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

M. Azim Ansari[#1,2], **Vincent Pedergnana**[#1], **Camilla Ip**[1], **Andrea Magri**[2], **Annette Von Delft**[3], **David Bonsall**[3], **Nimisha Chaturvedi**[4], **Istvan Bartha**[4], **David Smith**[3], **George Nicholson**[5], **Gilean McVean**[1,6], **Amy Trebes**[1], **Paolo Piazza**[1], **Jacques Fellay**[4], **Graham Cooke**[7], **Graham R Foster**[8], **STOP-HCV consortium**, **Emma Hudson**[3], **John McLauchlan**[9], **Peter Simmonds**[3], **Rory Bowden**[1], **Paul Klenerman**[3], **Eleanor Barnes**[3,*], and **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 [10]University of Nottingham, Queen's Medical Centre, Nottingham, NG7 2UH [11]Gilead Sciences, Inc., Foster City, California, United States [12]Conatus Pharmaceuticals, 16745 West Bernardo Drive, Suite 200, San Diego, CA 92127 [13]University of Dundee, Ninewells Hospital & Medical School, Dundee, DD1 9SY [14]Hepatitis C

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

Contributor information
STOP-HCV consortium:
Prof Eleanor Barnes[3], Prof Jonathan Ball[10], Dr Diana Brainard[11], Dr Gary Burgess[12], Dr Graham Cooke[7], Prof John Dillon[13], Prof Graham Foster[8], Mr Charles Gore[14], Dr Neil Guha[10], Ms Rachel Halford[14], Dr Cham Herath[15], Prof Chris Holmes[5], Dr Anita Howe[16], Dr Emma Hudson[3], Prof William Irving[10], Prof Salim Khakoo[17], Prof Paul Klenerman[3], Dr Diana Koletzki[18], Dr Natasha Martin[19], Dr Benedetta Massetto[11], Dr Tamyo Mbisa[20], Dr John McHutchison[11], Prof Jane McKeating[3], Dr John McLauchlan[9], Dr Alec Miners[21], Dr Andrea Murray[22], Dr Peter Shaw[23], Prof Peter Simmonds[3], Dr Chris Spencer[1], Dr Paul Targett-Adams[24], Dr Emma Thomson[9], Prof Peter Vickerman[25], Prof Nicole Zitzmann[26]

Trust, 27 Crosby Row, London SE1 3YD [15]Gilead Sciences, Stockley Park, 2 Roundwood Ave, Middlesex UB11 1AF [16]BC Centre for Excellence in HIV/AIDS, St. Paul's Hospital, 608–1081 Burrard Street, Canada V6Z 1Y6 [17]University of Southampton, University Road, Southampton, SO17 1BJ [18]Janssen Diagnostics, Turnhoutseweg, 30, 2340 Beerse, Belgium [19]UC San Diego, La Jolla, CA 92093-0507, USA [20]Public Health England, 61 Colindale Ave, London NW9 5EQ [21]London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH [22]OncImmune Limited, Clinical Sciences Building, Nottingham City Hospital, Hucknall Road, Nottingham NG5 1PB [23]Merck & Co., Inc., Kenilworth, NJ 07033, USA [24]Medivir AB, Box 1086, 141 22 Huddinge, Sweden [25]University of Bristol, Oakfield House, Oakfield Grove, Clifton BS8 2BN [26]University of Oxford, South Parks Road, Oxford OX1 3QU

[#] These authors contributed equally to this work.

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. Viral genotypes and distinct viral genetic motifs have been associated with the response to interferon based therapies7,8 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 lack of data addressing other genotypes. HCV genotype 3 is of particular interest as this genotype is known to infect 53 million people globally13 and is associated with a higher failure rate to DAA therapies14,15.

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 genome16,17. 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 ("escape" mutants) that abrogate T cell

recognition18. Understanding how host HLA molecules impact on viral selection has important implications for the development of HCV T-cell vaccines that aim to prevent infection19,20. A comprehensive host genome to viral genome analysis at scale will assess the relative contribution of host HLA molecules in driving HCV genetic change and this analysis may also reveal other host genes that play a key role in shaping the HCV genome.

We generated data from a cohort of 601 HCV infected patients (from the BOSON21 clinical trial) to systematically look for associations between host and virus genomes, exploiting the fact that while host genetics remain fixed, the virus mutates, allowing it to evolve during infection. For this we developed targeted viral enrichment22,23 to obtain whole HCV genomes, and used high-throughput genotyping arrays in combination with statistical imputation to obtain data on HLA alleles24,25 and nucleotide polymorphisms across the human genome. We provide evidence that polymorphisms relevant to the innate (*IFNL4*) and adaptive immune systems (HLA genes) are associated with HCV sequence polymorphisms. We show that an interaction between host *IFNL4* genotypes and a viral site in the NS5A protein determines HCV viral load. By assessing viral evolution in individuals with different *IFNL4* genotypes, we highlight systematic differences in the innate immune response and discuss how these might relate to previous associations with spontaneous clearance and clinical treatment. We demonstrate the potential for a joint analysis of host and viral genomic data to provide information on underlying molecular interactions and their importance in treating and preventing HCV, and other viral infections, in the era of genomic analysis.

Results

Sample description and genetic structure

DNA samples from 567 patients (out of 601 patients) were genotyped using the Affymetrix UK Biobank array. This array genotypes over 800,000 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 sequencing22.

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 (Supplementary Table 1). 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 1), which differed

significantly in some of the inferred HLA alleles frequencies (Supplementary Table 2) consistent with previous observations[26]. 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 2). 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)). We observed that patients of specific ancestries, measured either by genetic ancestry (PCs), or self-reported ethnicity, clustered together on the tree of viral diversity (Supplementary Figure 2). A PCA on virus nucleotides sequence data reflected the structure of the tree, specifically at the level of virus subtypes (Supplementary Figure 3).

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. We performed logistic regression assuming an additive model, and adjusting for sex, and population structure by including the first three PCs of the host and the first ten PCs of the virus as covariates. Failure to control for either covariates leads to a significant inflation in the association test statistics (Supplementary Figure 4), as would be expected given the observed correlation in population structure of the virus and the host (Supplementary Figure 2). Assuming a human genome-wide significance threshold[27] of 5×10^{-8}, and that amino acid variants in the viral genome are approximately uncorrelated once the population structure is accounted for, then a Bonferroni correction[28] results in a significance threshold of approximately 2×10^{-11}.

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) (Figure 1). Three other associations were observed between multiple SNPs in the host MHC and virus amino acids in NS3 and NS4B proteins ($P<4 \times 10^{-9}$, Figure 1). 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 ($P=1.98 \times 10^{-9}$, Figure 1). 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.

We observed 182 associations between human SNPs mapping to two loci and five HCV amino acid sites ($P<4 \times 10^{-9}$) (Supplementary Table 3). 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.

Host 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[29,30]. 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 ("escape mutations")[16,31–34]. Upon transmission to another host, with a different HLA repertoire, reversion to the wild type may occur ("reversion mutations"). To test for footprinting of the host HLA alleles along the viral genome (Figure 2a and Table 1), we inferred the changes on the terminal branches of the virus phylogenetic tree for each amino acid site and assessed association with the hosts' HLA alleles. We repeated the analysis on patients infected with genotype 3a and self-reporting as White (Table 1 and Supplementary Figure 5).

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 4). 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 epitopes[20] (Table 1), which represents a strong enrichment (odds ratio, OR=5.2, P=2.8x10^{-4}). We also observed that NS3 protein was strongly enriched for association signals with HLA alleles (OR=5.5, P=2.2x10^{-4}). The strongest HLA footprints 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*. (See Supplementary notes and Supplementary Figures 6, 7 and 8 for detailed description of the most significant association between *HLA-A* and Y1444F). 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 9). Using the 27 HLA and viral amino acid associations at a 20% FDR that 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 10). 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 host 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[2,3,35]. 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=1.98x10^{-9}). For SNP rs12979860, 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* expression[36]. In our cohort, only two individuals had discordant genotypes for rs12979860 and rs368234815 (which was imputed), resulting in a strong linkage disequilibrium (r^2=0.99) between these two SNPs.

Therefore, individuals with non-CC genotypes express *IFNL4* and have increased expression of hepatic interferon stimulated genes (ISGs)37.

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=1.94x10^{-11}, Figure 2b). We replicated this association in an independent study of 360 patients of European ancestry chronically infected with HCV genotypes 2 or 338,39 (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)32 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 5). Using a permutation approach, we found that the core protein is nominally (P<0.05) enriched, and the NS5A protein is nominally depleted, in the number of associations with the *IFNL4* genotypes. Genotype 3a viral sequences in individuals with the CC genotype did have more mutations away from the population consensus than non-CC individuals (See Supplementary Note and Supplementary Table 6). One of the sites (position 109 in the core protein) is also associated with *HLA-B*41:02* (P=4.3x10^{-6}, Table 1). However, we did not find any consistent evidence for interaction between by HLA alleles and *IFNL4* genotypes and their association with viral amino acids (Supplementary Figure 11). Nor did we find strong differences in the mean number of escape mutations in CC and non-CC patients when comparing HLA allele carriers and non-HLA allele carriers.

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 3). Whilst there was no difference in dS in CC and non-CC patients (P=0.68, Figure 3a), dN was significantly higher in CC patients (P=1.6 x10^{-8}, Figure 3b). The lower dN/dS ratio in non-CC patients (P=1.3x10^{-10}) is potentially indicative that the virus is under a stronger purifying selection (Figure 3c). 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 3d).

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 12). 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 13).

Host and virus genetic determinants of viral load

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 of pre-treatment viral load ($\log_{10}$PTVL) (Figure 4a). We replicated the known association between *IFNL4* variants on chromosome 19 and viral load3 (rs12979860, P=5.9x10^{-10}) with the non-CC genotypes conferring an approximately 0.45-fold decrease in viral load (mean for non-CC=3.4760x10^6 IU/mL and for CC=6.3447x10^6 IU/mL). We also performed a genome-wide association study to detect associations between viral amino acids of genotype 3a virus 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=9.21x10^{-7}, Figure 4b). 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 significant (P=9.37x10^{-9}, Figure 4c). 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 4d). 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 4c and 4d). The interaction is statistically significant when analysing either the whole cohort (P=0.017) or just patients with genotype 3a infections who self-report as being White (P=0.017). 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 =2.81x10^6 IU/mL and all other combinations =6.47x10^6 IU/mL). Introduction of the change from a serine to an asparagine at position 2414 into a modified S52 replicon of genotype 3a40 resulted in a approximately 10-fold increase in replication in Huh7.5 cell culture (Supplementary Note and Supplementary Figure 14).

We also observed that non-consensus amino acids, which are increased in frequency in individuals with a CC genotype, tend to also associate with increased viral load (r=0.42, P=0.005, Supplementary Figure 15). 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 16).

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. Confirmation of the specific associations reported here, and the extent to which they are specific to viral genotype, will require replication analysis in independent cohorts.

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 epitopes[20], 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 development[20].

In addition to the HLA, we now show that IFNL4 activity may significantly shape the HCV viral genome. Previous studies have shown that the "favourable" CC *IFNL4* genotype increases the chances of spontaneous resolution and interferon-based treatment success[2–6]. The *IFNL4* TT/TT genotype (rs368234815), which is strongly linked to the "favourable" CC genotype (rs12979860), abolishes the expression of *IFNL4*[46], whereas in individuals with the *IFNL4* ΔG/TT or ΔG/ΔG genotypes (linked to "unfavourable" non-CC genotypes), *IFNL4* is expressed, leading to the downstream up-regulation of hepatic ISGs expression via the JAK-STAT pathway[37]. The expression of ISGs has been shown to render the host less susceptible to exogenous IFNα/γ stimulation and is associated with more infected cells in the liver[42]. 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[42,43].

Our analysis adds to this hypothesis, with the observation that *IFNL4* variants also impact on the HCV 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 genotypes 2 and 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 study[32] in a single source infection cohort. *IFNL4* has also been associated with a viral mutation associated with DAA resistance in HCV genotype 1 infection[44] although this association has not been replicated in our HCV genotype 3 cohort[45]. The additional signals in NS5B protein associated with *IFNL4* genotypes (or HLA alleles) 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 exact tests could not be performed in an equivalent way (only NS5B sequences were available in the replication cohort). 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.

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 therapy46 (independent of *IFNL4* status). Paradoxically, the "favourable" *IFNL4* variants have also been associated with both an increase in disease progression47 and high viral load3,48. We report an association between a virus amino acid site (serine vs. non-serine at position 2414 in NS5A) and HCV viral load. This site is one of the 11 sites that are putatively associated with *IFNL4* genotypes in our dataset. 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 5). Our *in vitro* data from a genotype 3 replicon assay shows that a change from a serine to asparagine at site 2414 is associated with an increase in RNA replication and perhaps hyper-phosphorylation49, which is a negative regulator of virus replication.

Our interpretation of the data (Figure 5) is that the expression of *IFNL4* by ΔG/TT or ΔG/ΔG 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, 4 and Supplementary Figure 15). Our data suggest that this also leads to an overall increase in the strength of purifying selection (decrease in dN/dS, Figure 3), and together this leads to lower viral load. However, viruses that establish chronic infections in non-CC patients have evolved to survive in a 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 3). It is possible that this then permits a range of mutations which confer higher replicative fitness and therefore higher viral load (Supplementary Figure 15), 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.

In conclusion, we provide a comprehensive genome-to-genome analysis in chronic HCV infection. 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 amino acid polymorphisms 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.

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 infection21. 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:0224 accessed 22 March 2015. HLA amino acids were also imputed by SNP2HLA50 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 PLINK251 (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)22, 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 primers52 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 v7.012053 and adapter sequences with CutAdapt version 1.7.154 and subsequently discarded if either read had less than 50b remaining sequence or if both reads matched the human reference sequence using Bowtie version 2.2.455. The remaining read pool was screened against a BLASTn database containing 165 HCV genomes56 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 v1.357 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 MAFFT58 with default settings. This alignment was used to create a maximum likelihood tree using RAxML59, 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 RAxML59 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 PLINK251 (https://www.cog-genomics.org/plink2) adjusted for the human population structure (three first PCs assessed using EIGENSOFT v3.060) 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 Bhatacharaya *et al.*61. We used the inferred ancestral amino acid to estimate changes for each site on the terminal branches of the virus phylogenetic tree. Inferring the changes along the terminal branches of the tree aims to control for the confounding between host and virus population structures61 by looking at viral mutations after infection. We 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 had changed amino 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 PLINK251 (https://www.cog-genomics.org/plink2) to test for association between human SNPs and $\log_{10}$ 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))62 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 in CC and non-CC genotypes (2x2 contingency table, only using data in which the consensus is inferred to be ancestral). We also estimated the effect size of non-consensus

amino acids on the $\log_{10}$ 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 $\log_{10}$ 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.

Code Availability

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[38,39]. 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 reported[40] 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 0 enicillin, 100 100 lmM HEPES and 0.1M nonessential amino acids as described[63]. Huh7.5 cells are heterozygous CT for the *IFNL4* rs12979860 SNP[64].

HCV Mutant Replicons—The enhanced version of the subgenomic replicon of genotype 3a strain S52, lacking of neomycin resistance gene, has been previously described[40]. 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 *XbaI* 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 4×10^6 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).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology. 2013; 57:1333–1342. [PubMed: 23172780]

2. Thomas DL, et al. Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. Nature. 2009; 461:798–801. [PubMed: 19759533]

3. Ge D, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature. 2009; 461:399–401. [PubMed: 19684573]

4. Rauch A, et al. Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology. 2010; 138:1338–45. 1345–7. [PubMed: 20060832]

5. Suppiah V, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet. 2009; 41:1100–1104. [PubMed: 19749758]

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. 2009; 41:1105–1109. [PubMed: 19749757]

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. 1996; 334:77–82. [PubMed: 8531962]

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. 2004; 53:1345–1351. [PubMed: 15306598]

9. Halfon P, Locarnini S. Hepatitis C virus resistance to protease inhibitors. J Hepatol. 2011; 55:192–206. [PubMed: 21284949]

10. Halfon P, Sarrazin C. Future treatment of chronic hepatitis C with direct acting antivirals: Is resistance important? Liver International. 2012; 32:79–87.

11. Ahmed A, Felmlee DJ. Mechanisms of hepatitis C viral resistance to direct acting antivirals. Viruses. 2015; 7:6716–6729. [PubMed: 26694454]

12. Sarrazin C, et al. Prevalence of Resistance-Associated Substitutions in HCV NS5A, NS5B, or NS3 and Outcomes of Treatment with Ledipasvir and Sofosbuvir. Gastroenterology. 2016; doi: 10.1053/j.gastro.2016.06.002

13. Messina JP, et al. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology. 2014; 77:87.doi: 10.1002/hep.27259

14. Pol S, Vallet-Pichard A, Corouge M. Treatment of hepatitis C virus genotype 3-infection. Liver Int. 2014; 34(Suppl 1):18–23. [PubMed: 24373074]

15. Ampuero J, Romero-Gómez M, Reddy KR. Review article: HCV genotype 3 – the new treatment challenge. Aliment Pharmacol Ther. 2014; 39:686–698. [PubMed: 24612116]

16. Fitzmaurice K, et al. Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus infection. Gut. 2011; 60:1563–1571. [PubMed: 21551190]

17. Neumann-Haefelin C, et al. Human leukocyte antigen B27 selects for rare escape mutations that significantly impair hepatitis C virus replication and require compensatory mutations. Hepatology. 2011; 54:1157–1166. [PubMed: 22006856]

18. Heim MH, Thimme R. Innate and adaptive immune responses in HCV infections. J Hepatol. 2014; 61:S14–S25. [PubMed: 25443342]

19. 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; 6:261ra153.

20. von Delft A, et al. The broad assessment of HCV genotypes 1 and 3 antigenic targets reveals limited cross-reactivity with implications for vaccine design. Gut. 2016; 65:112–123. [PubMed: 26092843]

21. Foster GR, 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; 149:1462–1470. [PubMed: 26248087]

22. Bonsall D, et al. ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. F1000Research. 2015; 4:1062. [PubMed: 27092241]

23. Thomson E, et al. Comparison of next generation sequencing technologies for the comprehensive assessment of full-length hepatitis C viral genomes. J Clin Microbiol. 2016; doi: 10.1128/JCM. 00330-16

24. Dilthey A, et al. Multi-Population Classical HLA Type Imputation. PLoS Comput Biol. 2013; 9

25. Zheng X, et al. HIBAG--HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014; 14:192–200. [PubMed: 23712092]

26. Gonzalez-Galarza FF, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res. 2015; 43:D784–D788. [PubMed: 25414323]

27. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008; 32:381–5. [PubMed: 18348202]

28. Johnson RC, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010; 11:724. [PubMed: 21176216]

29. de Bakker PIW, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006; 38:1166–1172. [PubMed: 16998491]

30. Malkki M, Single R, Carrington M, Thomson G, Petersdorf E. MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies. Tissue Antigens. 2005; 66:114–124. [PubMed: 16029431]

31. Ruhl M, et al. CD8+ T-cell response promotes evolution of hepatitis c virus nonstructural proteins. Gastroenterology. 2011; 140:2064–2073. [PubMed: 21376049]

32. Merani S, et al. Effect of immune pressure on hepatitis C virus evolution: insights from a single-source outbreak. Hepatology. 2011; 53:396–405. [PubMed: 21246583]

33. Rauch A, et al. Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen-restricted immune pressure. Hepatology. 2009; 50:1017–1029. [PubMed: 19670417]

34. Gaudieri S, et al. Evidence of Viral Adaptation to HLA Class I-Restricted Immune Pressure in Chronic Hepatitis C Virus Infection. J Virol. 2006; 80:11094–11104. [PubMed: 17071929]

35. Patin E, et al. Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. Gastroenterology. 2012; 143:1212–1244.

36. Prokunina-Olsson L, et al. A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. Nat Genet. 2013; 45:164–171. [PubMed: 23291588]

37. Terczy ska-Dyla E, et al. Reduced IFNλ4 activity is associated with improved HCV clearance and reduced expression of interferon-stimulated genes. Nat Commun. 2014; 5:5699. [PubMed: 25534433]

38. Jacobson IM, et al. Sofosbuvir for Hepatitis C Genotype 2 or 3 in Patients without Treatment Options. N Engl J Med. 2013; 368:1867–1877. [PubMed: 23607593]

39. Lawitz E, et al. Sofosbuvir for previously untreated chronic hepatitis C infection. N Engl J Med. 2013; 368:1878–87. [PubMed: 23607594]

40. Witteveldt J, Martin-Gans M, Simmonds P. Enhancement of the Replication of Hepatitis C Virus Replicons of Genotypes 1 to 4 by Manipulation of CpG and UpA Dinucleotide Frequencies and Use of Cell Lines Expressing SECL14L2 for Antiviral Resistance Testing. Antimicrob Agents Chemother. 2016; 60:2981–2992. [PubMed: 26953209]

41. Bibert S, et al. IL28B expression depends on a novel TT/-G polymorphism which improves HCV clearance prediction. J Exp Med. 2013; 210:1109–1116. [PubMed: 23712427]

42. Sheahan T, et al. Interferon Lambda Alleles Predict Innate Antiviral Immune Responses and Hepatitis C Virus Permissiveness. Cell Host Microbe. 2014; 15:190–202. [PubMed: 24528865]

43. Ferraris P, et al. Cellular Mechanism for Impaired Hepatitis C Virus Clearance by Interferon Associated with IFNL3 Gene Polymorphisms Relates to Intrahepatic Interferon-λ Expression. Am J Pathol. 2016; 186:938–951. [PubMed: 26896692]

44. Peiffer K-H, et al. Interferon lambda 4 genotypes and resistance-associated variants in patients infected with hepatitis C virus genotypes 1 and 3. Hepatology. 2016; 63:63–73. [PubMed: 26406534]

45. Pedergnana V, et al. Interferon Lambda 4 variant rs12979860 is not associated with RAV NS5A Y93H in Hepatitis C Virus Genotype 3a. Hepatology. 2016; doi: 10.1002/hep.28533

46. McHutchison JG, et al. Peginterferon alfa-2b or alfa-2a with ribavirin for treatment of hepatitis C infection. N Engl J Med. 2009; 361:580–593. [PubMed: 19625712]

47. Bochud P-Y, et al. IL28B alleles associated with poor hepatitis C virus (HCV) clearance protect against inflammation and fibrosis in patients infected with non-1 HCV genotypes. Hepatology. 2012; 55:384–394. [PubMed: 22180014]

48. Thompson AJ, et al. Interleukin-28B Polymorphism Improves Viral Kinetics and Is the Strongest Pretreatment Predictor of Sustained Virologic Response in Genotype 1 Hepatitis C Virus. Gastroenterology. 2010; 139

49. Tellinghuisen TL, Foss KL, Treadaway J. Regulation of hepatitis C virion production via phosphorylation of the NS5A protein. PLoS Pathog. 2008; 4:e1000032. [PubMed: 18369478]

50. Jia X, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. PLoS One. 2013; 8

51. Chang CC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. [PubMed: 25722852]

52. Lamble S, et al. Improved workflows for high throughput library preparation using the transposome-based Nextera system. BMC Biotechnol. 2013; 13:104. [PubMed: 24256843]

53. Gaidatzis D, Lerch A, Hahne F, Stadler MB. QuasR: Quantification and annotation of short reads in R. Bioinformatics. 2015; 31:1130–1132. [PubMed: 25417205]

54. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17:10.

55. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]

56. Smith DB, et al. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. Hepatology. 2014; 59:318–327. [PubMed: 24115039]

57. Yang X, et al. De novo assembly of highly diverse viral populations. BMC Genomics. 2012; 13:475. [PubMed: 22974120]

58. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013; 30:772–780. [PubMed: 23329690]

59. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. [PubMed: 24451623]

60. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

61. Bhattacharya T, et al. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science. 2007; 315:1583–1586. [PubMed: 17363674]

62. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2016. URL https://www.R-project.org/

63. Magri A, et al. Rethinking the old antiviral drug moroxydine: Discovery of novel analogues as anti-hepatitis C virus (HCV) agents. Bioorg Med Chem Lett. 2015; 25:5372–5376. [PubMed: 26428870]

64. Rojas Á, et al. Hepatitis C virus infection alters lipid metabolism depending on IL28B polymorphism and viral genotype and modulates gene expression in vivo and in vitro. J Viral Hepat. 2014; 21:19–24. [PubMed: 24188401]
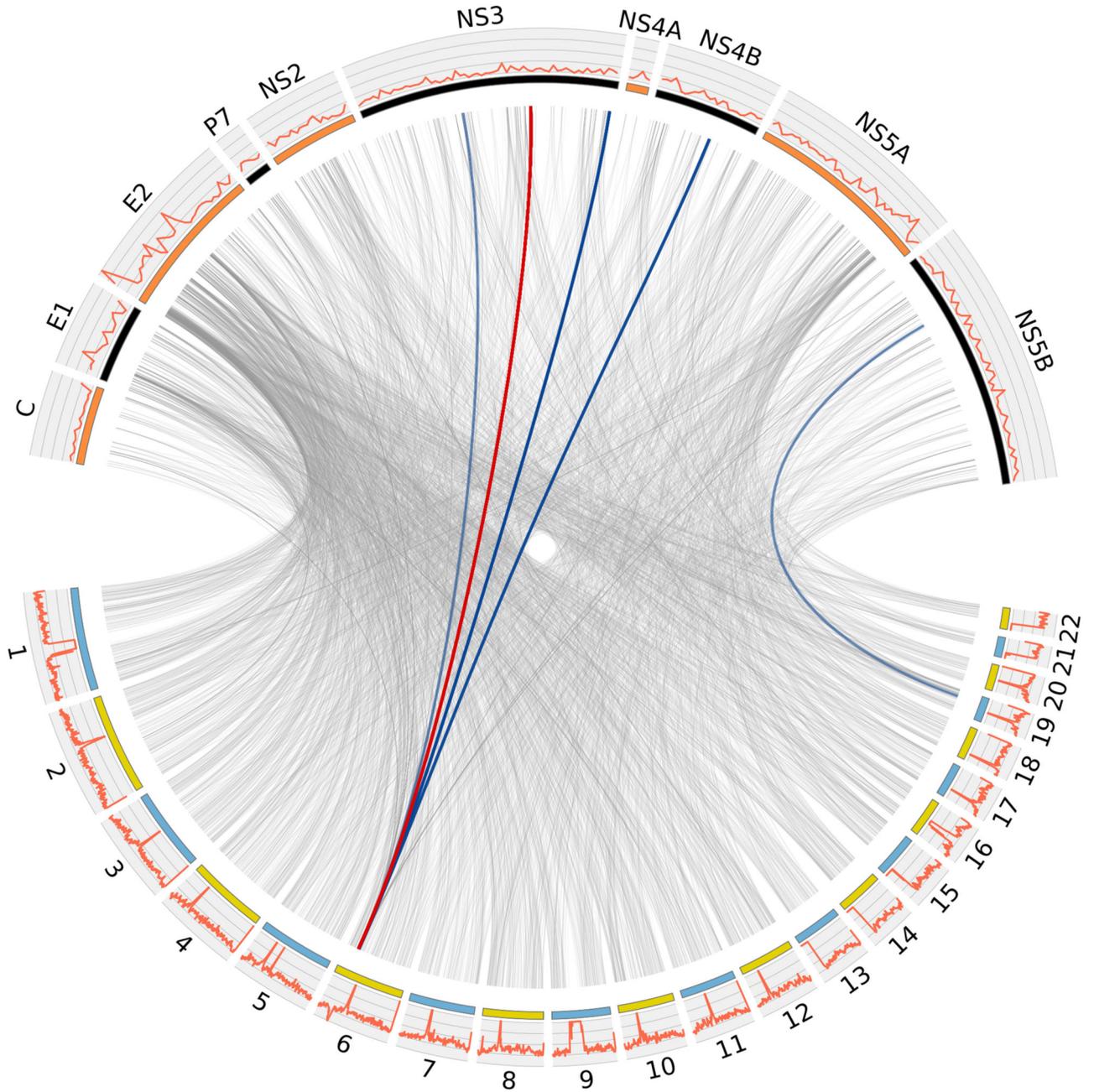
Figure 1. Human to hepatitis C virus genome-wide association study in 542 patients.
The lower arc shows the human autosomes from chromosome 1 to 22, and the upper arc
shows the HCV proteome from core (C) to NS5B. The red line represents the most
significant association, $P < 2 \times 10^{-11}$. The four blue 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.
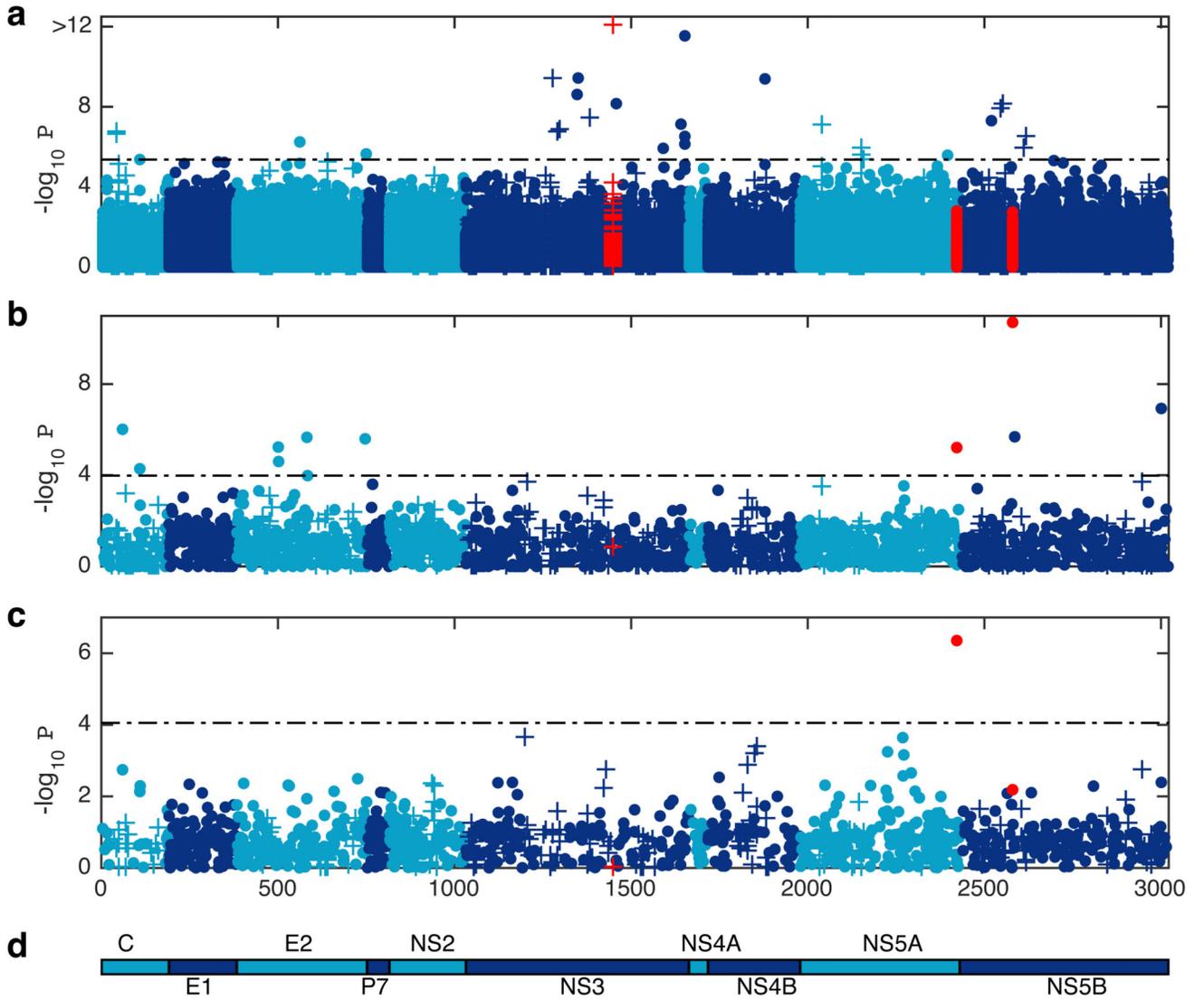
Figure 2. Hepatitis C virus genome-wide association studies.
Association between HCV amino acids and (a) HLA alleles (Fisher's exact test) (b) *IFNL4* genotypes (Fisher's exact test) and (c) pre-treatment viral load ($\log_{10}$PTVL, linear regression). Sites in experimentally validated epitopes in HCV genotype 3 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.
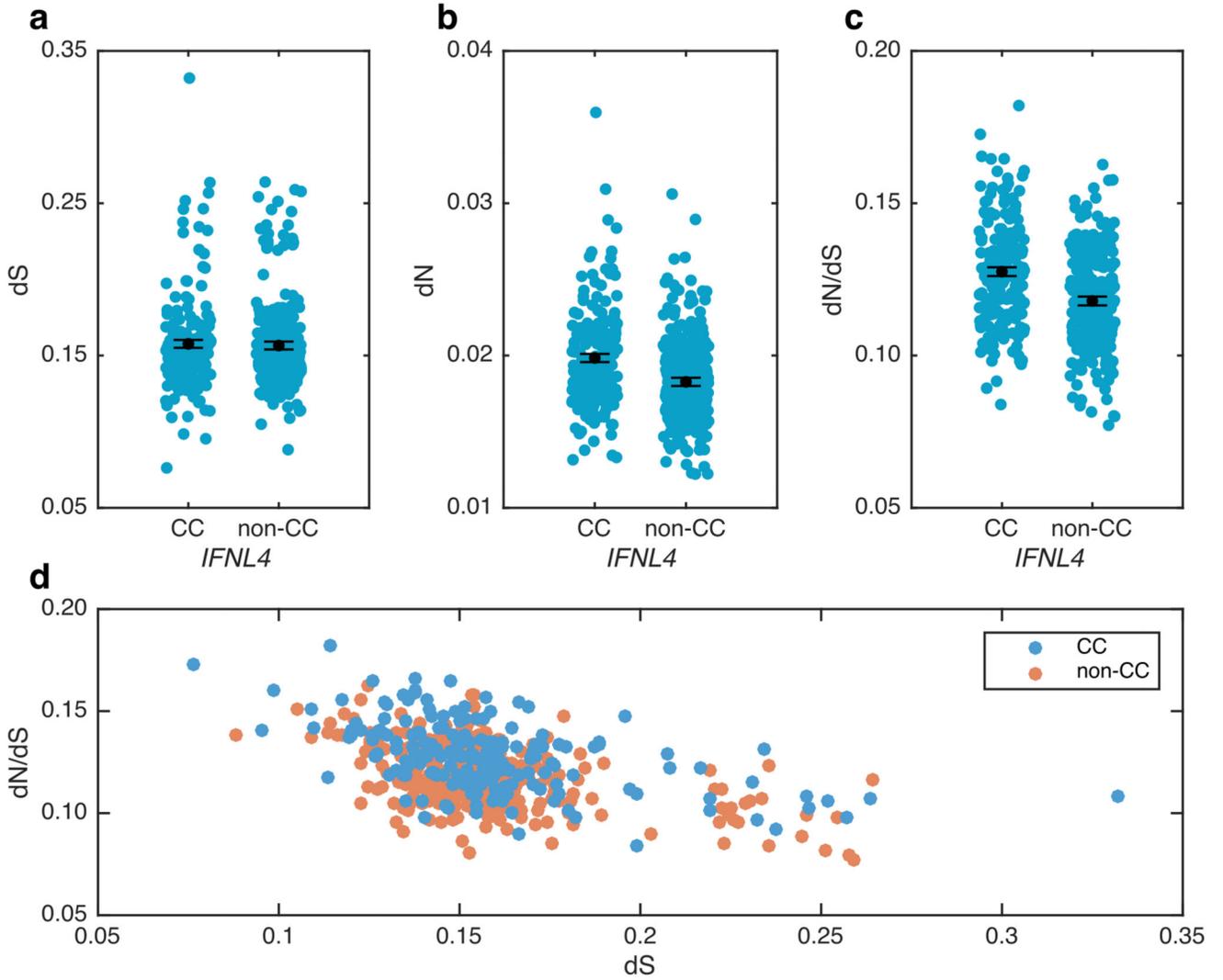
Figure 3. Association between *IFNL4* genotypes and substitution rates in genotype 3a infected patients.

Stratification of substitution rates by *IFNL4* CC and non-CC genotypes. (a) Rate of synonymous substitutions (dS, P=0.68), (b) Rate of non-synonymous substitutions (dN, P=1.6x10^{-8}) and (c) dN/dS ratio (P=1.28x10^{-10}). 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. (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).
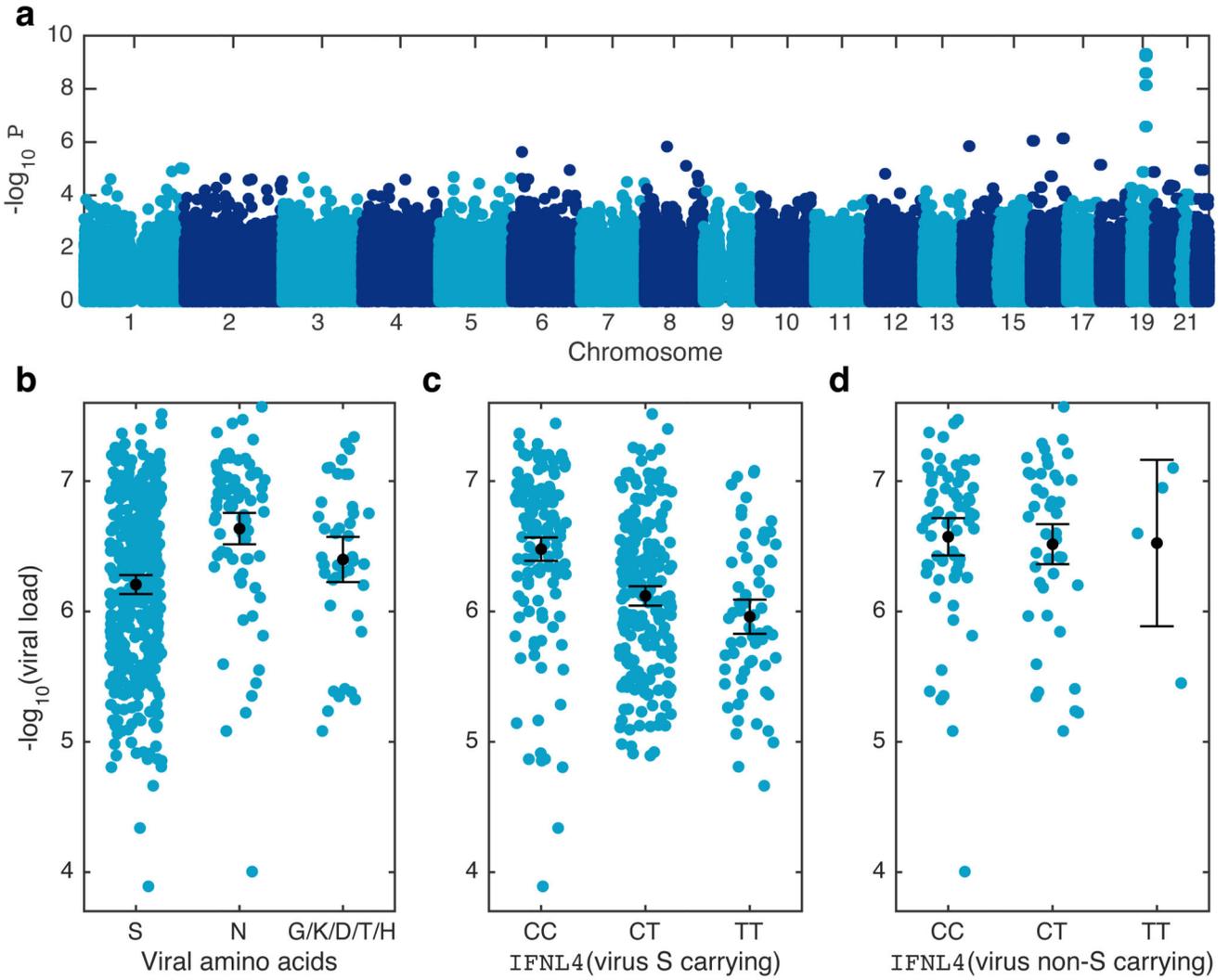
Figure 4. Association between viral load and human and virus genetic variants in genotype 3a infected patients.

(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=9.21×10^{-7}), (c) by *IFNL4* (CC, CT and TT) genotypes in patients whose virus carries a serine at position 2414 (P=9.37×10^{-9}) or (d) in patients whose virus does not carry a serine amino acid at position 2414 (P=0.9).
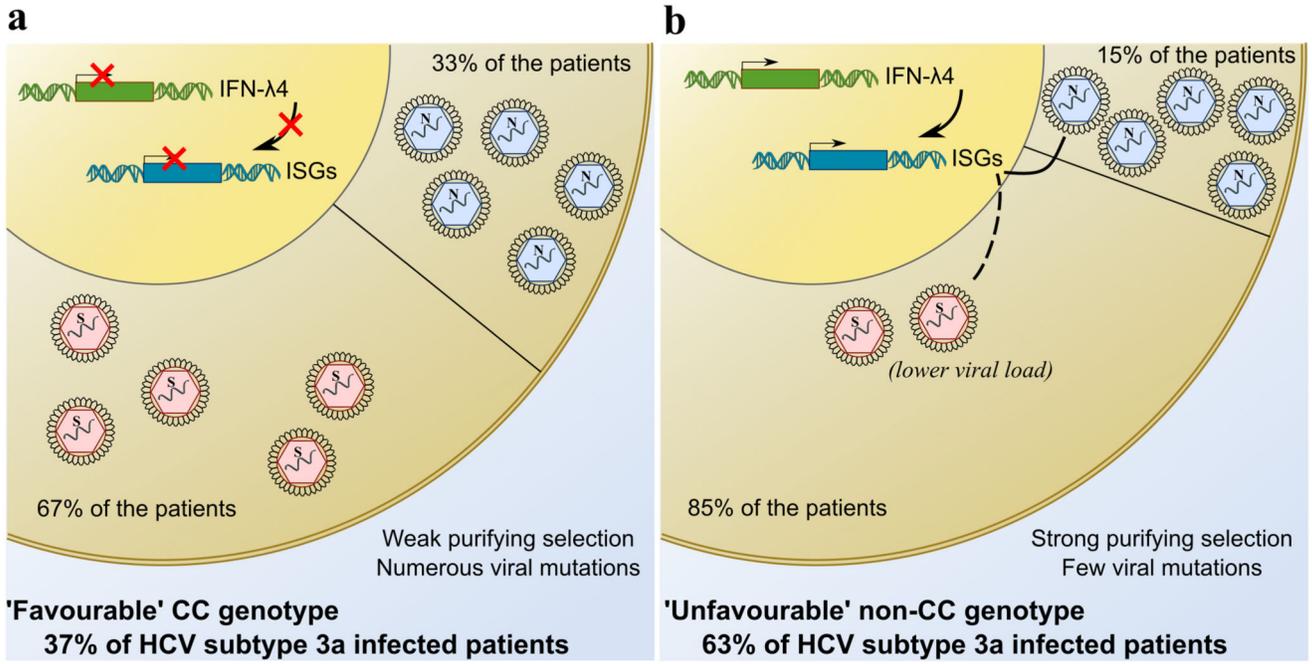
Figure 5. Overview of the observations relating to the interplay between innate immune response and the viral genome in hepatitis C virus control.
(a) Infected individuals with *IFNL4* CC genotype (37% of the HCV genotype 3a infected patients) show high rates of spontaneous and treatment-induced clearance 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 (85% are infected by a virus with a serine at position 2414 and 15% with a different amino-acid).

Table 1

Associations between HLA alleles and viral amino acids at a 5% false discovery rate.

In the whole cohort (All) or in HCV genotype 3a self-reported "White" patients only (G3a White). Significant associations are in bold. For each combination of HLA allele and viral site, only the most significant associated amino acid is reported. Amino acids are ordered by decreasing frequency in the column "variable amino acids at this site".

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	**1.55×10^{-07}**	**5.00×10^{-03}**	**8.07×10^{-07}**	**5.81×10^{-03}**	Yes
C*07:02	42	C	PL	P	**2.27×10^{-07}**	**6.10×10^{-03}**	**1.99×10^{-07}**	**1.33×10^{-03}**	Yes
B*41:02	109	C	PQSAL	S	4.29×10^{-06}	8.09×10^{-02}	**9.84×10^{-06}**	**4.93×10^{-02}**	No
B*13:02	348	E1	ILVM	I	8.13×10^{-06}	1.04×10^{-01}	**1.25×10^{-06}**	**8.78×10^{-03}**	No
C*08:02	372	E1	ATVIFC	I	7.33×10^{-04}	4.12×10^{-01}	**9.41×10^{-06}**	**4.93×10^{-02}**	No
A*29:02	444	E2	YHFRVWL	Y	1.85×10^{-04}	2.89×10^{-01}	**4.02×10^{-06}**	**2.39×10^{-02}**	No
C*15:02	561	E2	VTLI	L	**5.70×10^{-07}**	**1.50×10^{-02}**	6.80×10^{-04}	5.92×10^{-01}	No
DRB1*14:04	905	NS2	ATSC	S	4.45×10^{-03}	5.19×10^{-01}	**5.58×10^{-07}**	**4.20×10^{-03}**	No
A*31:01	1270	NS3	RKHS	R	**1.95×10^{-09}**	**4.44×10^{-04}**	**6.20×10^{-08}**	**3.64×10^{-04}**	Yes
A*33:03	1282	NS3	NVTSA	T	**1.73×10^{-07}**	**5.26×10^{-03}**	4.14×10^{-02}	9.13×10^{-01}	Yes
B*15:01	1290	NS3	KPARS	R	**1.25×10^{-07}**	**4.47×10^{-03}**	**3.64×10^{-06}**	**2.39×10^{-02}**	Yes
A*68:02	1341	NS3	VA	V	**2.43×10^{-09}**	**4.44×10^{-04}**	**9.93×10^{-07}**	**6.91×10^{-03}**	No
A*68:02	1344	NS3	TVA	T	**3.68×10^{-10}**	**$<4.00 \times 10^{-04}$**	**9.22×10^{-11}**	**$<2.50 \times 10^{-04}$**	No
B*51:01	1380	NS3	ILV	L	**3.22×10^{-08}**	**1.67×10^{-03}**	**2.10×10^{-07}**	**1.38×10^{-03}**	Yes
A*01:01	1444	NS3	FY	F	**9.63×10^{-33}**	**$<4.00 \times 10^{-04}$**	**7.18×10^{-24}**	**$<2.50 \times 10^{-04}$**	Yes
A*68:02	1452	NS3	IV	I	**6.98×10^{-09}**	**4.44×10^{-04}**	**3.84×10^{-07}**	**2.89×10^{-03}**	No
A*30:02	1585	NS3	YF	Y	**1.19×10^{-06}**	**3.28×10^{-02}**	6.89×10^{-05}	1.90×10^{-01}	No
B*13:02	1635	NS3	ITVLAF	T	**7.38×10^{-08}**	**3.13×10^{-03}**	**1.65×10^{-07}**	**1.00×10^{-03}**	No
B*08:01	1646	NS3	MTAVSI	T	**2.89×10^{-12}**	**$<4.00 \times 10^{-04}$**	**2.01×10^{-10}**	**$<2.50 \times 10^{-04}$**	No
C*07:01	1646	NS3	MTAVSI	T	**3.01×10^{-07}**	**7.39×10^{-03}**	**1.40×10^{-08}**	**$<2.50 \times 10^{-04}$**	No
DQA1*05:01	1646	NS3	MTAVSI	T	**7.24×10^{-07}**	**1.76×10^{-02}**	1.18×10^{-05}	5.64×10^{-02}	No
B*57:01	1759	NS4B	AVTGNSDI	A	3.38×10^{-05}	1.83×10^{-01}	**4.01×10^{-06}**	**2.39×10^{-02}**	No
A*02:01	1873	NS4B	LFKICVPRMTA	F	**4.06×10^{-10}**	**$<4.00 \times 10^{-04}$**	**2.58×10^{-08}**	**2.50×10^{-04}**	No
B*38:01	2034	NS5A	STNLPIAQVDKEMGRFW	P	**7.56×10^{-08}**	**3.13×10^{-03}**	**6.35×10^{-08}**	**3.64×10^{-04}**	Yes
C*12:03	2034	NS5A	STLPQKVAIRMFW	P	9.31×10^{-06}	1.05×10^{-01}	**3.99×10^{-06}**	**2.39×10^{-02}**	Yes
B*18:01	2144	NS5A	ED	E	**1.16×10^{-06}**	**3.28×10^{-02}**	**2.36×10^{-06}**	**1.67×10^{-02}**	Yes
B*51:01	2148	NS5A	MTVSLKIA	V	2.77×10^{-06}	6.07×10^{-02}	**4.60×10^{-07}**	**3.16×10^{-03}**	Yes
C*14:02	2148	NS5A	MTVSLKIA	V	1.91×10^{-05}	1.34×10^{-01}	**2.32×10^{-06}**	**1.67×10^{-02}**	Yes
B*40:01	2248	NS5A	TSKAN	T	5.76×10^{-04}	3.91×10^{-01}	**9.62×10^{-06}**	**4.93×10^{-02}**	No

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?
DQA1*01:01	2486	NS5B	TIANSV	T	4.71×10^{-05}	1.96×10^{-01}	**5.12×10^{-06}**	**2.90×10^{-02}**	Yes
A*31:01	2510	NS5B	AKQSEMGLTV	A	**4.96×10^{-08}**	**2.14×10^{-03}**	**1.79×10^{-06}**	**1.35×10^{-02}**	No
A*32:01	2537	NS5B	NDHSYEAI	N	**1.13×10^{-08}**	**7.27×10^{-04}**	**4.77×10^{-06}**	**2.75×10^{-02}**	Yes
A*32:01	2540	NS5B	RSKHNQC	R	**7.45×10^{-09}**	**6.00×10^{-04}**	**5.96×10^{-12}**	**$<2.50 \times 10^{-04}$**	Yes
A*02:11	2600	NS5B	QKRSAL	Q	**1.01×10^{-06}**	**3.12×10^{-02}**	NA	NA	Yes
A*26:01	2605	NS5B	EAGVK	G	**2.61×10^{-07}**	**6.64×10^{-03}**	5.91×10^{-05}	1.73×10^{-01}	Yes
B*51:01	2713	NS5B	ILMV	I	7.70×10^{-06}	1.03×10^{-01}	**1.25×10^{-08}**	**$<2.50 \times 10^{-04}$**	No
A*25:01	2821	NS5B	RCQL	R	8.31×10^{-06}	1.04×10^{-01}	**2.51×10^{-06}**	**1.73×10^{-02}**	No