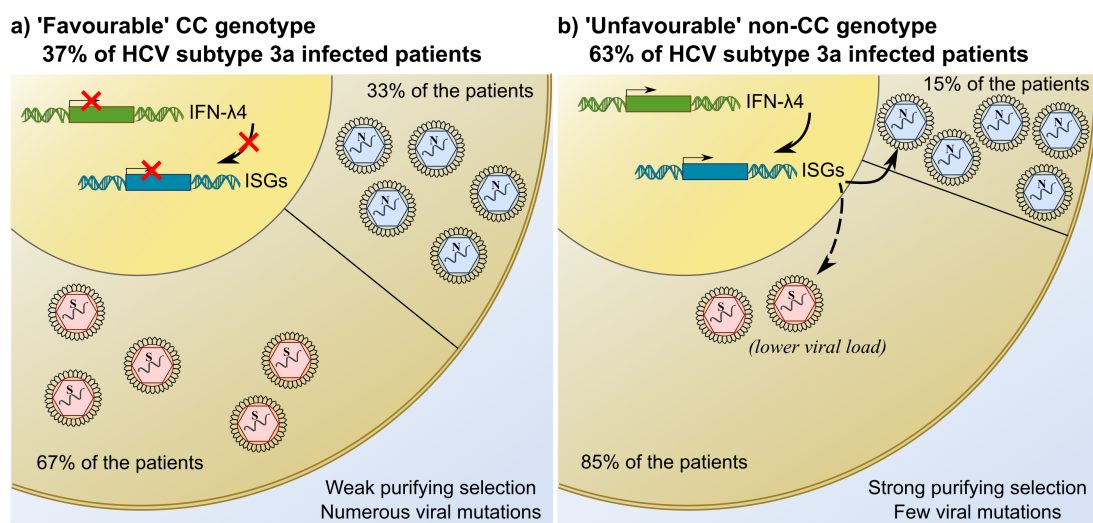


Figure 6: Overview of the observations relating to the interplay between innate immune response and the viral genome in hepatitis C virus (HCV) control. a) Infected individuals with *IFNL4* CC genotype (37% of the HCV genotype 3a infected patients) show high rates of spontaneous and treatment-induced clearance of HCV. *IFNL4* is not expressed, which in turn induces a weaker and possibly differential interferon-stimulated genes (ISG) expression. The host environment is associated with weaker purifying selection and allows viral mutations associated with a better replicative fitness to accumulate, leading to higher viral load. In this group of patients, 67% are infected by a virus with a serine at position 2414 and 33% with a different amino-acid. b) Infected individuals with *IFNL4* non-CC genotypes (63% of the HCV genotype 3a infected patients), have lower rates of spontaneous and treatment-induced clearance of HCV. *IFNL4* is expressed and induces ISGs that collectively establish an antiviral state hostile to viral replication. This hostile environment induces a high selective pressure and fewer viral mutations can accumulate. Although the serine at position 2414 (compared to non serine) is associated with lower viral load, it is highly prevalent in this group of patients with non-CC genotypes (85% are infected by a virus with a serine at position 2414 and 15% with a different amino-acid).

In conclusion, we provide a comprehensive genome-to-genome analysis in chronic HCV infections. Using this genome-wide, hypothesis-free, approach we show that the host's HLA alleles leave multiple footprints in the HCV genome and that the host's innate immune environment also influences the presence of amino acid polymorphism in the virus, both at specific loci, and genome-wide. We observe a common viral amino-acid residue that is associated with HCV viral load only in patients with the "unfavourable" non-CC *IFNL4* genotypes. These observations suggest that the innate and adaptive immune system jointly impact on HCV genome evolution and likely together determine the establishment of infection and its control over time. The new insights into the biological mechanisms that drive HCV evolution *in vivo*, and the identification of specific interactions between viral and host polymorphisms are relevant for future approaches to treatment stratification and vaccine development.

Materials & Correspondence.

Correspondence and material requests should be addressed to Chris Spencer (chris.spencer@well.ox.ac.uk) or by contacting STOP-HCV <http://www.stop-hcv.ox.ac.uk/contact>.

Author contributions

M.A.A and V.P contributed equally; E.B and C.C.A.S jointly supervised research; M.A.A, V.P, E.B and C.C.A.S conceived and designed the experiments; M.A.A, V.P, C.I, A.M, A.V, N.C, I.B, A.T and P.P performed the experiments; M.A.A, V.P, A.M, N.C, I.B, G.N and C.C.A.S performed statistical analysis; M.A.A, V.P, A.M, P.K, E.B, C.C.A.S analysed the data; C.I, G.N, A.V, D.B, G.M, A.T, G.N, P.P, J.F and J.M contributed reagents/materials/analysis tools; M.A.A, V.P, J.F, G.C, G.R.F, E.H, J.M, P.M, R.B, P.K, E.B and C.C.A.S wrote the paper.

Acknowledgements

The authors would like to thank Gilead Sciences for the provision of samples and data from the BOSON clinical study for use in these analyses. The authors would also like to thank HCV Research UK (funded by the Medical Research Foundation) for their assistance in handling and coordinating the release of samples for these analyses.

This work was funded by a grant from the Medical Research Council (MR/K01532X/1 – STOP-HCV Consortium). The work was supported by Core funding to the Wellcome Trust Centre for Human Genetics provided by the Wellcome Trust (090532/Z/09/Z). E.B is funded by the MRC as an MRC Senior clinical fellow with additional support from the Oxford NHR BRC and the Oxford Martin School. M.A.A. is funded by the Oxford Martin School. G.C is funded by the BRC of Imperial College NHS Trust. P.K is funded by the Oxford Martin School, NIHR Biomedical Research Centre, Oxford, by the Wellcome Trust (091663MA) and NIH (U19AI082630). C.C.A.S is funded by the Wellcome Trust (097364/Z/11/Z). G.M is funded by the Wellcome Trust grant 100956/Z/13/Z.

Conflicts of interest

The authors disclose the following: G.R.F: Grants Consulting and Speaker/Advisory Board: AbbVie, Alcura, Bristol-Myers Squibb, Gilead, Janssen, GlaxoSmithKline,

634 Merck, Roche, Springbank, Idenix, Tekmira, Novartis. G.M. is a partner in Peptide
635 Groove LLP, which commercializes HLA*IMP.
636

Online Methods

Patients and sample

Plasma and DNA samples came from patients enrolled in the Boson study. Boson study is a phase 3 randomized open-label trial to determine the efficacy and safety of sofosbuvir with and without pegylated-interferon-alfa, in treatment-experienced patients with cirrhosis and hepatitis C virus (HCV) genotype 2 infection and treatment-naïve or -experienced patients with HCV genotype 3 infection³⁰. 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. The study was conducted in accordance with the International Conference on Harmonization Good Clinical Practice Guidelines and the Declaration of Helsinki. The study reported here is not a clinical trial, but is based on the analysis of patients from a clinical trial (registration number: NCT01962441). Sample sizes were determined by the available data. All samples for which both viral sequencing and host genetics were available were included in the final analysis unless otherwise specified.

Host genotyping and imputation

Informed consent for host genetic analysis was obtained from 567 patients. Genotyping was performed using Affymetrix UK Biobank arrays. We imputed the MHC class I loci *HLA-A*, *HLA-B*, *HLA-C* and class II loci *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5* using HLA*IMP:02²⁷ accessed 22 March 2015. HLA amino acids were also imputed by SNP2HLA⁶⁶ using the T1DGC as the reference panel, which contains 5225 unrelated individuals (10,450 haplotypes). Logistic regression using posterior genotype probabilities (allele dosages) for each HLA allele from SNP2HLA were carried out using PLINK2³² (<https://www.cog-genomics.org/plink2>).

Virus sequencing

Sample collection and preparation

RNA was isolated from 500µl plasma using the NucliSENS magnetic extraction system (bioMerieux) and collected in 30µl of kit elution buffer for storage in aliquots at -80°C.

Sequencing library construction, enrichment and sequencing

Libraries were prepared for Illumina sequencing using the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Biolabs) with 5µl sample (maximum 10ng total RNA) and previously published modifications of the manufacturer's guidelines (v2.0)²³, briefly: fragmentation for 5 minutes at 94°C, omission of Actinomycin D at first-strand reverse transcription, library amplification for 18 PCR cycles using custom indexed primers⁶⁷ and post-PCR clean-up with 0.85× volume Ampure XP (Beckman Coulter).

Libraries were quantified using Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen) and analysed using Agilent TapeStation with D1K High Sensitivity kit (Agilent) for equimolar pooling, then re-normalized by qPCR using the KAPA SYBR® FAST qPCR Kit (Kapa Biosystems) for sequencing. A 500ng aliquot of the pooled library was enriched using the xGen® Lockdown® protocol from IDT (Rapid Protocol for DNA Probe Hybridization and Target Capture Using an Illumina TruSeq® or Ion Torrent® Library (v1.0), Integrated DNA Technologies) with equimolar-pooled 120nt DNA oligonucleotide probes (IDT) followed by a 12-cycle, modified, on-bead, post-enrichment PCR re-amplification. The cleaned post-enrichment ve-Seq library was normalized with the aid of qPCR and sequenced with 151b paired-end reads on a single run of the Illumina MiSeq using v2 chemistry.

Sequence data analysis

De-multiplexed sequence read-pairs were trimmed of low-quality bases using QUASR v7.0120⁶⁸ and adapter sequences with CutAdapt version 1.7.1⁶⁹ and subsequently discarded if either read had less than 50b remaining sequence or if both reads matched the human reference sequence using Bowtie version 2.2.4⁷⁰. The remaining read pool was screened against a BLASTn database containing 165 HCV genomes⁷¹ covering its diversity both to choose an appropriate reference and to select those reads which formed a majority population for de novo assembly with Vicuna v1.3⁷² and finishing with V-FAT v1.0 (<http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/v-fat>). Population consensus sequence at each site is defined as the most common variant at that site among all the patients.

Phylogenetic and ancestral sequence reconstruction

Whole genome viral consensus sequences for each patient were aligned using MAFFT⁷³ with default settings. This alignment was used to create a maximum likelihood tree using RAxML⁷⁴, assuming a general time reversible model of nucleotide substitution under the gamma model of rate heterogeneity. The resulting

tree was rooted at midpoint. Maximum likelihood ancestral sequence reconstruction was performed using RAxML⁷⁴ with the maximum likelihood tree and HCV polyprotein sequences as input.

Association analysis

To test for association between human SNPs and HCV amino acids at genome-to-genome level, we performed logistic regression using PLINK2³² (<https://www.cog-genomics.org/plink2>) adjusted for the human population structure (three first PCs assessed using EIGENSOFT v3.0⁷⁵) and the virus population structure (10 first PCs). For the viral data PCA was performed on the nucleotide data. Tri and quad-allelic sites were converted to binary variables and the amino acid frequencies 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.

To test for association between imputed HLA alleles and HCV amino acids, we used Fisher's exact test, correcting for the virus population structure as described in Bhattacharaya *et al.*⁴². We used the inferred ancestral amino acid to construct a 2x4 contingency table where rows denote presence or absence of a host HLA allele, and the columns denote changes to and away from a specific amino acid in viruses with and without the amino acid inferred to be ancestral. To test for association between *IFNL4* SNP rs12979860 genotypes and HCV amino acids, we used the same Fisher's exact test with a dominant model for *IFNL4* rs12979860 by encoding genotypes as CC and non-CC.

Permutation was used to estimate the FDR for association tests that used Fisher's exact test. The rows of matrix M (where rows correspond to study participants and columns to HLA alleles or the *IFNL4* genotypes) were randomly permuted 500 times and in each case the p-values of the associations were calculated. For each threshold t, the expected number of false significant associations was estimated by the mean number of false positives across permuted null data sets. The FDR for threshold t was then estimated as the mean number of false positives divided by the observed number of significant associations in the actual data at threshold t.

To test for enrichment of association signals in epitope regions, viral proteins or with a specific HLA allele we used Fisher's exact test. Each site is either within a target region (epitope regions or a specific viral protein) or not and the most associated test

with it, is significant or not with an FDR of 5%. The resulting contingency table was tested using Fisher's exact test to assess enrichment or depletion of signals of association.

To assess the relationship between rates of escape and reversion in HLA presentation, we estimated the odds ratio for each 2x2 sub-table used in the Fisher's exact test. This was done only for viral sites associated with HLA alleles at 20% FDR and which there were sufficient observations in both tables to estimate the odds ratio where confidence interval did not go to infinity. Pearson's correlation coefficient was used to assess the relationship between the $\log_{10}(\text{OR})$ of escape and reversion.

To test for enrichment of viral amino acid associations with host *IFNL4* genotypes in viral proteins, the null distribution of number of association in each protein was estimated using 10,000 permutations of *IFNL4* labels and performing the same tests. The estimated null distribution of number of associations for each viral protein was compared to the observed number of associations in the data to test for enrichment or depletion of number of associations. To test if HVR1, HVR2, HVR3, ISDR and PKR-BD regions show differences in the number of changes away from the population consensus in CC and non-CC hosts, we used a Poisson regression. In each individual and each locus we determined the number of differences to the population consensus. We then estimated the effect of *IFNL4* genotypes on the mean number of differences to the population consensus using Poisson regression. The same procedure was used to test if the total number of differences across the whole poly-protein relative to the population consensus was influenced by *IFNL4* genotypes.

To estimate the rate of synonymous and non-synonymous mutations, we used `dndsml` function from MATLAB (Release 2015a, The MathWorks) that uses Goldman and Yang's method. It estimates (using maximum likelihood) an explicit model for codon substitution that takes into account transition/transversion rate bias and base/codon frequency bias. Then it uses the model to correct synonymous and non-synonymous counts to account for multiple substitutions at the same site. To estimate dN and dS, each sequence was compared to the population consensus which indicates the most common nucleotide observed in our data set at each position along the genome.

To determine whether *IFNL4* genotypes impact on HLA alleles presentation of epitopes, logistic regression with interaction term was used. The outcome for individuals was whether on the terminal branches of the tree a specific amino acid had changed or not. We tested for interaction between presence and absence of the associated HLA allele and *IFNL4* genotypes for all combinations of HLA alleles and viral sites associated at a 20% FDR. In addition, we tested for an overall effect of *IFNL4* genotypes on HLA alleles' presentation of epitopes. We used our 2x4 contingency tables and the odds ratios estimated from the 2x2 sub-tables to infer the antigenic amino acids. If the odds ratio indicates that in individuals with HLA allele present, the "X ancestral amino acid -> any other amino acid" element is enriched relative to individuals without the HLA allele then we assume the X amino acid is the antigenic amino acid and escape occurs away from X to any other amino acid. For these cases, we count how many of "X ancestral amino acid -> any other amino acid" occur in CC and non-CC individuals across all combinations of HLA alleles and viral sites associated at 20% FDR. If *IFNL4* genotypes have no impact on the HLA presentation (null hypothesis), then the mean number of escape mutations in CC and non-CC hosts should be proportional to the frequency of hosts with CC and non-CC genotypes (null distribution is binomial with parameters n equal to the total number of observed escape mutations and p equal to the proportion of CC hosts).

We used linear regression in PLINK2³² (<https://www.cog-genomics.org/plink2>) to test for association between human SNPs and log₁₀ transformed pre-treatment viral load (PTVL) including sex and first three PCs of the host genome as covariates. We used linear regression to test for association between HCV amino acids (with a minimal count of 10 at each site) and viral load. We used linear regression in R (version 3.2.4 (2016-03-10))⁷⁶ to analyse the interaction between *IFNL4* genotypes and amino acids at viral site 2414 and to quantify their impact on viral load.

For all viral sites associated with *IFNL4* genotypes at 20% FDR, we assessed if there is a relationship between effect size of non-consensus amino acids on viral load and effect size of *IFNL4* genotypes and changes to non-consensus amino acids. We estimated the odds ratio of enrichment of changes away from the consensus amino acid on the terminal branches of the virus tree. We also estimated the effect size of non-consensus amino acids on the log₁₀ of viral load using a linear regression with *IFNL4* genotype as a covariate. Additionally, we estimated the effect size of non-consensus amino acid on the log₁₀ viral load in CC and non-CC hosts. We used Pearson's correlation coefficient to measure the strength of relationship between the

effect size of non-consensus amino acids on viral load and log of odds ratio of enrichment of non-consensus amino acid changes in CC hosts.

Data and Code Availability

The consensus viral sequences will be made available via GENBANK (<http://www.ncbi.nlm.nih.gov/genbank/>), and the human genotype data will be made available via EGA (<https://www.ebi.ac.uk/ega/>) through the STOP-HCV data access committee <http://www.stop-hcv.ox.ac.uk>. R and MATLAB code used to generate the results and figures from the primary analyses described above are available from the authors on request.

Replication study

To replicate the *IFNL4* SNP rs12979860 results, we ran the association analysis on an independent HCV infected population that was recruited to the FISSION, FUSION and POSITRON phase 3 clinical studies^{50,51}. The Material Transfer Agreements under which the data were shared limited analysis to the NS5B protein. Paired human genome-wide genotyping and HCV sanger sequencing data for NS5B amplicon were obtained from DNA and plasma samples collected from 360 Caucasian patients chronically infected with HCV genotype 2 (N=153) or 3 (N=208). We searched for association between the *IFNL4* SNP rs12979860 and viral position 2570 using logistic regression with the outcome indicated by the presence or absence of amino acid V at position 2570. To help prevent spurious associations due to host and viral stratification, we included human principal components and viral genotype as covariates.

Replicon assay

Cell Culture

Huh7.5-Sec14L2 cells, previously reported⁵⁵ were grown in Dulbecco's modified Eagle's medium (DMEM, Life Technologies) supplemented with 10% fetal calf serum, 2 mM L-glutamine, 100 U/ml penicillin, 100 U/ml streptomycin, 100 mM HEPES and 0.1M nonessential amino acids as described⁷⁷. Huh7.5 cells are heterozygous CT for the *IFNL4* rs12979860 SNP⁷⁸.

HCV Mutant Replicons

The enhanced version of the subgenomic replicon of genotype 3a strain S52, lacking of neomycin resistance gene, has been previously described⁵⁵. Site-specific mutations were introduced using QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies) following manufacturer instructions and confirmed by direct

sequencing. HCV mutant plasmids were linearized by *Xba*I digestion (New England Biolabs; NEB), mung-bean treated (NEB) and purified. Linearized DNA was then used as template for *in vitro* RNA transcription (IVT) (Megascript T7, Life Technologies) according to manufacturer protocol. Finally, IVT RNA has been DNase treated, purified and stored at -80°C.

Electroporation and Luciferase Detection

For electroporation, cells were counted and then washed twice in ice-cold PBS. Typically, for each mutant 4x10⁶ cells were mixed with 1 µg of replicon RNA in a 4mm cuvette and electroporated in the Gene Pulser Xcell (Bio-Rad) at 250 V, 950 µF using exponential decay setting. Cells were immediately recovered in pre-warmed complete DMEM, seeded in a 24-well plate and incubated at 37°C. After 5, 24, 48 or 72 hours, medium was removed and cells lysed with Glo Lysis Buffer (Promega). Cell lysates were then transferred in a white 96-well plate (Corning) and the luciferase expression was quantitated in a luminometer (GloMax 96 Microplate Luminometer, Promega) using Bright-Glo assay system (Promega).

References

1. Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. & Wiersma, S. T. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* **57**, 1333–1342 (2013).
2. Thomas, D. L. *et al.* Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature* **461**, 798–801 (2009).
3. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
4. Rauch, A. *et al.* Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. *Gastroenterology* **138**, 1338–45, 1345–7 (2010).
5. Suppiah, V. *et al.* IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nat. Genet.* **41**, 1100–1104 (2009).
6. Tanaka, Y. *et al.* Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nat. Genet.* **41**, 1105–1109 (2009).
7. Enomoto, N. *et al.* Mutations in the Nonstructural Protein 5a Gene and Response to Interferon in Patients with Chronic Hepatitis C Virus 1b Infection. *N. Engl. J. Med.* **334**, 77–82 (1996).
8. Pascu, M. *et al.* Sustained virological response in hepatitis C virus type 1b infected patients is predicted by the number of mutations within the NS5A-ISDR: a meta-analysis focused on geographical differences. *Gut* **53**, 1345–1351 (2004).
9. Halfon, P. & Locarnini, S. Hepatitis C virus resistance to protease inhibitors. *J. Hepatol.* **55**, 192–206 (2011).
10. Halfon, P. & Sarrazin, C. Future treatment of chronic hepatitis C with direct acting antivirals: Is resistance important? *Liver International* **32**, 79–87 (2012).
11. Ahmed, A. & Felmlee, D. J. Mechanisms of hepatitis C viral resistance to

- 892 direct acting antivirals. *Viruses* **7**, 6716–6729 (2015).
- 893 12. Sarrazin, C. *et al.* Prevalence of Resistance-Associated Substitutions in HCV
894 NS5A, NS5B, or NS3 and Outcomes of Treatment with Ledipasvir and
895 Sofosbuvir. *Gastroenterology* (2016). doi:10.1053/j.gastro.2016.06.002
- 896 13. Messina, J. P. *et al.* Global distribution and prevalence of hepatitis C virus
897 genotypes. *Hepatology* **77**–87 (2014). doi:10.1002/hep.27259
- 898 14. Pol, S., Vallet-Pichard, A. & Corouge, M. Treatment of hepatitis C virus
899 genotype 3-infection. *Liver Int.* **34 Suppl 1**, 18–23 (2014).
- 900 15. Ampuero, J., Romero-Gómez, M. & Reddy, K. R. Review article: HCV
901 genotype 3 – the new treatment challenge. *Aliment. Pharmacol. Ther.* **39**,
902 686–698 (2014).
- 903 16. Fitzmaurice, K. *et al.* Molecular footprints reveal the impact of the protective
904 HLA-A*03 allele in hepatitis C virus infection. *Gut* **60**, 1563–1571 (2011).
- 905 17. Neumann-Haefelin, C. *et al.* Human leukocyte antigen B27 selects for rare
906 escape mutations that significantly impair hepatitis C virus replication and
907 require compensatory mutations. *Hepatology* **54**, 1157–1166 (2011).
- 908 18. Heim, M. H. & Thimme, R. Innate and adaptive immune responses in HCV
909 infections. *J. Hepatol.* **61**, S14–S25 (2014).
- 910 19. Swadling, L. *et al.* A human vaccine strategy based on chimpanzee adenoviral
911 and MVA vectors that primes, boosts, and sustains functional HCV-specific T
912 cell memory. *Sci. Transl. Med.* **6**, 261ra153 (2014).
- 913 20. von Delft, A. *et al.* The broad assessment of HCV genotypes 1 and 3 antigenic
914 targets reveals limited cross-reactivity with implications for vaccine design. *Gut*
915 **65**, 112–123 (2016).
- 916 21. Simmonds, P. Genetic diversity and evolution of hepatitis C virus--15 years on.
917 *J. Gen. Virol.* **85**, 3173–3188 (2004).
- 918 22. Lauck, M. *et al.* Analysis of hepatitis C virus intrahost diversity across the
919 coding region by ultradeep pyrosequencing. *J. Virol.* **86**, 3952–3960 (2012).
- 920 23. Bonsall, D. *et al.* ve-SEQ: Robust, unbiased enrichment for streamlined
921 detection and whole-genome sequencing of HCV and other highly diverse
922 pathogens. *F1000Research* **4**, 1062 (2015).
- 923 24. Thomson, E. *et al.* Comparison of next generation sequencing technologies for
924 the comprehensive assessment of full-length hepatitis C viral genomes. *J.*
925 *Clin. Microbiol.* (2016). doi:10.1128/JCM.00330-16
- 926 25. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**,
927 68–74 (2015).
- 928 26. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev.*
929 *Genomics Hum. Genet.* **10**, 387–406 (2009).
- 930 27. Dilthey, A. *et al.* Multi-Population Classical HLA Type Imputation. *PLoS*
931 *Comput. Biol.* **9**, (2013).
- 932 28. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging.
933 *Pharmacogenomics J.* **14**, 192–200 (2014).
- 934 29. Bartha, I. *et al.* A genome-to-genome analysis of associations between human
935 genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2**, e01123
936 (2013).
- 937 30. Foster, G. R. *et al.* Efficacy of sofosbuvir plus ribavirin with or without
938 peginterferon-alfa in patients with hepatitis C virus genotype 3 infection and
939 treatment-experienced patients with cirrhosis and hepatitis C virus genotype 2
940 infection. *Gastroenterology* **149**, 1462–1470 (2015).
- 941 31. Gonzalez-Galarza, F. F. *et al.* Allele frequency net 2015 update: new features
942 for HLA epitopes, KIR and disease and HLA adverse drug reaction
943 associations. *Nucleic Acids Res.* **43**, D784–D788 (2015).
- 944 32. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger
945 and richer datasets. *Gigascience* **4**, 7 (2015).
- 946 33. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple

947 testing burden for genomewide association studies of nearly all common
948 variants. *Genet. Epidemiol.* **32**, 381–5 (2008).

949 34. Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide
950 association study (GWAS). *BMC Genomics* **11**, 724 (2010).

951 35. de Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for
952 disease association studies in the extended human MHC. *Nat. Genet.* **38**,
953 1166–1172 (2006).

954 36. Malkki, M., Single, R., Carrington, M., Thomson, G. & Petersdorf, E. MHC
955 microsatellite diversity and linkage disequilibrium among common HLA-A,
956 HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic
957 transplantation and disease association studies. *Tissue Antigens* **66**, 114–124
958 (2005).

959 37. Ruhl, M. *et al.* CD8+ T-cell response promotes evolution of hepatitis c virus
960 nonstructural proteins. *Gastroenterology* **140**, 2064–2073 (2011).

961 38. Merani, S. *et al.* Effect of immune pressure on hepatitis C virus evolution:
962 insights from a single-source outbreak. *Hepatology* **53**, 396–405 (2011).

963 39. Rauch, A. *et al.* Divergent adaptation of hepatitis C virus genotypes 1 and 3 to
964 human leukocyte antigen-restricted immune pressure. *Hepatology* **50**, 1017–
965 1029 (2009).

966 40. Gaudieri, S. *et al.* Evidence of Viral Adaptation to HLA Class I-Restricted
967 Immune Pressure in Chronic Hepatitis C Virus Infection. *J. Virol.* **80**, 11094–
968 11104 (2006).

969 41. Pagel, M. The maximum likelihood approach to reconstructing ancestral
970 character states of discrete characters on phylogenies. *Syst. Biol.* **48**, 612–622
971 (1999).

972 42. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV
973 polymorphisms and HLA allele associations. *Science* **315**, 1583–1586 (2007).

974 43. Westfall, P. H. & Young, S. S. *Resampling-Based Multiple Testing: Examples
975 and Methods for P-Value Adjustment*. (Wiley, 1993).

976 44. Meinshausen, N., Maathuis, M. H. & Bühlmann, P. Asymptotic optimality of the
977 Westfall–Young permutation procedure for multiple testing under dependence.
978 *Ann. Stat.* **39**, 3369–3391 (2011).

979 45. Neumann-Haefelin, C. *et al.* Analysis of the evolutionary forces in an
980 immunodominant CD8 epitope in hepatitis C virus at a population level. *J.*
981 *Virol.* **82**, 3438–3451 (2008).

982 46. Bronke, C. *et al.* HIV escape mutations occur preferentially at HLA-binding
983 sites of CD8 T-cell epitopes. *AIDS* **27**, 899–905 (2013).

984 47. Patin, E. *et al.* Genome-wide association study identifies variants associated
985 with progression of liver fibrosis from HCV infection. *Gastroenterology* **143**,
986 1212–1244 (2012).

987 48. Prokunina-Olsson, L. *et al.* A variant upstream of IFNL3 (IL28B) creating a
988 new interferon gene IFNL4 is associated with impaired clearance of hepatitis C
989 virus. *Nat. Genet.* **45**, 164–171 (2013).

990 49. Terczyńska-Dyla, E. *et al.* Reduced IFNL4 activity is associated with improved
991 HCV clearance and reduced expression of interferon-stimulated genes. *Nat.*
992 *Commun.* **5**, 5699 (2014).

993 50. Jacobson, I. M. *et al.* Sofosbuvir for Hepatitis C Genotype 2 or 3 in Patients
994 without Treatment Options. *N. Engl. J. Med.* **368**, 1867–1877 (2013).

995 51. Lawitz, E. *et al.* Sofosbuvir for previously untreated chronic hepatitis C
996 infection. *N. Engl. J. Med.* **368**, 1878–87 (2013).

997 52. Wong, M.-T. & Chen, S. S.-L. Emerging roles of interferon-stimulated genes in
998 the innate immune response to hepatitis C virus infection. *Cell. Mol. Immunol.*
999 (2014). doi:10.1038/cmi.2014.127

1000 53. Saito, T. *et al.* Sequence analysis of PePHD within HCV E2 region and
1001 correlation with resistance of interferon therapy in Japanese patients infected

1002 with HCV genotypes 2a and 2b. *Am. J. Gastroenterol.* **98**, 1377–1383 (2003).

1003 54. Macquillan, G. C. *et al.* Does sequencing the PKRBD of hepatitis C virus

1004 NS5A predict therapeutic response to combination therapy in an Australian

1005 population? *J. Gastroenterol. Hepatol.* **19**, 551–557 (2004).

1006 55. Witteveldt, J., Martin-Gans, M. & Simmonds, P. Enhancement of the

1007 Replication of Hepatitis C Virus Replicons of Genotypes 1 to 4 by Manipulation

1008 of CpG and UpA Dinucleotide Frequencies and Use of Cell Lines Expressing

1009 SECL14L2 for Antiviral Resistance Testing. *Antimicrob. Agents Chemother.*

1010 **60**, 2981–2992 (2016).

1011 56. Yu, M. *et al.* In vitro efficacy of approved and experimental antivirals against

1012 novel genotype 3 hepatitis C virus subgenomic replicons. *Antiviral Res.* **100**,

1013 439–445 (2013).

1014 57. Bibert, S. *et al.* IL28B expression depends on a novel TT/-G polymorphism

1015 which improves HCV clearance prediction. *J. Exp. Med.* **210**, 1109–1116

1016 (2013).

1017 58. Sheahan, T. *et al.* Interferon Lambda Alleles Predict Innate Antiviral Immune

1018 Responses and Hepatitis C Virus Permissiveness. *Cell Host Microbe* **15**, 190–

1019 202 (2014).

1020 59. Ferraris, P. *et al.* Cellular Mechanism for Impaired Hepatitis C Virus Clearance

1021 by Interferon Associated with IFNL3 Gene Polymorphisms Relates to

1022 Intrahepatic Interferon- λ Expression. *Am. J. Pathol.* **186**, 938–951 (2016).

1023 60. Peiffer, K.-H. *et al.* Interferon lambda 4 genotypes and resistance-associated

1024 variants in patients infected with hepatitis C virus genotypes 1 and 3.

1025 *Hepatology* **63**, 63–73 (2016).

1026 61. Pedergrana, V. *et al.* Interferon Lambda 4 variant rs12979860 is not

1027 associated with RAV NS5A Y93H in Hepatitis C Virus Genotype 3a.

1028 *Hepatology* (2016). doi:10.1002/hep.28533

1029 62. McHutchison, J. G. *et al.* Peginterferon alfa-2b or alfa-2a with ribavirin for

1030 treatment of hepatitis C infection. *N. Engl. J. Med.* **361**, 580–593 (2009).

1031 63. Bochud, P.-Y. *et al.* IL28B alleles associated with poor hepatitis C virus (HCV)

1032 clearance protect against inflammation and fibrosis in patients infected with

1033 non-1 HCV genotypes. *Hepatology* **55**, 384–394 (2012).

1034 64. Thompson, A. J. *et al.* Interleukin-28B Polymorphism Improves Viral Kinetics

1035 and Is the Strongest Pretreatment Predictor of Sustained Virologic Response

1036 in Genotype 1 Hepatitis C Virus. *Gastroenterology* **139**, (2010).

1037 65. Tellinghuisen, T. L., Foss, K. L. & Treadaway, J. Regulation of hepatitis C

1038 virion production via phosphorylation of the NS5A protein. *PLoS Pathog.* **4**,

1039 e1000032 (2008).

1040 66. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte

1041 Antigens. *PLoS One* **8**, (2013).

1042 67. Lamble, S. *et al.* Improved workflows for high throughput library preparation

1043 using the transposome-based Nextera system. *BMC Biotechnol.* **13**, 104

1044 (2013).

1045 68. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: Quantification and

1046 annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).

1047 69. Martin, M. Cutadapt removes adapter sequences from high-throughput

1048 sequencing reads. *EMBnet.journal* **17**, 10 (2011).

1049 70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.

1050 *Nat Methods* **9**, 357–359 (2012).

1051 71. Smith, D. B. *et al.* Expanded classification of hepatitis C virus into 7 genotypes

1052 and 67 subtypes: updated criteria and genotype assignment web resource.

1053 *Hepatology* **59**, 318–327 (2014).

1054 72. Yang, X. *et al.* De novo assembly of highly diverse viral populations. *BMC*

1055 *Genomics* **13**, 475 (2012).

1056 73. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software

1057 version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**,
1058 772–780 (2013).

1059 74. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-
1060 analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

1061 75. Price, A. L. *et al.* Principal components analysis corrects for stratification in
1062 genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

1063 76. R Core Team (2016). R: A language and environment for statistical computing.
1064 R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-](https://www.R-project.org/)
1065 [project.org/](https://www.R-project.org/).

1066 77. Magri, A. *et al.* Rethinking the old antiviral drug moroxydine: Discovery of
1067 novel analogues as anti-hepatitis C virus (HCV) agents. *Bioorg. Med. Chem.*
1068 *Lett.* **25**, 5372–5376 (2015).

1069 78. Rojas, Á. *et al.* Hepatitis C virus infection alters lipid metabolism depending on
1070 IL28B polymorphism and viral genotype and modulates gene expression in
1071 vivo and in vitro. *J. Viral Hepat.* **21**, 19–24 (2014).

1072

1073

1074 **Table 1.** Associations between HLA alleles and viral amino acids at a 5% false
1075 discovery rate in the whole cohort (All) or in HCV genotype 3a self-reported “White”
1076 patients only (G3a White). Associations found in the whole cohort only are coloured
1077 in green, in the G3a White cohort only in orange and in both cohorts in blue. For
1078 each combination of HLA allele and viral site, only the most significant associated
1079 amino acid is reported. The first amino acid at “variable amino acids at this site”
1080 column indicates the most common amino acid at the site in our data.
1081

HLA allele	HCV amino acid position	Viral protein	Variable amino acids at this site	Associated amino acid	P value All	q value All	P value G3a White	q value G3a White	In a known epitope ?
B*07:02	42	C	PL	P	1.55x10 ⁻⁰⁷	5.00x10 ⁻⁰³	8.07x10 ⁻⁰⁷	5.81x10 ⁻⁰³	Yes
C*07:02	42	C	PL	P	2.27x10 ⁻⁰⁷	6.10x10 ⁻⁰³	1.99x10 ⁻⁰⁷	1.33x10 ⁻⁰³	Yes
B*41:02	109	C	PQSAL	S	4.29x10 ⁻⁰⁶	8.09x10 ⁻⁰²	9.84x10 ⁻⁰⁶	4.93x10 ⁻⁰²	No
B*13:02	348	E1	ILVM	I	8.13x10 ⁻⁰⁶	1.04x10 ⁻⁰¹	1.25x10 ⁻⁰⁶	8.78x10 ⁻⁰³	No
C*08:02	372	E1	ATVIFC	I	7.33x10 ⁻⁰⁴	4.12x10 ⁻⁰¹	9.41x10 ⁻⁰⁶	4.93x10 ⁻⁰²	No
A*29:02	444	E2	YHFRVWL	Y	1.85x10 ⁻⁰⁴	2.89x10 ⁻⁰¹	4.02x10 ⁻⁰⁶	2.39x10 ⁻⁰²	No
C*15:02	561	E2	VTLI	L	5.70x10 ⁻⁰⁷	1.50x10 ⁻⁰²	6.80x10 ⁻⁰⁴	5.92x10 ⁻⁰¹	No
DRB1*14:04	905	NS2	ATSC	S	4.45x10 ⁻⁰³	5.19x10 ⁻⁰¹	5.58x10 ⁻⁰⁷	4.20x10 ⁻⁰³	No
A*31:01	1270	NS3	RKHS	R	1.95x10 ⁻⁰⁹	4.44x10 ⁻⁰⁴	6.20x10 ⁻⁰⁸	3.64x10 ⁻⁰⁴	Yes
A*33:03	1282	NS3	NVTSA	T	1.73x10 ⁻⁰⁷	5.26x10 ⁻⁰³	4.14x10 ⁻⁰²	9.13x10 ⁻⁰¹	Yes
B*15:01	1290	NS3	KPARS	R	1.25x10 ⁻⁰⁷	4.47x10 ⁻⁰³	3.64x10 ⁻⁰⁶	2.39x10 ⁻⁰²	Yes
A*68:02	1341	NS3	VA	V	2.43x10 ⁻⁰⁹	4.44x10 ⁻⁰⁴	9.93x10 ⁻⁰⁷	6.91x10 ⁻⁰³	No
A*68:02	1344	NS3	TVA	T	3.68x10 ⁻¹⁰	<4.00x10 ⁻⁰⁴	9.22x10 ⁻¹¹	<2.50x10 ⁻⁰⁴	No
B*51:01	1380	NS3	ILV	L	3.22x10 ⁻⁰⁸	1.67x10 ⁻⁰³	2.10x10 ⁻⁰⁷	1.38x10 ⁻⁰³	Yes
A*01:01	1444	NS3	FY	F	9.63x10 ⁻³³	<4.00x10 ⁻⁰⁴	7.18x10 ⁻²⁴	<2.50x10 ⁻⁰⁴	Yes
A*68:02	1452	NS3	IV	I	6.98x10 ⁻⁰⁹	4.44x10 ⁻⁰⁴	3.84x10 ⁻⁰⁷	2.89x10 ⁻⁰³	No
A*30:02	1585	NS3	YF	Y	1.19x10 ⁻⁰⁶	3.28x10 ⁻⁰²	6.89x10 ⁻⁰⁵	1.90x10 ⁻⁰¹	No
B*13:02	1635	NS3	ITVLAF	T	7.38x10 ⁻⁰⁸	3.13x10 ⁻⁰³	1.65x10 ⁻⁰⁷	1.00x10 ⁻⁰³	No
B*08:01	1646	NS3	MTAVSI	T	2.89x10 ⁻¹²	<4.00x10 ⁻⁰⁴	2.01x10 ⁻¹⁰	<2.50x10 ⁻⁰⁴	No
C*07:01	1646	NS3	MTAVSI	T	3.01x10 ⁻⁰⁷	7.39x10 ⁻⁰³	1.40x10 ⁻⁰⁸	<2.50x10 ⁻⁰⁴	No
DQA1*05:01	1646	NS3	MTAVSI	T	7.24x10 ⁻⁰⁷	1.76x10 ⁻⁰²	1.18x10 ⁻⁰⁵	5.64x10 ⁻⁰²	No
B*57:01	1759	NS4B	AVTGNSDI	A	3.38x10 ⁻⁰⁵	1.83x10 ⁻⁰¹	4.01x10 ⁻⁰⁶	2.39x10 ⁻⁰²	No
A*02:01	1873	NS4B	LFKICVPRM TA	F	4.06x10 ⁻¹⁰	<4.00x10 ⁻⁰⁴	2.58x10 ⁻⁰⁸	2.50x10 ⁻⁰⁴	No
B*38:01	2034	NS5A	STNLPIAQV DKEMGRFW	P	7.56x10 ⁻⁰⁸	3.13x10 ⁻⁰³	6.35x10 ⁻⁰⁸	3.64x10 ⁻⁰⁴	Yes
C*12:03	2034	NS5A	STLPQKVAIR MFW	P	9.31x10 ⁻⁰⁶	1.05x10 ⁻⁰¹	3.99x10 ⁻⁰⁶	2.39x10 ⁻⁰²	Yes
B*18:01	2144	NS5A	ED	E	1.16x10 ⁻⁰⁶	3.28x10 ⁻⁰²	2.36x10 ⁻⁰⁶	1.67x10 ⁻⁰²	Yes
B*51:01	2148	NS5A	MTVSLKIA	V	2.77x10 ⁻⁰⁶	6.07x10 ⁻⁰²	4.60x10 ⁻⁰⁷	3.16x10 ⁻⁰³	Yes
C*14:02	2148	NS5A	MTVSLKIA	V	1.91x10 ⁻⁰⁵	1.34x10 ⁻⁰¹	2.32x10 ⁻⁰⁶	1.67x10 ⁻⁰²	Yes
B*40:01	2248	NS5A	TSKAN	T	5.76x10 ⁻⁰⁴	3.91x10 ⁻⁰¹	9.62x10 ⁻⁰⁶	4.93x10 ⁻⁰²	No
DQA1*01:01	2486	NS5B	TIANSV	T	4.71x10 ⁻⁰⁵	1.96x10 ⁻⁰¹	5.12x10 ⁻⁰⁶	2.90x10 ⁻⁰²	Yes
A*31:01	2510	NS5B	AKQSEMGLT V	A	4.96x10 ⁻⁰⁸	2.14x10 ⁻⁰³	1.79x10 ⁻⁰⁶	1.35x10 ⁻⁰²	No
A*32:01	2537	NS5B	NDHSYEA	N	1.13x10 ⁻⁰⁸	7.27x10 ⁻⁰⁴	4.77x10 ⁻⁰⁶	2.75x10 ⁻⁰²	Yes
A*32:01	2540	NS5B	RSKHNQC	R	7.45x10 ⁻⁰⁹	6.00x10 ⁻⁰⁴	5.96x10 ⁻¹²	<2.50x10 ⁻⁰⁴	Yes

<i>A*02:11</i>	2600	NS5B	QKRSAL	Q	1.01×10^{-06}	3.12×10^{-02}	NA	NA	Yes
<i>A*26:01</i>	2605	NS5B	EAGVK	G	2.61×10^{-07}	6.64×10^{-03}	5.91×10^{-05}	1.73×10^{-01}	Yes
<i>B*51:01</i>	2713	NS5B	ILMV	I	7.70×10^{-06}	1.03×10^{-01}	1.25×10^{-08}	$< 2.50 \times 10^{-04}$	No
<i>A*25:01</i>	2821	NS5B	RCQL	R	8.31×10^{-06}	1.04×10^{-01}	2.51×10^{-06}	1.73×10^{-02}	No

1082

1083

1084

1085