nature genetics

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

M Azim Ansari[1–3,11], Vincent Pedergnana[1,11], Camilla L C Ip[1,3], Andrea Magri[3], 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[9], Emma Hudson[3], John McLauchlan[10], Peter Simmonds[3], Rory Bowden[1], Paul Klenerman[3], Eleanor Barnes[3], Chris C A Spencer[1]

Outcomes of hepatitis C virus (HCV) infection and treatment depend on viral and host genetic factors. Here 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 who were chronically infected with HCV, predominantly genotype 3. We show that both alleles of genes encoding human leukocyte antigen molecules and genes encoding components of the interferon lambda innate immune system drive viral polymorphism. Additionally, we show that *IFNL4* genotypes determine HCV viral load through a mechanism dependent on a specific amino acid residue in the HCV NS5A protein. These findings highlight the interplay between the innate immune system and the viral genome in HCV control.

HCV infection presents a major health burden, with more than 185 million people being infected worldwide[1], which can lead to liver failure and hepatocellular cancer in infected individuals. Genetic variations in both the host and the virus are associated with important clinical outcomes. Genetic polymorphisms in the host, most notably in the interferon (IFN) lambda 3 (*IFNL3*) and *IFNL4* loci, are associated with spontaneous clearance of the virus, response to treatment, viral load and progression of liver disease[2–6]. Viral genotypes and distinct viral genetic motifs have been associated with the response to interferon-based therapies[7,8], whereas resistance-associated substitutions (RASs) have been identified for most of the new oral direct-acting antiviral (DAA) drugs[9–12]. HCV can be divided into seven major genotypes, and most of the genetic data acquired to date has focused on HCV genotype 1, with a lack of data for other genotypes. HCV genotype 3 is of particular interest, as this genotype is known to infect 53 million people globally[13] and is associated with a higher failure rate to DAA therapies[14,15].

Previous work, including candidate gene studies of the association between the human leukocyte antigen (HLA) type I proteins and the HCV genome[16,17], has shown that within-host virus diversity evolves in response to the host adaptive immune system. HLA molecules are expressed on most cell types, and they present viral peptides (epitopes) to cytotoxic T lymphocytes (CTLs), which kill infected cells. CTL-mediated killing of virus-infected cells drives the selection of viral polymorphisms ('escape' mutations) that abrogate T cell recognition[18]. Understanding how host HLA molecules affect viral selection has important implications for the development of HCV-specific T cell vaccines that aim to prevent infection[19,20]. A comprehensive host genome to viral genome analysis at scale will assess the relative contribution of host HLA molecules in driving changes in the HCV genome, and it might also identify other host genes that have a key role in shaping the HCV genome.

We generated data from a cohort of 601 HCV-infected patients (from the BOSON[21] clinical trial) to systematically look for associations between host and virus genomes, exploiting the fact that while the host genome remains fixed the virus mutates, allowing it to evolve during infection. For this, we developed a targeted viral enrichment methodology[22,23] to obtain whole HCV genomes, and we used high-throughput genotyping arrays in combination with statistical imputation to obtain data for nucleotide polymorphisms across the human genome and the alleles of genes encoding HLA molecules[24,25] (hereafter referred to as HLA genes). 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 an amino acid residue in the HCV NS5A protein determines HCV viral load. By assessing viral evolution in individuals with different *IFNL4*

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Oxford Martin School, University of Oxford, Oxford, UK. [3]Peter Medawar Building for Pathogen Research, Nuffield Department of Medicine and the NIHR Oxford BRC, University of Oxford, Oxford, UK. [4]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [5]Department of Statistics, University of Oxford, Oxford, UK. [6]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [7]Wright–Fleming Institute, Imperial College London, London, UK. [8]Queen Mary University of London, London, UK. [9]A full list of members and affiliations appears at the end of the paper. [10]MRC–University of Glasgow Centre for Virus Research, Glasgow, UK. [11]These authors contributed equally to this work. Correspondence should be addressed to C.C.A.S. (chris.spencer@well.ox.ac.uk) or E.B. (ellie.barnes@ndm.ox.ac.uk).

Received 9 August 2016; accepted 10 March 2017; published online 10 April 2017; doi:10.1038/ng.3835

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, as well as other viral infections, in the era of genomic analysis.

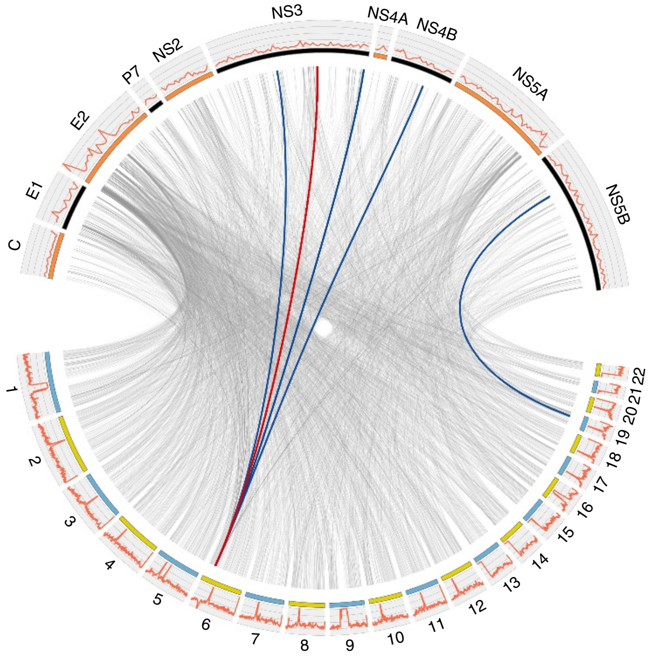

Figure 1 Human-to-HCV genome-wide association study in 542 patients. The lower arc shows the human autosomes from chromosomes 1 to 22, and the upper arc shows the HCV proteome from the core protein (C) to NS5B. The red line represents the most significant association ($P < 2 \times 10^{-11}$). The four blue lines represent suggestive associations ($P < 4 \times 10^{-9}$). The thin gray 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 1 Mb, with higher values further away from the center for both the upper and lower arcs.

RESULTS

Sample description and genetic structure

DNA samples from 567 patients (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. Pretreatment plasma samples from 583 patients in the same study were analyzed to obtain consensus HCV whole-genome sequences using a high-throughput HCV-targeted sequence-capture approach coupled with Illumina sequencing[22].

Both full-length HCV genome sequences and human genome-wide SNP data were obtained on a total of 542 patients of mainly self-reported white and Asian ancestry who were infected with HCV genotypes 2 or 3 (**Supplementary Table 1**). After quality control analysis and filtering of the human genotype data, approximately 330,000 common SNPs with a minor allele frequency >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, which corresponds to >3,000 encoded amino acids. In our data set, we defined 1,226 sites of the HCV proteome to be variable (where at least ten isolates had an amino acid that differed from the consensus amino acid, giving adequate statistical power for analysis).

We characterized human genetic diversity in the cohort via principal component analysis (PCA). The first two principal components (PCs) corresponded to the sample's (self-reported) 83% white and 14% Asian ethnicity (**Supplementary Fig. 1**), which differed significantly ($P < 0.05$) in some of the inferred HLA allele frequencies (**Supplementary Table 2**), consistent with previous observations[26]. The third PC separates individuals of self-reported black ethnicity from the rest of the cohort. We summarized virus diversity by constructing a maximum-likelihood tree of the consensus sequences from each patient (**Supplementary Fig. 2**). Major clades in the tree separated HCV genotypes 2 (8% of the total number of samples) and 3 (which in turn comprised clades representing samples with subtype 3a (90% of the samples) and non-subtype 3a (2% of the samples)). We observed that patients of specific ancestries, as measured either by genetic ancestry (PCs) or self-reported ethnicity, clustered together on the tree of viral diversity (**Supplementary Fig. 2**). A PCA of virus nucleotide sequence data reflected the structure of the tree, specifically at the level of virus subtypes (**Supplementary Fig. 3**).

Systematic host genome to virus genome analysis

We used the genotyped autosomal SNPs in the host genome to undertake genome-wide association studies, for which the traits of interest were the presence or absence of each amino acid at the variable sites of the virus proteome, which resulted 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 of the covariates led to a substantial inflation in the association test statistics (**Supplementary Fig. 4**), as would be expected given the observed correlation in population structure of the virus and the host (**Supplementary Fig. 2**). We assumed that there was a human genome-wide significance threshold[27] of 5×10^{-8} and that amino acid variants in the viral genome were approximately uncorrelated once the population structure was accounted for, and we found that a Bonferroni correction[28] resulted 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) locus (on chromosome 6) and a virus amino acid variant in non-structural protein 3 (NS3) (**Fig. 1**). Three other associations were observed between multiple SNPs in the host MHC locus and the virus amino acids in the NS3 and NS4B proteins ($P < 4 \times 10^{-9}$; **Fig. 1**). Outside of the MHC locus, the strongest association between host and virus was detected between the SNP rs12979860 in the host *IFNL4* gene (on chromosome 19) and the amino acids at position 2,570 in the HCV NS5B protein ($P = 1.98 \times 10^{-9}$; **Fig. 1**). The observed variability in the density of nominally significant associations (**Fig. 1**) can largely be explained by variability in the host and virus sequences, for example, in the hypervariable region (HVR) of the HCV E2 protein.

We observed 182 associations between human SNPs that mapped to two loci and five HCV amino acid sites ($P < 4 \times 10^{-9}$) (**Supplementary Table 3**). Because these associations represent places where host genetic diversity has had an effect on virus sequence diversity, we refer to them as 'footprints'. We interpreted the signals of association in the MHC region to indicate the effect of the adaptive immune system on genetic diversity in the virus genome. Although the effect of MHC was anticipated, the strong signal of association of the interferon lambda region with viral diversity indicated that there were additional effects of the innate immune response.

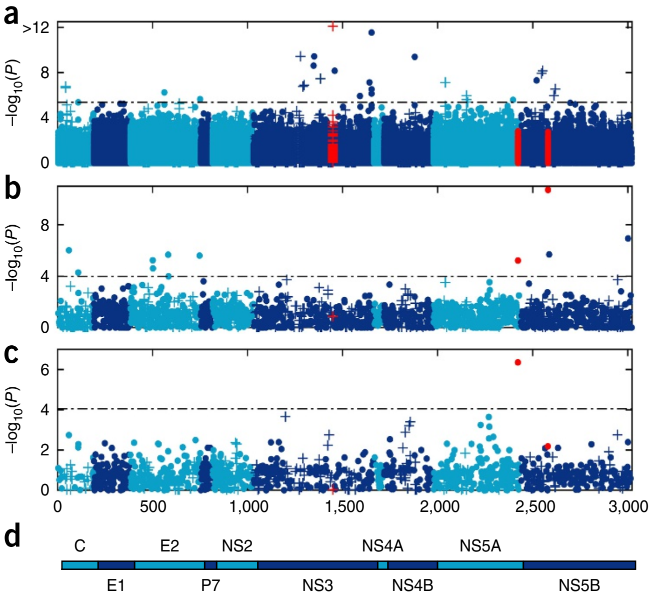

Figure 2 HCV genome-wide association studies. (**a**–**c**) Association between HCV amino acids (on *x* axis) and HLA alleles (by Fisher's exact test) (**a**), *IFNL4* genotypes (by Fisher's exact test) (**b**) and pre-treatment viral load ($\log_{10}$(PTVL); by linear regression) (**c**). Sites in experimentally validated epitopes in HCV genotype 3 are indicated by a plus sign. Viral sites at positions 1,444, 2,414 and 2,570 are colored red. Dashed lines represent a 5% FDR. (**d**) Schematic of the HCV polyprotein.

Host HLA alleles to virus genome

SNPs in the MHC-encoding region that show strong association with viral amino acids are likely to be correlated with alleles at the 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 this mechanism can lead to the selection of viral 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 (**Fig. 2a** and **Table 1**), we inferred the changes on the terminal branches of the virus phylogenetic tree for each amino acid site and assessed their association with the hosts' HLA alleles. We repeated the analysis on patients who were infected with HCV genotype 3a and who self-reported as white (**Table 1** and **Supplementary Fig. 5**).

In the whole cohort, at a false discovery rate (FDR) of 5%, 24 combinations of HLA alleles and HCV sites were significant (**Table 1**), and this increased to 153 associations at a FDR of 20% (**Supplementary Table 4**). Of 21 viral amino acid positions that showed 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; $P = 2.8 \times 10^{-4}$). 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 footprints were found with common alleles at *HLA-A* and *HLA-B*, although signals were also found in *HLA-C* and the class II gene *HLA-DQA1* (see **Supplementary Note** and **Supplementary Figs. 6–8** for a detailed description of the most significant association between *HLA-A* and Tyr1,444Phe). At amino acid position 1,646 in the HCV polyprotein (in the NS3 protein), the footprint is seen with multiple HLA alleles (*HLA-B*08:01*, *HLA-C*07:01* and *HLA-DQA1*05:01*), although this is potentially a result of linkage disequilibrium between these HLA alleles (r^2 between *HLA-B*08:01* and *HLA-C*07:01* = 0.49, and r^2 between *HLA-B**08:01 and *HLA-DGA1*05:01* = 0.22; **Supplementary Fig. 9**). At a 20% FDR, using the 27 HLA and viral amino acid associations that have sufficient observations to estimate the ORs, we observed a negative correlation between the ORs of escape and reversion amino acids ($r = -0.65$; $P < 0.01$; **Supplementary Fig. 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, outside of the MHC region, variants around the interferon lambda region showed the strongest association with HCV amino acids (the top associated SNP is rs12979860, $P = 1.98 \times 10^{-9}$). For SNP rs12979860, the CC genotype is associated with higher rates of spontaneous clearance and interferon-based treatment response, putatively owing to the fact that the C allele tags a dinucleotide insertion polymorphism (rs368234815-TT) that 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 at position 2,570 in the virus NS5B protein ($P = 1.94 \times 10^{-11}$; **Fig. 2b**). We replicated this association in an independent study of 360 patients of European ancestry who were chronically infected with HCV genotypes 2 or 3 (one-sided $P = 0.005$)[38,39]. 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 2,609)[32], reinforcing the potential role of the locus in interactions with the host innate immune system.

Overall, using a 5% FDR, we observed 11 significant associations between *IFNL4* genotypes and amino acid positions located in the HCV core, E2, NS5A and NS5B proteins (**Supplementary Table 5**). Using a permutation approach, we found that the core protein was nominally ($P < 0.05$) enriched, and the NS5A protein was 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 individuals with the non-CC genotype (**Supplementary Note** and **Supplementary Table 6**). One of the sites (position 109 in the core protein) was also associated with *HLA-B*41:02* ($P = 4.3 \times 10^{-6}$; **Table 1**). However, we did not find any consistent evidence for interaction between *IFNL4* genotypes, the HLA alleles and their associated viral amino acid sites (**Supplementary Fig. 11**) nor did we find strong differences in the mean number of escape mutations in patients with the CC or non-CC genotypes when comparing across all combinations of HLA alleles and viral sites associated at 20% FDR.

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 nonsynonymous (dN) substitutions in patients infected with HCV genotype 3a (**Fig. 3**). Although there was no

Table 1 Associations between HLA alleles and viral amino acids

HLA allele	HCV amino acid position	Viral protein	Variable amino acids at this site	Associated amino acid	*P* value Whole	*q* value Whole	*P* value G3a white	*q* value G3a white	In a known epitope?
B*07:02	42	C	PL	P	**1.55×10^{-7}**	**5.00×10^{-3}**	**8.07×10^{-7}**	**5.81×10^{-3}**	Yes
C*07:02	42	C	PL	P	**2.27×10^{-7}**	**6.10×10^{-3}**	**1.99×10^{-7}**	**1.33×10^{-3}**	Yes
B*41:02	109	C	PQSAL	S	4.29×10^{-6}	8.09×10^{-2}	**9.84×10^{-6}**	**4.93×10^{-2}**	No
B*13:02	348	E1	ILVM	I	8.13×10^{-6}	1.04×10^{-1}	**1.25×10^{-6}**	**8.78×10^{-3}**	No
C*08:02	372	E1	ATVIFC	I	7.33×10^{-4}	4.12×10^{-1}	**9.41×10^{-6}**	**4.93×10^{-2}**	No
A*29:02	444	E2	YHFRVWL	Y	1.85×10^{-4}	2.89×10^{-1}	**4.02×10^{-6}**	**2.39×10^{-2}**	No
C*15:02	561	E2	VTLI	L	**5.70×10^{-7}**	**1.50×10^{-2}**	6.80×10^{-4}	5.92×10^{-1}	No
DRB1*14:04	905	NS2	ATSC	S	4.45×10^{-3}	5.19×10^{-1}	**5.58×10^{-7}**	**4.20×10^{-3}**	No
A*31:01	1,270	NS3	RKHS	R	**1.95×10^{-9}**	**4.44×10^{-4}**	**6.20×10^{-8}**	**3.64×10^{-4}**	Yes
A*33:03	1,282	NS3	NVTSA	T	**1.73×10^{-7}**	**5.26×10^{-3}**	4.14×10^{-2}	9.13×10^{-1}	Yes
B*15:01	1,290	NS3	KPARS	R	**1.25×10^{-7}**	**4.47×10^{-3}**	**3.64×10^{-6}**	**2.39×10^{-2}**	Yes
A*68:02	1,341	NS3	VA	V	**2.43×10^{-9}**	**4.44×10^{-4}**	**9.93×10^{-7}**	**6.91×10^{-3}**	No
A*68:02	1,344	NS3	TVA	T	**3.68×10^{-10}**	**$<4.00 \times 10^{-4}$**	**9.22×10^{-11}**	**$<2.50 \times 10^{-4}$**	No
B*51:01	1,380	NS3	ILV	L	**3.22×10^{-8}**	**1.67×10^{-3}**	**2.10×10^{-7}**	**1.38×10^{-3}**	Yes
A*01:01	1,444	NS3	FY	F	**9.63×10^{-33}**	**$<4.00 \times 10^{-4}$**	**7.18×10^{-24}**	**$<2.50 \times 10^{-4}$**	Yes
A*68:02	1,452	NS3	IV	I	**6.98×10^{-9}**	**4.44×10^{-4}**	**3.84×10^{-7}**	**2.89×10^{-3}**	No
A*30:02	1,585	NS3	YF	Y	**1.19×10^{-6}**	**3.28×10^{-2}**	6.89×10^{-5}	1.90×10^{-1}	No
B*13:02	1,635	NS3	ITVLAF	T	**7.38×10^{-8}**	**3.13×10^{-3}**	**1.65×10^{-7}**	**1.00×10^{-3}**	No
B*08:01	1,646	NS3	MTAVSI	T	**2.89×10^{-12}**	**$<4.00 \times 10^{-4}$**	**2.01×10^{-10}**	**$<2.50 \times 10^{-4}$**	No
C*07:01	1,646	NS3	MTAVSI	T	**3.01×10^{-7}**	**7.39×10^{-3}**	**1.40×10^{-8}**	**$<2.50 \times 10^{-4}$**	No
DQA1*05:01	1,646	NS3	MTAVSI	T	**7.24×10^{-7}**	**1.76×10^{-2}**	1.18×10^{-5}	5.64×10^{-2}	No
B*57:01	1,759	NS4B	AVTGNSDI	A	3.38×10^{-5}	1.83×10^{-1}	**4.01×10^{-6}**	**2.39×10^{-2}**	No
A*02:01	1,873	NS4B	LFKICVPRMTA	F	**4.06×10^{-10}**	**$<4.00 \times 10^{-4}$**	**2.58×10^{-8}**	**2.50×10^{-4}**	No
B*38:01	2,034	NS5A	STNLPIAQVD KEMGRFW	P	**7.56×10^{-8}**	**3.13×10^{-3}**	**6.35×10^{-8}**	**3.64×10^{-4}**	Yes
C*12:03	2,034	NS5A	STLPQKVAIRMFW	P	9.31×10^{-6}	1.05×10^{-1}	**3.99×10^{-6}**	**2.39×10^{-2}**	Yes
B*18:01	2,144	NS5A	ED	E	**1.16×10^{-6}**	**3.28×10^{-2}**	**2.36×10^{-6}**	**1.67×10^{-2}**	Yes
B*51:01	2,148	NS5A	MTVSLKIA	V	2.77×10^{-6}	6.07×10^{-2}	**4.60×10^{-7}**	**3.16×10^{-3}**	Yes
C*14:02	2,148	NS5A	MTVSLKIA	V	1.91×10^{-5}	1.34×10^{-1}	**2.32×10^{-6}**	**1.67×10^{-2}**	Yes
B*40:01	2,248	NS5A	TSKAN	T	5.76×10^{-4}	3.91×10^{-1}	**9.62×10^{-6}**	**4.93×10^{-2}**	No
DQA1*01:01	2,486	NS5B	TIANSV	T	4.71×10^{-5}	1.96×10^{-1}	**5.12×10^{-6}**	**2.90×10^{-2}**	Yes
A*31:01	2,510	NS5B	AKQSEMGLTV	A	**4.96×10^{-8}**	**2.14×10^{-3}**	**1.79×10^{-6}**	**1.35×10^{-2}**	No
A*32:01	2,537	NS5B	NDHSYEAI	N	**1.13×10^{-8}**	**7.27×10^{-4}**	**4.77×10^{-6}**	**2.75×10^{-2}**	Yes
A*32:01	2,540	NS5B	RSKHNQC	R	**7.45×10^{-9}**	**6.00×10^{-4}**	**5.96×10^{-12}**	**$<2.50 \times 10^{-4}$**	Yes
A*02:11	2,600	NS5B	QKRSAL	Q	**1.01×10^{-6}**	**3.12×10^{-2}**	NA	NA	Yes
A*26:01	2,605	NS5B	EAGVK	G	**2.61×10^{-7}**	**6.64×10^{-3}**	5.91×10^{-5}	1.73×10^{-1}	Yes
B*51:01	2,713	NS5B	ILMV	I	7.70×10^{-6}	1.03×10^{-1}	**1.25×10^{-8}**	**$<2.50 \times 10^{-4}$**	No
A*25:01	2,821	NS5B	RCQL	R	8.31×10^{-6}	1.04×10^{-1}	**2.51×10^{-6}**	**1.73×10^{-2}**	No

Significant associations at a 5% FDR (shown in bold) in the whole cohort (Whole) or in the self-reported white patients infected with HCV genotype 3a (G3a white) are shown. 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".

difference in the dS values in patients with the CC or non-CC genotype ($P = 0.68$; **Fig. 3a**), the dN value was significantly higher in patients with the CC genotype ($P = 1.6 \times 10^{-8}$; **Fig. 3b**). The lower dN/dS ratio in patients with the non-CC genotype ($P = 1.3 \times 10^{-10}$) is potentially indicative of the idea that the virus is under a stronger purifying selection than in patients with the CC genotype (**Fig. 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 in patients with the CC genotype (**Fig. 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 Fig. 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 Fig. 13**).

Host and virus genetic determinants of viral load

We performed a genome-wide association study (GWAS) in patients who were infected with HCV genotype 3a, using an additive linear regression model adjusted for sex and the three first host PCs for $\log_{10}$-transformed pre-treatment viral load ($\log_{10}$(PTVL)) (**Fig. 4a**). We replicated the known association between *IFNL4* variants on chromosome 19 and viral load[3] (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 GWAS to detect associations between amino acids of HCV genotype 3a and viral load (**Fig. 2c**). The only amino acid that was significantly associated with $\log_{10}$(PTVL) at a 5% FDR was residue 2,414 (in the NS5A protein), which showed a change from a serine to an asparagine

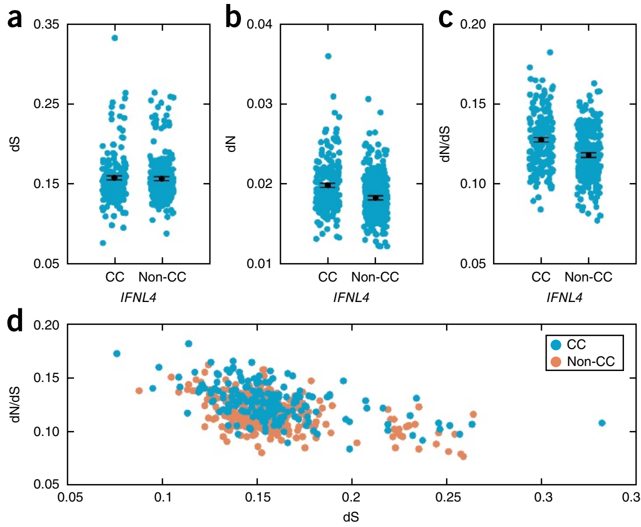

Figure 3 Association between *IFNL4* genotypes and nucleotide substitution rates in patients infected with HCV genotype 3a. (**a**) Rate of synonymous substitutions (dS, $P = 0.68$; by linear regression). (**b**) Rate of nonsynonymous substitutions (dN, $P = 1.6 \times 10^{-8}$; by linear regression). (**c**) The dN/dS ratio ($P = 1.28 \times 10^{-10}$; by linear regression). In **a**–**c**, the mean and 95% confidence intervals are shown as black dots and bars, respectively. (**d**) The joint distribution of dS and dN/dS values in individuals with either the *IFNL4* non-CC genotypes (red dots) or the *IFNL4* CC genotype (blue dots).

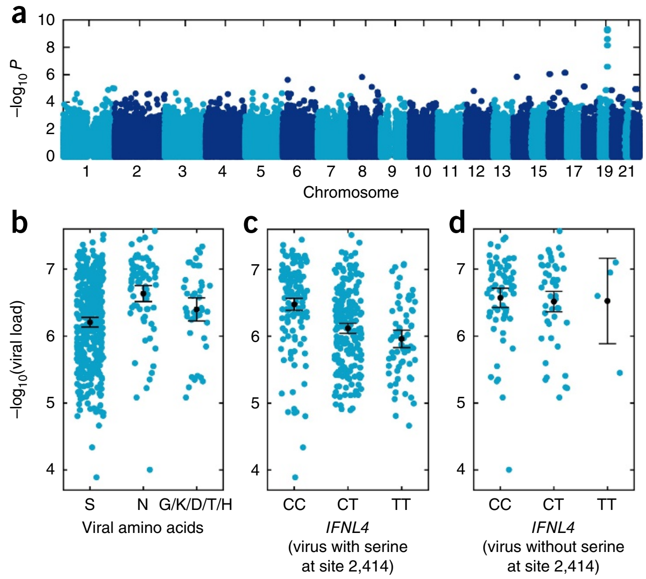

Figure 4 Association between viral load and the human and virus genetic variants in patients infected with HCV genotype 3a. (**a**) Association between human SNPs and $\log_{10}$(PTVL). (**b**–**d**) Distribution (blue dots), estimated mean and 95% confidence interval (shown as black dots and bars, respectively) of $\log_{10}$(PTVL) stratified by the amino acids present at position 2,414 in HCV ($P = 9.21 \times 10^{-7}$; by linear regression) (**b**), by *IFNL4* genotypes (CC, CT and TT) in patients whose virus carries a serine at position 2,414 ($P = 9.37 \times 10^{-9}$; by linear regression) (**c**) or in patients whose virus does not carry a serine at position 2,414 ($P = 0.9$; by linear regression) (**d**).

($P = 9.21 \times 10^{-7}$; **Fig. 4b**). This site is one of the 11 sites that were significantly associated (5% FDR) with *IFNL4* genotypes (**Fig. 2b**,**c**).

In patients infected with HCV that had a serine at position 2,414, the association between *IFNL4* genotypes and $\log_{10}$(PTVL) was significant ($P = 9.37 \times 10^{-9}$; **Fig. 4c**). However, we observed no association ($P = 0.9$) between *IFNL4* genotypes and $\log_{10}$(PTVL) in patients who were infected with a virus that had a different amino acid at that position (**Fig. 4d**). In other words, the host's *IFNL4* genotypes determined viral load only if the individuals were infected by a virus with a serine at position 2,414 in the NS5A protein (**Fig. 4c**,**d**). The interaction was statistically significant when analyzing either the whole cohort ($P = 0.017$) or just patients who had been infected with HCV genotype 3a and who self-reported as being white ($P = 0.017$). Taken together, the combinations of patients with non-CC genotypes and HCV with serine at site 2,414 are inferred to result in a 0.57-fold decrease in viral load as compared to that for all other combinations (mean viral load for non-CC and serine at site 2,414 = 2.81×10^6 IU/ml, and mean viral load for all other combinations = 6.47×10^6 IU/ml). Introduction of the change from a serine to an asparagine at position 2,414 into a modified S52 replicon of HCV genotype 3a (ref. 40) resulted in an approximately tenfold increase in replication of the replicon in Huh7.5 cell culture (**Supplementary Note** and **Supplementary Fig. 14**).

We also observed that non-consensus amino acids, which are increased in frequency in individuals with a CC genotype, tended also to associate with increased viral load ($r = 0.42$; $P = 0.005$; **Supplementary Fig. 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 with non-CC genotypes only ($r = 0.4$; $P = 0.007$). A nominally significant trend was observed when the analysis was repeated for all of the variant positions in the viral genome ($r = 0.099$; $P = 0.04$) (**Supplementary Fig. 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 numbers of patients for a reasonable cost (~£100 per sample), as well as host genetic data at millions of directly assayed and imputed polymorphisms (~£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 PCs of the genome-wide data as covariates (60 h for ~2,500 association studies of 330,000 SNPs). We also applied a contingency table analysis based on the inferred amino acid changes in the virus after 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 for 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 previously unknown T cell escape mutations. Assuming that our analysis removes biases associated with population structure and incorrect ancestral inference, 5% of the viral amino acids (153/3,021) are associated with HLA alleles (at 20% FDR). These data highlight the importance of the adaptive immune system in

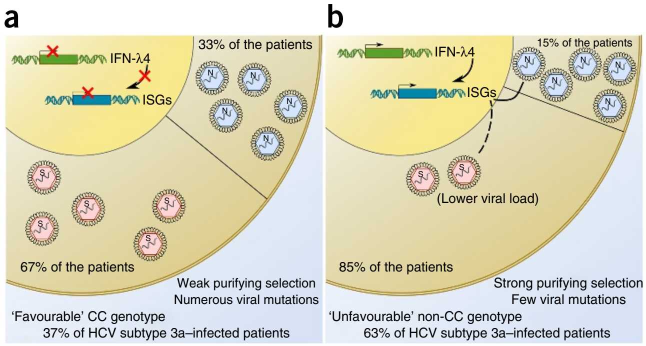

Figure 5 Overview of the observations relating to the interplay between the host innate immune response and the viral genome in HCV control. (**a**) Infected individuals who have the *IFNL4* CC genotype (37% of the HCV genotype 3a–infected patients) show high rates of spontaneous and treatment-induced clearance of HCV. *IFNL4* is not expressed, which in turn induces a weaker, and possibly differential, expression of ISGs. 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% were infected by a virus with a serine at position 2,414 and 33% with a different amino acid at that position. (**b**) Infected individuals with *IFNL4* non-CC genotypes (63% of the HCV genotype 3a–infected patients) had lower rates of spontaneous and treatment-induced clearance of HCV. *IFNL4* is expressed and induces expression of ISGs that collectively establish an antiviral state that is hostile to viral replication. This hostile environment induces a high selective pressure, and fewer viral mutations can accumulate. Although the serine at position 2,414 (as compared to non-serine residues) 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 2,414 and 15% with a different amino acid at that position).

driving viral evolution and serve 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 HLAs expressed by an individual, we now show that host IFNL4 activity may significantly shape the HCV viral genome. Previous studies have shown that the 'favorable' 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 favorable CC genotype (rs12979860), abolishes the expression of *IFNL4* (ref. 36), whereas in individuals with the *IFNL4* ΔG/TT or ΔG/ΔG genotypes (which are linked to 'unfavorable' non-CC genotypes), *IFNL4* is expressed, leading to the downstream upregulation of hepatic ISG expression via the JAK–STAT signaling pathway[37]. The expression of ISGs has been shown to render the host less susceptible to stimulation by exogenously added IFN-α and IFN-γ, and it is associated with more infected cells in the liver[41]. 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[41–43].

Our analysis adds to this hypothesis with the observation that *IFNL4* variants also have an effect on the HCV genome at multiple amino acid sites. Indeed, these were the strongest footprinting signals in our systematic analysis outside of the HLA region. The most significant association was at residue 2,570 in the viral NS5B protein, an association that we replicated in an independent cohort, whose members were infected with HCV genotypes 2 and 3. The additional signals in the NS5B protein that were associated with *IFNL4* genotypes (or HLA alleles) at a 5% FDR showed no significant association ($P > 0.05$) in our replication cohort, 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 manner (only NS5B sequences were available in the replication cohort). An association between *IFNL4* genotypes and amino acid variability in the same region of the NS5B protein has previously been reported in a candidate gene study[32] in a cohort with a single-source infection. *IFNL4* has also been associated with a viral mutation associated with DAA drug resistance in HCV genotype 1 infection[44], although this association has not been replicated in our HCV genotype 3–infected cohort[45]. However, the broader effect of the interferon lambda host genes that are associated with, and potentially select for, specific viral variants has not been previously recognized.

HCV viral load is an important and clinically relevant parameter, as patients with higher HCV viral loads have lower response rates to IFN- and DAA-drug-based therapy[46] (independent of *IFNL4* status). Paradoxically, the favorable *IFNL4* variants have also been associated with both an increase in disease progression[47] and high viral load[3,48]. We report an association between a virus amino acid site (serine versus non-serine at position 2,414 in NS5A) and HCV viral load. This site is one of the 11 sites that are putatively associated with *IFNL4* genotypes in our data set. Furthermore, our data show that a decrease in viral load was observed only in those patients with non-CC genotypes whose virus encoded a serine at amino acid residue 2,414. Because the Ser2,414 variant is found in 85% of patients with the non-CC genotypes (as compared to 67% of patients with the CC genotype), this interplay between host and viral genes helps explain the previous observation that patients with the non-CC genotypes have a lower HCV viral load (**Fig. 5**). Our *in vitro* data from a genotype 3 replicon assay showed that a change from a serine to asparagine at site 2,414 was associated with an increase in RNA replication and perhaps hyper-phosphorylation of the HCV NS5A protein[49], which is a negative regulator of virus replication.

Our interpretation of the data (**Fig. 5**) is that the expression of *IFNL4* by the ΔG/TT or ΔG/ΔG genotypes (tagged in our cohort by the 'unfavorable' 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 encoded by the viral genome (most notably amino acid 2,414 in NS5A, which has a significant impact on viral load; **Figs. 2b** and **4** and **Supplementary Fig. 15**). Our data suggest that this also leads to an overall increase in the strength of purifying selection (decrease in dN/dS; **Fig. 3**), and together this leads to lower viral load. However, viruses that establish chronic infections in patients with non-CC genotypes have evolved to survive in a more hostile environment (for example, by mutating the Ser2,414 of NS5A), which makes them less likely to respond to interferon-based therapy. In contrast, our analysis suggests that the favorable CC genotype and the inactivation of the *IFNL4* gene (by the *IFNL4* TT genotype) disables components of the immune response (therefore removing the effect of amino acid residue 2,414 in NS5A on viral load), which leads to a reduced level of purifying selection (**Fig. 3**). It is possible that this then permits a range of mutations that confer higher replicative fitness and, therefore, higher viral load (**Supplementary Fig. 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 effect of favorable CC and unfavorable 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. By 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 in only patients with the unfavorable non-CC *IFNL4* genotypes. These observations suggest that the innate and adaptive immune systems jointly have an effect on HCV genome evolution and together probably 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.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank Gilead Sciences for providing samples and data from the BOSON clinical study for use in these analyses. We also 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 (MRC) (MR/K01532X/1; to the 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; C.C.A.S. and R.B.). E.B. is funded by the MRC as an MRC Senior Clinical Fellow and received additional support from the NIHR Oxford BRC and the Oxford Martin School. M.A.A. is funded by the Oxford Martin School. G.C. is funded by the BRC of the Imperial College NHS Trust. P.K. is funded by the Oxford Martin School, the NIHR Oxford BRC, the Wellcome Trust (109965MA) and the US National Institutes of Health (U19AI082630). J.M. has funding support from the MRC (MC.UU.12014/1) and the Medical Research Foundation (C0365). C.C.A.S. is funded by Wellcome Trust grant 097364/Z/11/Z. G.M. is funded by Wellcome Trust grant 100956/Z/13/Z.

AUTHOR CONTRIBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

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**, 1338–1345 (2010).
5. Suppiah, V. *et al.* *IL28B* is associated with response to chronic hepatitis C interferon-α and ribavirin therapy. *Nat. Genet.* **41**, 1100–1104 (2009).
6. Tanaka, Y. *et al.* Genome-wide association of *IL28B* with response to pegylated interferon-α 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 Int.* **32**, 79–87 (2012).
11. Ahmed, A. & Felmlee, D.J. Mechanisms of hepatitis C viral resistance to direct-acting antivirals. *Viruses* **7**, 6716–6729 (2015).
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* **151**, 501–512 (2016).
13. Messina, J.P. *et al.* Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* **61**, 77–87 (2015).
14. Pol, S., Vallet-Pichard, A. & Corouge, M. Treatment of hepatitis C virus genotype 3 infection. *Liver Int.* **34** (Suppl. 1), 18–23 (2014).
15. Ampuero, J., Romero-Gómez, M. & Reddy, K.R. HCV genotype 3—the new treatment challenge. *Aliment. Pharmacol. Ther.* **39**, 686–698 (2014).
16. Fitzmaurice, K. *et al.* Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus infection. *Gut* **60**, 1563–1571 (2011).
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* **54**, 1157–1166 (2011).
18. Heim, M.H. & Thimme, R. Innate and adaptive immune responses in HCV infections. *J. Hepatol.* **61** (Suppl. 1), S14–S25 (2014).
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.* **6**, 261ra153 (2014).
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* **65**, 112–123 (2016).
21. Foster, G.R. *et al.* Efficacy of sofosbuvir plus ribavirin, with or without peginterferon-α, in patients with hepatitis C virus genotype 3 infection and treatment-experienced patients with cirrhosis and hepatitis C virus genotype 2 infection. *Gastroenterology* **149**, 1462–1470 (2015).
22. Bonsall, D. *et al.* ve-SEQ: robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000 Res.* **4**, 1062 (2015).
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.* **54**, 2470–2484 (2016).
24. Dilthey, A. *et al.* Multi-population classical HLA type imputation. *PLoS Comput. Biol.* **9**, e1002877 (2013).
25. Zheng, X. *et al.* HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).
26. Gonzalez-Galarza, F.F. *et al.* Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease, and HLA adverse drug reaction associations. *Nucleic Acids Res.* **43**, D784–D788 (2015).
27. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple-testing burden for genome-wide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
28. Johnson, R.C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11**, 724 (2010).
29. de Bakker, P.I.W. *et al.* A high-resolution HLA and SNP haplotype map for disease-association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
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* **66**, 114–124 (2005).
31. Ruhl, M. *et al.* CD8$^+$ T cell response promotes evolution of hepatitis C virus nonstructural proteins. *Gastroenterology* **140**, 2064–2073 (2011).
32. Merani, S. *et al.* Effect of immune pressure on hepatitis C virus evolution: insights from a single-source outbreak. *Hepatology* **53**, 396–405 (2011).
33. Rauch, A. *et al.* Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen–restricted immune pressure. *Hepatology* **50**, 1017–1029 (2009).
34. Gaudieri, S. *et al.* Evidence of viral adaptation to HLA class I–restricted immune pressure in chronic hepatitis C virus infection. *J. Virol.* **80**, 11094–11104 (2006).
35. Patin, E. *et al.* Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. *Gastroenterology* **143**, 1244–1252 (2012).
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.* **45**, 164–171 (2013).

37. Terczyńska-Dyla, E. *et al.* Reduced IFN-λ4 activity is associated with improved HCV clearance and reduced expression of interferon-stimulated genes. *Nat. Commun.* **5**, 5699 (2014).
38. Jacobson, I.M. *et al.* Sofosbuvir for hepatitis C genotype 2 or 3 in patients without treatment options. *N. Engl. J. Med.* **368**, 1867–1877 (2013).
39. Lawitz, E. *et al.* Sofosbuvir for previously untreated chronic hepatitis C infection. *N. Engl. J. Med.* **368**, 1878–1887 (2013).
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.* **60**, 2981–2992 (2016).
41. Sheahan, T. *et al.* Interferon-λ alleles predict innate antiviral immune responses and hepatitis C virus permissiveness. *Cell Host Microbe* **15**, 190–202 (2014).
42. Bibert, S. *et al.* *IL28B* expression depends on a novel TT/–G polymorphism, which improves HCV clearance prediction. *J. Exp. Med.* **210**, 1109–1116 (2013).
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.* **186**, 938–951 (2016).
44. Peiffer, K.-H. *et al.* Interferon-λ4 genotypes and resistance-associated variants in patients infected with hepatitis C virus genotypes 1 and 3. *Hepatology* **63**, 63–73 (2016).
45. Pedergnana, V. *et al.* Interferon-λ4 variant rs12979860 is not associated with RAV NS5A Y93H in hepatitis C virus genotype 3a. *Hepatology* **64**, 1377–1378 (2016).
46. McHutchison, J.G. *et al.* Peginterferon-α2b or α2a with ribavirin for treatment of hepatitis C infection. *N. Engl. J. Med.* **361**, 580–593 (2009).
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* **55**, 384–394 (2012).
48. Thompson, A.J. *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* **139**, 120–129 (2010).
49. Tellinghuisen, T.L., Foss, K.L. & Treadaway, J. Regulation of hepatitis C virion production via phosphorylation of the NS5A protein. *PLoS Pathog.* **4**, e1000032 (2008).

STOP-HCV Consortium:

Eleanor Barnes[3], Jonathan Ball[12], Diana Brainard[13], Gary Burgess[14], Graham Cooke[7], John Dillon[15], Graham R Foster[8], Charles Gore[16], Neil Guha[12], Rachel Halford[16], Cham Herath[17], Chris Holmes[5], Anita Howe[18], Emma Hudson[3], William Irving[12], Salim Khakoo[19], Paul Klenerman[3], Diana Koletzki[20], Natasha Martin[21], Benedetta Massetto[13], Tamyo Mbisa[22], John McHutchison[13], Jane McKeating[3], John McLauchlan[10], Alec Miners[23], Andrea Murray[24], Peter Shaw[25], Peter Simmonds[3], Chris C A Spencer[1], Paul Targett-Adams[26], Emma Thomson[10], Peter Vickerman[27] & Nicole Zitzmann[28]

[12]University of Nottingham, Queen's Medical Centre, Nottingham, UK. [13]Gilead Sciences, Inc., Foster City, California, USA. [14]Conatus Pharmaceuticals, San Diego, California, USA. [15]University of Dundee, Ninewells Hospital and Medical School, Dundee, UK. [16]Hepatitis C Trust, London, UK. [17]Gilead Sciences, Middlesex, UK. [18]BC Centre for Excellence in HIV–AIDS, St. Paul's Hospital, Vancouver, British Columbia, Canada. [19]University of Southampton, Southampton, UK. [20]Janssen Diagnostics, Beerse, Belgium. [21]University of California, San Diego, La Jolla, California, USA. [22]Public Health England, London, UK. [23]London School of Hygiene and Tropical Medicine, London, UK. [24]OncImmune Limited, Nottingham City Hospital, Nottingham, UK. [25]Merck and Company, Inc., Kenilworth, New Jersey, USA. [26]Medivir AB, Huddinge, Sweden. [27]University of Bristol, Clifton, UK. [28]University of Oxford, Oxford, UK.

ONLINE METHODS

Patients and samples. Plasma and DNA samples came from patients enrolled in the BOSON study. The BOSON study is a phase 3 randomized open-label trial to determine the efficacy and safety of treatment with sofosbuvir, with and without pegylated IFN-α, in treatment-experienced patients with cirrhosis and HCV genotype 2 infection and treatment-naive or treatment-experienced patients with HCV genotype 3 infection[21]. 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* and *HLA-C* and the MHC class II loci *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4* and *HLA-DRB5* using HLA*IMP:02 (accessed 22 March 2015)[24]. HLA amino acids were also imputed by SNP2HLA[50] using the T1DGC as the reference panel, which contains data from 5,225 unrelated individuals (10,450 haplotypes). Logistic regression using posterior genotype probabilities (allele dosages) for each HLA allele from SNP2HLA were carried out using PLINK2 (https://www.cog-genomics.org/plink2)[51].

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

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 5 µl sample (maximum 10 ng total RNA) and previously published modifications of the manufacturer's guidelines (v2.0)[22], including fragmentation for 5 min at 94 °C, omission of actinomycin D at first-strand reverse transcription, library amplification for 18 PCR cycles using custom indexed primers[52] 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 analyzed using Agilent TapeStation with D1K High Sensitivity Kit (Agilent) for equimolar pooling; they were then re-normalized by qPCR using the KAPA SYBR FAST qPCR Kit (Kapa Biosystems) for sequencing. A 500-ng aliquot of the pooled library was enriched using the xGen Lockdown protocol from Integrated DNA Technologies (IDT) (Rapid Protocol for DNA Probe Hybridization and Target Capture Using an Illumina TruSeq or Ion Torrent Library (v1.0)) with equimolar-pooled 120-nt DNA oligonucleotide probes (IDT) followed by a 12-cycle, modified, on-bead, post-enrichment PCR re-amplification step. The cleaned post-enrichment library was normalized with the aid of qPCR and sequenced with 151-base paired-end reads on a single run of the Illumina MiSeq using v2 chemistry.

Sequence data analysis. De-multiplexed sequence-read pairs were trimmed of low-quality bases using QUASR (v7.0120)[53] and of adaptor sequences using CutAdapt (version 1.7.1)[54], and they were subsequently discarded if either the read had less than 50 bases of remaining sequence or if both reads matched the human reference sequence using Bowtie (version 2.2.4)[55]. The remaining read pool was screened against a BLASTn database containing 165 HCV genomes[56], which covered its diversity both to choose an appropriate reference and to select those reads that formed a population for *de novo* assembly with Vicuna (v1.3)[57]. The assembly was finished with V-FAT v1.0 (http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/v-fat). The population consensus sequence at each site was defined as the most common variant at that site among all of the patients.

Phylogenetic and ancestral sequence reconstruction. Whole-genome viral consensus sequences for each patient were aligned using MAFFT[58] with default settings. This alignment was used to create a maximum-likelihood tree using RAxML[59], 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 RAxML[59] 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 the genome-to-genome level, we performed logistic regression using PLINK2 (https://www.cog-genomics.org/plink2)[51], adjusted for the human population structure (three first PCs as assessed using EIGENSOFT v3.0)[60] and the virus population structure (ten first PCs). For the viral data, PCA was performed on the nucleotide data as follows. Tri- and quad-allelic sites were converted to binary variables, and the amino acid frequencies were standardized 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 structures[61] by looking at viral mutations after infection. We constructed a 2 × 4 contingency table, where rows denoted presence or absence of a host HLA allele, and the columns denoted 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 or 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 corresponded 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 a 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 the enrichment of association signals in epitope regions, viral proteins or with a specific HLA allele, we used Fisher's exact test. Each site was either within a target region (epitope regions or a specific viral protein) or not, and the most associated test with it was 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 2 × 2 subtable used in the Fisher's exact test. This was done only for viral sites associated with HLA alleles at 20% FDR and for which there were a sufficient number of observations in both tables to estimate the odds ratio (OR) for which the confidence interval did not go to infinity. Pearson's correlation coefficient was used to assess the relationship between the $\log_{10}$(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 the number of associations in each protein was estimated using 10,000 permutations of *IFNL4* labels and performing the same tests. The estimated null distribution of the 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 the number of associations. To test whether the hypervariable region 1 (HVR1), HVR2, HVR3, the interferon-sensitivity-determining region (ISDR) and the RNA-dependent protein kinase–binding domain (PKR-BD) region showed differences in the number of changes away from the population consensus in hosts with CC or non-CC genotypes, we used a Poisson regression analysis. For each individual and each locus, we determined the number of differences relative to the population consensus. We then estimated the effect of *IFNL4* genotypes on the mean number of differences relative to the population consensus using Poisson regression. The same procedure was used to test whether the total number of differences across the whole polyprotein relative to the population consensus was influenced by *IFNL4* genotypes.

doi:10.1038/ng.3835

To estimate the rate of synonymous and non-synonymous mutations, we used the 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 and transversion rate bias, as well as base and codon frequency bias. It then 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 sequence, which indicates the most common nucleotide observed in our data set at each position along the genome.

To determine whether *IFNL4* genotypes have an effect on HLA allele presentation of epitopes, we used logistic regression with an interaction term. The outcome variable for each individual was whether there was any change in the specific amino acid on the terminal branch of the virus tree associated with that individual. We tested for interaction between presence or absence of the associated HLA allele and *IFNL4* genotypes for all of the combinations of HLA alleles and viral sites associated at a 20% FDR. In addition, we tested for an overall effect of *IFNL4* genotypes on the HLA alleles' presentation of epitopes. We used our 2 × 4 contingency tables and the ORs estimated from the 2 × 2 subtables to infer the antigenic amino acids. If the OR indicated that in individuals with HLA allele present the 'X ancestral amino acid any other amino acid' element was enriched relative to individuals without the HLA allele, then we assumed that the X amino acid was the antigenic amino acid and that escape occurred away from X to any other amino acid. For these cases, we count how many of 'X ancestral amino acid any other amino acid' incidences occurred in individuals with CC or non-CC genotypes across all combinations of HLA alleles and viral sites associated at 20% FDR. If the *IFNL4* genotypes have no effect on HLA presentation (null hypothesis), then the mean number of escape mutations in hosts with the CC or non-CC genotypes should be proportional to the frequency of hosts with the CC or 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 hosts with the CC genotype).

We used linear regression in PLINK2 (https://www.cog-genomics.org/plink2)[51] to test for association between human SNPs and $\log_{10}$-transformed pretreatment viral load (PTVL), including sex and the 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 ten at each site) and viral load. We used linear regression in R version 3.2.4 (2016-03-10)[62] to analyze the interactions between *IFNL4* genotypes and amino acids at viral site 2,414 and to quantify their effect on viral load.

For all of the viral sites associated with *IFNL4* genotypes at 20% FDR, we assessed whether there was a relationship between the effect size of non-consensus amino acids on viral load and the effect size of *IFNL4* genotypes and changes to non-consensus amino acids. We estimated the OR of enrichment of changes away from the consensus amino acid on the terminal branches of the virus tree in individuals with CC and non-CC genotypes (2 × 2 contingency table; using only data in which the consensus was inferred to be ancestral). We also estimated the effect size of non-consensus amino acids on the $\log_{10}$ value of viral load using a linear regression analysis with *IFNL4* genotype as a covariate. Additionally, we estimated the effect size of the non-consensus amino acids on the $\log_{10}$(viral load) in hosts with CC or non-CC genotypes. We used Pearson's correlation coefficient to measure the strength of the relationship between the effect size of non-consensus amino acids on viral load and the log of the OR of enrichment of non-consensus amino acid changes in hosts with a CC genotype relative to that in hosts with non-CC genotypes.

Code availability. The 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 population of HCV-infected individuals 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 the NS5B amplicon were obtained from DNA and plasma samples collected from 360 self-reported white patients who were chronically infected with HCV genotype 2 ($n = 153$) or HCV genotype 3 ($n = 208$). We searched for association between the *IFNL4* SNP rs12979860 and viral amino acid position 2,570 using logistic regression with the outcome indicated by the presence or absence of a valine at that residue. To help prevent spurious associations due to host and viral stratification, we included human PCs and viral genotype as covariates.

Replicon assay. *Cell culture*. Huh7.5-Sec14L2 human cells, obtained from J. Witteveldt[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 mM HEPES and 0.1 M non-essential amino acids as described[63]. The genotype of Huh7.5 cell line for the *IFNL4* SNP (rs12979860) is heterozygous CT[64].

HCV mutant replicons. The enhanced version of the subgenomic replicon of genotype 3a strain S52, which lacks the neomycin resistance gene, has been previously described[40]. Site-specific mutations were introduced using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies), following the manufacturer's instructions, and confirmed by direct sequencing. Plasmids with mutated HCV sequences were linearized by *XbaI* digestion (New England BioLabs; NEB), treated with mung bean nuclease (NEB) and purified. Linearized DNA was then used as a template for *in vitro* RNA transcription (IVT) (Megascript T7, Life Technologies) according to the manufacturer's protocol. Finally, the IVT RNA was 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 4-mm cuvette and electroporated using the Gene Pulser Xcell (Bio-Rad) at 250 V, 950 μF and the 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 h, medium was removed and cells were lysed with Glo Lysis Buffer (Promega). Cell lysates were then transferred in a white 96-well plate (Corning), and the amount of luciferase expression was quantified in a luminometer (GloMax 96 Microplate Luminometer, Promega) using the Bright-Glo assay system (Promega).

Data availability. Human genotype data underlying this manuscript are deposited in the European Genome–phenome Archive under accession code EGAS00001002324. HCV sequence data underlying this manuscript are deposited in GenBank under accession codes KY620313–KY620880. Information on access to the study data is available at http://www.stop-hcv.ox.ac.uk/data-access.

50. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
51. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer data sets. *Gigascience* **4**, 7 (2015).
52. Lamble, S. *et al.* Improved workflows for high-throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol.* **13**, 104 (2013).
53. Gaidatzis, D., Lerch, A., Hahne, F., Stadler, M.B. & Quas, R. Quantification and annotation of short reads in R. *Bioinformatics* **31**, 1130–1132 (2015).
54. Martin, M. Cutadapt removes adapter sequences from high-throughput-sequencing reads. *EMBnet.journal* **17**, 10 (2011).
55. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
56. Smith, D.B. *et al.* Expanded classification of hepatitis C virus into seven genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology* **59**, 318–327 (2014).
57. Yang, X. *et al. De novo* assembly of highly diverse viral populations. *BMC Genomics* **13**, 475 (2012).
58. Katoh, K. & Standley, D.M. MAFFT multiple-sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
59. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
60. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
61. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315**, 1583–1586 (2007).
62. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2016).
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.* **25**, 5372–5376 (2015).
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.* **21**, 19–24 (2014).