SUPPLEMENTARY INFORMATION

I. Patient characteristics

A total of 1671 genotype 1 hepatitis C patients were included in the study. We initially excluded 2 individuals that had an undetectable HCV RNA result at baseline, 27 individuals with insufficient phenotype data and 27 individuals of Asian or unknown ancestry. Table S1 lists the clinical characteristics of the remaining 1615 patients. Treatment response and non-response (NR) were defined according to standard definitions [1]. Sustained virological response (SVR) was defined as undetectable serum HCV RNA using a sensitive RT-PCR assay 24 weeks after cessation of treatment (or undetectable viral levels at 12 weeks follow-up if no further follow-up was available). Non-response was defined either as a failure to achieve at least a 2-$\log_{10}$ reduction in serum HCV RNA at week 12 of treatment, or as detectable serum HCV RNA at the end of follow-up. Subjects with insufficient viral response at 12 or 24 weeks discontinued therapy per protocol as treatment failures. All patients who achieved an SVR were included in the analysis as responders. In order to ensure that only non-responders with adequate drug exposure were evaluated (true biological non-responders), we included only patients with a minimum of 12 weeks of therapy and compliance of greater than 80% for

both PegIFN and RBV. We also applied a series of quality control procedures to ensure the data quality (part III). A total of 1,137 patients with sufficient treatment response data fulfilled the inclusion (described in the main text) and quality control (described below) criteria, and were then included in the main association analysis (Table S2). We also carried out a genome-wide association test on baseline viral load. A total of 1,475 patients with sufficient viral load data were included in that analysis (Table S3). All subjects signed informed consent to participate. In the final analysis 62 were from the study population of Muir et al.[5]

Table S1. Clinical characteristics of the overall hepatitis C populations

	Populations		
	Caucasians	African Americans	Hispanics
N	1186	299	130
Sex (F/M)	456/730	131/168	42/88
Age (yrs)	47.3(7.6)	50.0(7.0)	44.(9.4)
BMI (kg/m^2)	27.9(4.6)	29.8(5.0)	29.1(5.2)
Fasting glucose level (mmol/L)	5.3(0.9)	5.4(1.0)	5.3(0.7)
Baseline ALT *	2.2(1.7)	1.7(1.3)	2.6(1.9)
Baseline viral load (log_{10} IU/mL)	6.4(0.6)	6.3(0.5)	6.2(0.7)
(<600,000 IU/mL / >=600,000 IU/mL)	191/995	55/244	37/93
Baseline liver fibrosis stage (n, %) **			
F0	18 (1.6%)	2 (0.7%)	3 (2.3%)
F1	804 (70.8%)	191 (67.7%)	92 (71.9%)
F2	177 (15.6%)	60 (21.3%)	15 (11.7%)
F3	62 (5.5%)	8 (2.8%)	8 (6.3%)
F4	75 (6.6%)	21 (7.5%)	10 (7.8%)
Steatosis (none/any) **	451/685	97/185	36/92
IFN treatment regimen ***			
PegIFN alfa-2a 180ug/wk	387 (32.6%)	95 (31.8%)	39 (30.0%)
PegIFN alfa-2b 1.0 mcg/kg/wk	383 (32.3%)	89 (29.7%)	45 (34.6%)
PegIFN alfa-2b 1.5 mcg/kg/wk	416 (35.1%)	115 (38.5%)	46 (35.4%)
Assigned RBV dose	525/661	175/124	53/77
(<13 mg/kg/d / >=13 mg/kg/d)			

SVR, sustained virological response (SVR24 = 539, SVR12 = 36) ; NR, Non-response; BMI, body mass index. Basal viral load is logarithmically transformed. Fibrosis was scored by METAVIR stage on a baseline centrally evaluated liver biopsy [3, 4]. Steatosis was scored as none (0%) vs any (including 0 – 5%, and >5% [4]) . Data are mean (SD) unless otherwise indicated. Ethnicity was self-reported by the patient.

* Ratio of the measured ALT levels over the upper limit of the normal (ULN), as described previously [4].

** METAVIR fibrosis score and Steatosis data are only available for 1546 participants where the liver biopsy specimen was deemed adequate for assessment by the central pathologist.

*** The IDEAL study [4] compared the effectiveness of three treatment regimens involving peginterferon alfa-2b (PegIFN alfa-2b) or peginterferon alfa-2a (PegIFN alfa-2a) combined with ribavirin (RBV). The study results demonstrated similar efficacy of the two IFN preparations and of the three treatment regimens.

Table S2. Clinical characteristics of the hepatitis C populations for studying SVR

	Populations		
	Caucasians	African Americans	Hispanics
N	871	191	75
Sex (F/M)	331/540	71/120	29/46
Age (yrs)	47.5(7.2)	50.1(6.5)	45.3(9.2)
BMI (kg/m^2)	28.0(4.4)	29.9(4.8)	29.3(5.5)
Baseline viral load (log_{10} IU/mL)	6.4(0.6)	6.3(0.5)	6.2(0.7)
Baseline liver fibrosis stage (n, %)	770 (88.4%)	174 (91.1%)	63 (84.0%)
Minimal (F0-2)	101 (11.6%)	17 (8.9%)	12 (16.0%)
Advanced (F3-4)			
SVR/NR (SVR%)	**488/383 (56.0%)**	**45/146 (23.6%)**	**38/37 (50.7%)**

SVR, sustained virological response (SVR24 = 535, SVR12 = 36) ; NR, Non-response; BMI, body mass index.

Baseline viral load is logarithmically transformed. Fibrosis was scored by METAVIR stage on a baseline centrally evaluated liver biopsy [3, 4]. Data are mean (SD) unless otherwise indicated. Ethnicity was self-reported by the patient.

Table S3. Clinical characteristics of the hepatitis C populations for studying baseline viral load

	Populations		
	Caucasians	African Americans	Hispanics
N	1116	263	96
Sex (F/M)	434/682	110/153	33/63
Age (yrs)	47.4(7.5)	50.1(6.7)	44.4(9.6)
BMI (kg/m^2)	27.9(4.6)	29.8(5.0)	29.3(5.4)
Baseline viral load (log_{10} IU/mL)	**6.4 (0.6)**	**6.3 (0.5)**	**6.2 (0.7)**

BMI, body mass index.

Baseline viral load is logarithmically transformed. Data are mean (SD) unless otherwise indicated. Ethnicity

II. An additional random multi-ethnic population sample

We used a random multi-ethnic population sample to check the C allele frequency in different populations with unknown hepatitis C status, including an East Asian population. These participants were apparently healthy individuals recruited from in and around the Duke and NC State University campuses in North Carolina. Table S4 shows the general characteristics of this population sample. All subjects signed informed consent to participate in genetic studies. They were genotyped using the Illumina Human 610-Quad BeadChip and the genotype data were subject to quality control procedures as described for the HCV cohort.

Table S4. General characteristics of the random multi-ethnic population sample

	Populations			
	Caucasians	African Americans	Hispanics	East Asians
N	271	61	16	107
Sex (F/M)	141/130	41/20	7/9	59/48
Age (yrs)	23.5(7.4)	28.5(12.2)	23.3(6.8)	21.1(2.6)

Data are mean (SD) unless otherwise indicated. Ethnicity was self-reported by the participant.

III. Quality control of data flow

The following quality control steps were taken to make sure genotypes were correctly called.

1. Infinium BeadStudio Raw Data Analysis

All samples were brought into a single BeadStudio file using the standard Illumina cluster file. Any sample that had very low intensity or a very low call rate using the Illumina cluster (<95%) was deleted. All SNPs that had a call frequency below 100% were then reclustered. Any sample that was below a 98% call rate after the reclustering was deleted. Next, a "1% rule" was applied where all SNPs that had a call frequency below 99% were deleted. Any SNPs where more than 1% of samples were not called or were ambiguously called were deleted. We have shown (unpublished data) that SNPs with many samples not called (or potentially miscalled) can lead to false positives in statistical associations. A total of 24 samples were deleted during this procedure.

The reclustering step creates SNP calling errors, but we have identified a procedure to prevent the errant calls from being released in the final report. The SNP data is screened within BeadStudio by looking at two criteria. First, all SNPs with a cluster separation value below 0.3 are manually checked to ensure correct calls. Many of these SNPs can be manually fixed, but some have to be deleted. Next, any SNP (excluding X chromosome SNPs) with a Het Excess value between -1.0 to -0.1 and 0.1 to 1.0 are evaluated to determine if the raw and normalized data show a clean call. Any SNP cluster that doesn't appear normal is deleted. This includes SNPs that appear to show a deletion (hemizygotes and homozygous deletion). The rational behind is to avoid artifacts from either the chemistry or an interfering SNP during hybridization.

These procedures resulted in a success rate of genotyping calls of 97.5%.

2. Minor allele frequency (MAF) check for data handling accuracy

This step performs a basic check of the data accuracy on the data flow pipeline from the output of the Illumina genotyping facility to the analytical process. We checked the MAF report from PLINK software [6] (http://pngu.mgh.harvard.edu/~purcell/plink/) against the original locus report generated by genotyping facility. We checked that the two MAF reports match exactly.

3. Specification of gender

This step performs a check on the gender specification obtained from the phenotype database, using the observed genotypes of SNPs on chromosome X and Y. All individuals who were marked as "male" but with significant amount of heterozygous X genotypes (>=1%), or who were marked as "female" but with high frequency of homozygous X genotypes (>=80%) or Y genotype readings, were individually inspected against original data source. If no satisfactory correction could be obtained these individuals were excluded from further analyses. Two subjects were excluded at this step.

4. Cryptic relatedness

This step performs a check on the cryptic relatedness between study participants. We estimated the sharing of genetic information by estimating identity by descent (IBD) using the PLINK software. All pairs of DNA samples showing >=0.125 (estimated proportion of alleles IBD) were individually inspected and one sample in each pair was excluded from further analyses. One subject was excluded at this step.

5. Genotype missing

This step performs a check whether the genotype missing is skewed towards high or low phenotype values and hence may give rise to spurious association P-values. We used PLINK software to perform this check on the top SNPs discussed in the paper.

6. Low MAF

We removed all SNPs with a MAF<0.01. This criterion ensured that at least 6 individuals of the rare genotype are present in the dataset, to control for error in the estimation of asymptotic P-values (as alleles with MAF this low or lower have no chance of approaching significance).

7. Hardy-Weinberg Equilibrium (HWE)

This step performs a check whether the observed genotype data deviate from HWE. We performed this check using PLINK software on the top SNPs showing genome-wide significant association. We defined a deviation from HWE with a criterion of P-value less than 0.05.

8. Recheck of the genotyping quality

The top SNPs showing genome-wide significant association were subject to a double check for their genotyping quality. This is an individual recheck on the raw and normalized data to be sure that it is called correctly as described in “Infinium BeadStudio Raw Data Analysis” process.

IV. Modified EIGENSTRAT [7] method to control for stratification

This method derives the principal components (PC) of the correlations among gene variants and corrects for those correlations in the association tests. In principle, the principal components in the analyses should reflect population ancestry. We have noticed however that some of the leading axes appear to depend on other sources of correlation, such as sets of variants near one another that show extended association. We have documented the potential for inversions to create this effect and it may be created by other causes of extended LD as well. For this reason we inspected the SNP 'loadings' for each of the leading axes to determine if they depended on many or relatively few SNPs, as would be expected if the given axis reflected population ancestry or a more localized LD effect respectively.

We selected EIGENSTRAT axes for use as covariates to adjust for ancestry in subsequent linear regression analyses as follows (procedure repeated within each ancestry group).

1. To find EIGENSTRAT axes, we started with autosomal SNPs with MAF>0.01.

2. On inspection of SNP loadings for each PC axis (the "gamma" coefficients of Price et al [7]), we found several of the top axes to be dominated by a small number of SNPs all mapping to the same region of the genome. For example, one axis was found to be dominated by SNPs mapping to a region of chr8p22-23.1 coinciding with a known inversion polymorphism.

3. To correct for these LD effects, and ensure that EIGENSTRAT axes reflected only effects that applied equally across the whole genome (as ancestry effects should), we re-applied principal components analysis to a reduced SNP set in which (i) certain known high LD regions were excluded (chr8:8000000..12000000, chr6:25000000..33500000, chr11:45000000..57000000, chr5:44000000..51500000); (ii) SNPs were thinned using the "--indep-pairwise" option in PLINK, such that all SNPs within a window size of 1500 (step size of 150) were required to have r2<0.2; (iii)

Each SNP was regressed on the previous 5 SNPs, and the residual entered into the PCA analysis, as suggested by Patterson et al (2).

4. Inspection of SNP loadings on all axes deemed significant by the Tracy-Widom method of Patterson et al (2), using Q-Q plots against Normal expectation, revealed no axes dominated by single high-LD regions of the genome.

5. Tracy Widom tests nominated the first 5 resulting PC axes as significant ($p<0.05$) in Caucasians, the first 4 axes in African Americans and the first 3 axes in Hispanics. We then included the PC values emerging from the EIGENSTRAT analyses as covariates in all the regression models.

V. Generation of CNV Calls Using the Intensity Data and QC

All subjects that passed SNP QC procedures were entered into the CNV analysis. The CNV calls were generated using the PennCNV software [8] using the Log R ratio (LRR) and B allele frequency (BAF) measures automatically computed from the signal intensity files by BeadStudio, and the standard hg18 PennCNV hidden Markov model (hmm) and population frequency of B allele (pfb) files for the Infinium Human610-Quad BeadChip. We implemented the gcmodel wave adjustment procedure and used the PennCNV quality control procedures to exclude samples that failed quality control. These included samples that have a LRR standard deviation >0.28, BAF median>0.55 or <0.45, BAF drift > 0.002 or WF>0.04 or <-0.04. Due to the complications of hemizygosity in males and X-chromosome inactivation in females, this analysis was restricted to autosomes. Additionally, to ensure that we worked with high-confidence CNVs, we excluded any CNV for which the difference of the log likelihood of the most likely copy number state and the less likely copy number state was less than 10 (generated using the -conf function in PennCNV). Finally, some centromeric and telomeric regions are not well mapped, and this can potentially result in CNV-calling errors in these regions (Dr. Kai Wang, personal communication). Also, genomic regions coding for immunoglobulin genes have previously been shown to be potential sites of false-positive PennCNV calls [8]. Our own research has shown that calls in both of these types of region differed significantly depending on the sample type used for DNA extraction (significant difference $p<10^{-10}$ for deletion and/or duplication frequencies between samples genotyped on DNA extracted from blood or saliva, data not shown). We therefore excluded any CNV that overlapped any of the following regions by 50% or more of its length: chr2: 87.0-92.0, chr14: 18-23.6 Mb, chr14: 104.5-106.5 Mb, chr15: 17.0-21.0, chr16: 31.8 -36.0Mb, chr22: 20.5-21.8 Mb (immunoglobin regions); chr1:0-4Mb, 240-247Mb; chr2: 87.0-92.0 Mb; chr4: 0-1.43 Mb, 48.75-49 Mb; 190.7-191.3 Mb; chr7:0-200kb; 56.5-62.5Mb; chr8: 39-45Mb, 145-146.3Mb; chr9: 44.5-70.1Mb;

138-140.2Mb; chr10: 38.5-42Mb; 134-135.4; chr11: 0-1.8Mb; chr14: 18-23.6 Mb, 104.5-106.5 Mb; chr15: 17.0-21.100-100.3Mb; chr16: 0-2.1Mb, 31.8 -36.0Mb, 86.6-88.9Mb;chr17:0-1Mb, 76.5-78.8Mb; chr18: 14-16Mb, 75.5-76Mb; chr19: 0-2.1Mb, 25.7-28.3Mb, 61.5-62.5Mb; chr20: 25.7-28.3Mb, 61.5-62.5Mb;chr21:9.7-14.3Mb; chr22:14.4-14.7Mb, 20.5-21.8 Mb (centromeric and telomeric regions, some overlapping immunoglobin regions as above). We also removed CNVs that spanned centromeres by searching for those larger than 1Mb with fewer than 50SNPs and checking their genomic locations.

VI. Statistical models for association tests in individual populations

Our primary association tests on sustained virological response (SVR) involved single-marker genotype trend tests performed in three independent groups, using logistic regression models implemented in the PLINK software [6] correcting for EIGENSTRAT [7] PC axes representing population ancestry information, and a number of clinical covariates, including baseline (pre-treatment) HCV viral load (≤600,000 IU/mL vs > 600,000 IU/mL) and fibrosis severity in the pre-treatment liver biopsy (mild/moderate = METAVIR F0-2, advanced = F3-4) [1, 9]. Other covariates considered for this model included gender, age, BMI, baseline ALT (alanine aminotransferase) levels, baseline liver steatosis status (0% vs >0%), baseline blood sugar levels, and weight-adjusted ribavirin dose (≤ 13mg/kg/d vs > 13mg/kg/d).

Our additional association tests on baseline viral load were performed using linear regression models, also correcting for EIGENSTRAT [7] PC axes representing population ancestry information. Other covariates considered for this model included sex, age, and BMI as above. Viral load were Box-Cox transformed to achieve normality.

VII. The importance of the phenotype definition

It is worth noting the importance of the phenotype definition in this study. As indicated, we included only patients with a minimum of 12 weeks of therapy and a compliance of greater than 80% for both PegIFN and RBV, to ensure that only non-responders with adequate drug exposure were evaluated (true biological non-responders). In comparison, if we ignored this quality control procedure, the association signal dropped dramatically from 10^{-25} to 10^{-8} level. This observation documents the critical role of accurate phenotyping in a successful genetic discovery.

VIII. Correction for multiple testing

We used the Bonferroni adjustment as the primary method to correct for multiple testing for single marker tests. We did evaluate the usage of permutation procedures in correcting multiple testing and found the results were a bit less conservative but largely very similar. We acknowledge that the Bonferroni adjustment is a conservative approach to interpret the results because of the dependency between the single marker tests. However, with the inconsistent results of previous complex disease genetic association studies, we contend that a conservative approach is appropriate. In addition, we firstly used a linkage disequilibrium pruning procedure to remove the dependent markers (defined as $r^2=1$) between the tests, and then use the Bonferroni adjustment based on the reduced dataset, so that the dependency will be better controlled than a conventional Bonferroni correction. We showed in our previous study [10] that the *P* value cutoff estimated using this method is almost identical to the practical guideline for correcting multiple testing burden of nearly all common variants in European populations, as proposed by Mark Daly [11].

IX. SNPs showing genome-wide significant association with SVR in the *IL28B* region

We observed seven SNPs within a 17-Kb region around the *IL28B* gene showing genome-wide association with SVR (main text Figure 1, Table S5), including the top hit, rs12979860. The other six SNPs displays different degrees of linkage disequilibrium with rs12979860 (Table S5), and their effects can be largely explained by rs12979860, tested by adding the rs12979860 genotype into the individual logistic regression model for each SNP. The most strongly associated SNP in this region after accounting for rs12979860 is rs8099917 (7Kb downstream to the *IL28B* gene, rs12979860-corrected $P=1.18\times10^{-6}$, rs12979860-uncorrected $P = 4.37 \times10^{-26}$). One SNP rs12980602 is located relatively close to the *IL28A* gene (6Kb upstream to the *IL28A* gene, 17Kb downstream to the *IL28B* gene). The effect of this SNP is largely explained (corrected P=0.032) by the *IL28B* SNP rs12979860, but not vice versa (rs12979860 corrected $P= 2.54\times10^{-21}$). Therefore we concluded that the association signals in this region are mainly due to the *IL28B* gene variants, although a minimal effect from the *IL28A* gene cannot be entirely ruled out.

Table S5. SNPs in the *IL28B* region showing genome-wide association with SVR

		Caucasians		African Americans		Hispanics	
SNP	**P**	**r^2**	**D'**	**r^2**	**D'**	**r^2**	**D'**
rs12979860	1.21×10^{-28}	-	-	-	-	-	-
rs12980275	2.82×10^{-27}	0.88	0.96	0.56	0.90	0.88	1.00
rs8099917	4.37×10^{-26}	0.52	0.99	0.07	1.00	0.78	1.00
rs12972991	1.88×10^{-21}	0.63	0.96	0.08	1.00	0.78	1.00
rs8109886	1.32×10^{-16}	0.61	1.00	0.38	0.97	0.77	1.00
rs4803223	7.87×10^{-16}	0.26	0.80	0.04	0.82	0.66	0.90
rs12980602	5.94×10^{-9}	0.15	0.50	0.01	0.22	0.52	0.72

Linkage disequilibrium measures (r2 and D') are between each SNP and rs12979860.

X. Effects of treatment regimen on SVR

We also test the effect of treatment regimen on SVR across the three genotypes. The treatment regimen did not show any significant effect on SVR in our study population (P=0.53), confirming what was already observed in the clinical analysis of the IDEAL trial [4]. Upon addition of treatment regimen to the regression model, the genotypic effects of rs12979860 remained unchanged: therefore we did not include it as a covariate in the final model.

XI. Evaluating the inter-individual difference of SVR explained by predictors

To quantitatively compare the magnitude of different predictors of response for the patients studied here, we developed a simple logistic regression model which relates clinical predictors, as well as the rs12979860 genotype, to response rates (Box S1). We did not include in this model any host genetic variant that was previously reported as associated with treatment response [12, 13], because none of them has succeeded in gaining wide acceptance or explaining much of the variation in response to anti-HCV treatment.

Box S1. The logistic regression model for assessing probability of achieving SVR

$$P = \frac{1}{1 + e^{-(1.4\times G + 1.7\times V + 1.1\times E + 1.1\times F - 3.8)}}$$, where

P: Probability of achieving SVR;

G: rs12979860 genotype: TT=0, CT=1, CC=2;

V: Baseline viral load : ≥ 600,000 IU/mL =0, <600,000 IU/mL=1;

E: Ethnicity: African Ancestry =0, Caucasian =1;

F: Baseline fibrosis: METAVIR F3-4 = 0, F0-2 =1

Logistic regression does not have a direct equivalent to the R^2 that is found in ordinary least squares (OLS) regression that represents the proportion of variance explained by the predictors. However, it is possible to use an analog, so-called a pseudo-R^2, to mimic the OLS-R^2 in evaluating the goodness-of-fit and the variability explained, which is the approach we used [14].

Using this approach we estimated that rs12979860 could account for 58% of the ethnicity-explained variability by estimating the difference between the expected variability if the *IL28B* SNP does not account for the variability explained by ethnicity at all, and the observed variability explained by both ethnicity and rs12979860.

XII. GWAS P-value distribution

Figure S1 shows a Q-Q plot of the GWAS P-values for SVR, suggesting no inflation of association signals after correction for population ancestries, and significant deviations from the expected for the SNPs around the *IL28B* gene, with the top one rs12979860.

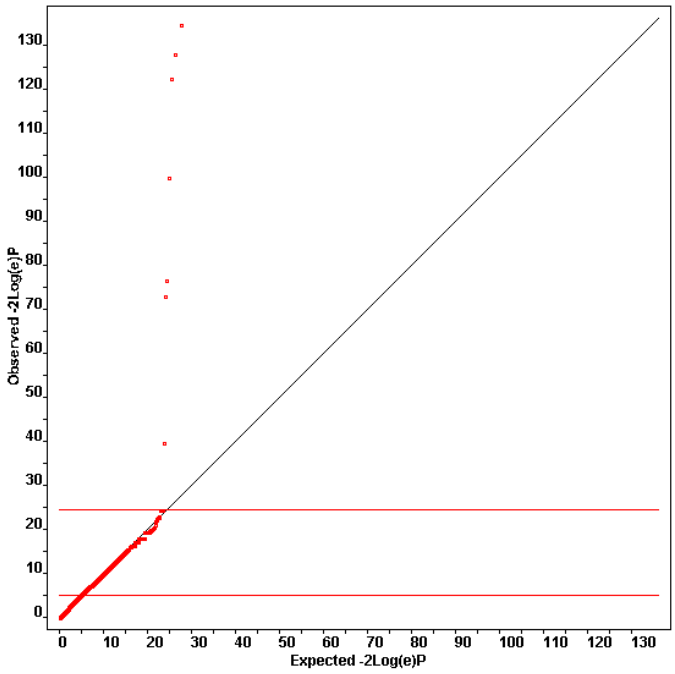

Figure S1. Q-Q plot of the GWAS P-values for SVR. Y axis is the observed -2Log(P), while X axis is the expected -2Log(P) values. The lambda value is 0.9974, suggesting no inflation of association signals after correction for population ancestries. Showed also include data points significantly lifting from the expected values, around the *IL28B* gene on chromosome 19, with the top one rs12979860.

XIII. Association with baseline HCV viral load

We tested whether rs12979860 influences baseline (pre-treatment) viral load and found a significant association in all groups (Table S6). Interestingly, the C allele, associated with better treatment response, is also associated with higher baseline viral load ($P = 1.08 \times 10^{-10}$). While counter-intuitive in that lower basal viral load predicts a better response to treatment, our finding may relate to recent observations of the role of interferon stimulated gene (ISG) expression in modulating response to PegIFN [15]. Sarasin-Filipowicz [15] showed that infected individuals with low basal levels of hepatic ISG expression had stronger up-regulation of ISG upon administration of PegIFN and a better response to treatment. In contrast, non-responders to PegIFN / RBV therapy were characterized by high level baseline hepatic ISG expression, that was refractory to further stimulation. The authors then argued that activation of the endogenous IFN system in chronic hepatitis C patients not only was ineffective in clearing the infection but also might impede the response to therapy, most likely by inducing a refractory state of the IFN signaling pathway. It is possible that the *IL28B* polymorphism plays a role in the regulation of intra-hepatic ISG expression activated by either the endogenous IFN system or the externally administrated IFN, with independent consequences both for pre-treatment viral load and treatment response.

In clinical practice it is customary to divide subjects into high and low viral load groups which has been shown at a population level to be predictive of treatment response, with a threshold set usually somewhere between 400,000 and 800,000 (IU, corresponding to about 2.5 logarithmically transformed IU/mL) per ml [4]. The IDEAL trial itself used a threshold of 600,000 [4]. For this reason, we first tested whether rs12979860 associates with whether individuals are above or below this threshold and found no significant association (P= 0.21). *Since this threshold appears most relevant to response, it indicates that the effect of the polymorphism on SVR is independent from its association with*

quantitative viral load. We next tested the polymorphism for an association with quantitative viral load within the high and low viral load groups (Table S7). Strangely, we found that within the high viral group the C allele positively associated with viral load, while within the lower viral load group the relationship was reversed. In both groups, however, the C allele was associated with an increased rate of response to treatment (Table S8). These analyses would suggest that however the polymorphism effects viral load, this effect is unrelated to how the polymorphism influences response to the treatment.

Table S6. Results of the GWAS on baseline HCV viral load in individual and combined Hepatitis C populations

	Baseline HCV viral load	
	N	P value
Caucasians	1116	3.48×10^{-8}
African Americans	263	8.46×10^{-5}
Hispanics	96	1.84×10^{-2}
Combined	1475	**1.21×10^{-10} ***

* Effects were consistent in direction and P values were combined using the Stouffer's weight Z-method [2].

Table S7. The association between the *IL28B* gene polymorphism and baseline viral load in high and low viral load groups

	Caucasians			African Americans			Hispanics		
	Beta (95% CI)	P	n	Beta (95% CI)	P	n	Beta (95% CI)	P	n
High viral load (>600,000 IU/ml)	14.8 (11.1-18.6)	2.24×10^{-14}	935	14.8 (7.9- 21.8)	4.0×10^{-5}	218	16.1 (2.8-29.4)	0.019	68
Low viral load (<=600,000 IU/ml)	-7.8 (-13.2--2.4)	0.005	181	-1.5 (-9.2-6.2)	0.696	45	-9.7 (-25.8-6.4)	0.224	28

CI, confidence interval.
The beta and 95% confidence intervals are generated from the linear regression model for Box-Cox transformed baseline viral load.

Table S8. The association between the *IL28B* gene polymorphism and SVR in high and low viral load groups

	Caucasians			African Americans			Hispanics		
	OR (95% CI)	P	n	OR (95% CI)	P	n	OR (95% CI)	P	n
High viral load (>600,000 IU/ml)	7.2 (4.9-10.4)	1.37×10^{-22}	728	5.2 (11.9-14.4)	0.006	160	4.0 (0.8-19.7)	0.09	54
Low viral load (<=600,000 IU/ml)	15.7 (3.3-73.6)	7.50×10^{-5}	143	Undefined(predicting SVR perfectly)	-	31	Undefined(predicting SVR perfectly)	-	21

SVR, sustained virological response. CI, confidence interval.
The odds ratios and 95% confidence intervals are generated from the logistic regression model.

XIV. Top 100 SNPs revealed in the GWAS

Table S7. Top SNPs revealed in the GWAS for SVR

SNP	Rank	Combined* P	P (Caucasians)	P (AAs)	P(Hispanics)	Chr.	Coordinate	Type	Ancestral allele	Closest gene	Distance to gene	Distance to exon boundary
rs12979860	1	1.37E-28	1.06E-25	0.0021	0.0044	19	44430627	UPSTREAM	T	*IL28B*	-3018	-
rs12980275	2	2.54E-27	2.36E-25	0.0304	0.0036	19	44423623	DOWNSTREAM	G	*IL28B*	2410	-
rs8099917	3	3.70E-26	8.58E-26	0.6448	0.0015	19	44435005	DOWNSTREAM	T	*IL28B /AC011445.6*	2391	-
rs12972991	4	1.72E-21	3.56E-20	0.1097	0.0063	19	44423587	DOWNSTREAM	A	*IL28B*	2446	-
rs8109886	5	1.54E-16	5.93E-15	0.0399	0.0078	19	44434602	DOWNSTREAM	A	*IL28B /AC011445.6*	2794	-
rs4803223	6	7.85E-16	7.52E-14	0.0177	0.0009	19	44438059	UPSTREAM	A	*IL28B /AC011445.6*	-310	-
rs12980602	7	6.10E-09	1.93E-08	0.7209	0.001	19	44444660	INTERGENIC	C	*IL28A*	-6337	-
rs9400317	8	6.77E-06	3.77E-05	0.0666	0.2237	6	110114340	INTRONIC	G	*AKD2*	0	-
rs17067123	9	7.99E-06	3.76E-05	0.0432	0.7912	4	180285356	INTERGENIC	C	*AC017087.5*	-441252	-
rs892015	10	1.06E-05	0.0001	0.0238	0.0914	19	44376749	UPSTREAM	C	*NCCRP1*	-2692	-
rs1421385	11	1.52E-05	8.79E-06	0.9966	0.9829	2	159702294	INTRONIC	G	*TANC1*	0	-
rs6875273	12	2.12E-05	8.51E-05	0.0811	0.5002	5	2325248	INTERGENIC	A	*AC138982.2*	87314	-
rs8106521	13	2.28E-05	0.0003	0.0146	0.038	19	44375555	UPSTREAM	G	*NCCRP1*	-3886	-
rs7952846	14	2.36E-05	3.00E-05	0.6303	0.4483	12	65140621	INTRONIC	T	*GRIP1*	0	-
rs873354	15	2.90E-05	0.0003	0.0081	0.3061	8	66373771	INTERGENIC	G	*AC060775.5*	-78634	-
rs3798004	16	3.48E-05	6.99E-05	0.1282	0.4861	5	9423945	INTRONIC	C	*SEMA5A*	0	-
rs10492508	17	3.61E-05	0.0002	0.0196	0.5064	13	19727582	INTERGENIC	C	*GJB6*	-23126	-
rs11160804	18	3.73E-05	9.09E-05	0.1443	0.9791	14	103863299	INTERGENIC	G	*AL512357.4*	75025	-
rs6444847	19	5.06E-05	9.14E-06	0.1963	0.9202	3	170387210	INTERGENIC	G	*EVI1*	-40156	-
rs17182166	20	5.23E-05	0.0004	0.0051	0.718	2	158304443	INTRONIC	G	*ACVR1*	0	-
rs11917147	21	5.49E-05	0.0005	0.0158	0.1399	3	171571142	INTRONIC	T	*SKIL*	0	-
rs7317784	22	5.68E-05	0.0005	0.0632	0.008	13	109753075	INTRONIC	T	*COL4A1*	0	-
rs1038923	23	5.98E-05	6.13E-05	0.9142	0.0972	12	65135675	INTRONIC	C	*GRIP1*	0	-
rs12425018	24	6.03E-05	7.38E-05	0.621	0.5607	12	129733134	INTERGENIC	T	*STX2*	106968	-
rs2272572	25	6.42E-05	0.0003	0.1125	0.1536	7	100123824	5PRIME_UTR	C	*GIGYF1*	0	-
rs970600	26	6.73E-05	3.81E-05	0.7961	0.3254	2	142595367	INTRONIC	C	*LRP1B*	0	-
rs6850226	27	7.10E-05	4.10E-05	0.6894	0.4288	4	180289571	INTERGENIC	T	*AC017087.5*	-445467	-
rs7487758	28	7.19E-05	6.58E-05	0.9516	0.3791	12	65166109	INTRONIC	T	*GRIP1*	0	-
rs1548928	29	7.41E-05	0.0002	0.0926	0.4992	16	47574271	INTERGENIC	A	*AC023827.7-2*	-251058	-
rs4660762	30	7.53E-05	9.34E-05	0.7714	0.2519	1	39699411	INTRONIC	A	*MACF1*	0	-
rs735214	31	7.78E-05	3.59E-05	0.8145	0.778	14	28282369	INTERGENIC	C	*FOXG1;FOXG1B*	-22432	-
rs11106877	32	8.19E-05	9.10E-05	0.8109	0.3387	12	76745666	UPSTREAM	A	*NAV3*	-3534	-
rs1011347	33	8.25E-05	0.0002	0.1466	0.859	12	129730850	INTERGENIC	G	*STX2*	109252	-
rs1036767	34	8.35E-05	4.37E-05	0.6561	0.5417	4	180305411	INTERGENIC	G	*AC017087.5*	-461307	-

rs13185011	35	8.45E-05	6.08E-05	0.9628	0.6494	5	165137993	DOWNSTREAM	T	*AC008562.5*	4390	-
rs1030460	36	8.48E-05	6.48E-05	0.6289	0.4445	12	20249480	INTERGENIC	C	*PDE3A*	-163984	-
rs4906450	37	9.06E-05	0.0002	0.2274	0.8882	14	103871245	INTERGENIC	A	*AL512357.4*	67079	-
rs8055576	38	9.39E-05	0.0003	0.3241	0.0821	16	80478031	INTRONIC	N/A	*PLCG2*	0	-
rs2899130	39	9.43E-05	0.0002	0.1156	0.6714	4	59252574	INTERGENIC	A	*N/A*	-9	-
rs4913307	40	9.44E-05	7.80E-05	0.9328	0.3526	12	65150558	INTRONIC	T	*GRIP1*	0	-
rs7076247	41	9.49E-05	0.0003	0.0291	0.1425	10	18799635	INTRONIC	C	*CACNB2*	0	-
rs12622986	42	0.0001	0.0007	0.0186	0.7047	2	10136397	INTRONIC	G	*CYS1*	0	-
rs4583650	43	0.0001	0.0001	0.8016	0.7909	3	63641857	INTERGENIC	T	*SNTN*	15939	-
rs16842883	44	0.0001	0.0002	0.3076	0.3645	3	136464085	DOWNSTREAM	G	*EPHB1*	2090	-
rs2166999	45	0.0001	0.0002	0.1139	0.6714	4	59247632	INTERGENIC	A	*N/A*	-9	-
rs6953952	46	0.0001	0.0001	0.9986	0.5179	7	105246348	INTRONIC	T	*ATXN7L1*	0	-
rs2440445	47	0.0001	0.0002	0.4754	0.1882	8	6330927	INTRONIC	T	*MCPH1*	0	-
rs10820447	48	0.0001	0.0002	0.2296	0.8516	9	98171865	INTRONIC	C	*SLC35D2*	0	-
rs10820564	49	0.0001	0.0003	0.0913	0.6019	9	98211452	INTRONIC	G	*ZNF367*	0	-
rs12355771	50	0.0001	0.0003	0.5071	0.112	10	75564410	INTRONIC	C	*AP3M1*	0	-
rs1805	51	0.0001	0.0001	0.8709	0.3613	11	117581279	INTRONIC	T	*AMICA1*	0	-
rs10492617	52	0.0001	0.0001	0.8512	0.2345	13	91039109	INTRONIC	A	*GPC5*	0	-
rs17115706	53	0.0001	0.0004	0.1369	0.9187	14	81228036	INTERGENIC	A	*AL160192.3*	30329	-
rs8032020	54	0.0001	3.63E-05	0.5487	0.7516	15	69362671	INTRONIC	N/A	*THSD4*	0	-
rs2285531	55	0.0001	0.0003	0.4516	0.0416	15	89543088	INTRONIC	G	*SV2B*	0	-
rs7197069	56	0.0001	0.0002	0.9998	0.008	16	17250425	INTRONIC	C	*XYLT1*	0	-
rs2254434	57	0.0001	0.0005	0.0199	0.9444	21	18257502	INTRONIC	C	*CHODL*	0	-
rs2824583	58	0.0001	0.0007	0.0178	0.9887	21	18262152	INTRONIC	G	*CHODL*	0	-
rs13056064	59	0.0001	0.0003	0.6738	0.0457	22	35669211	DOWNSTREAM	C	*CSF2RB*	2774	-
rs1279820	60	0.0001	0.0004	0.0614	0.9795	X	123210800	INTERGENIC	G	*RP13-158L7.3*	-31261	-
rs3121892	61	0.0002	0.0004	0.2153	0.9686	1	39478595	INTRONIC	G	*MACF1*	0	-
rs1187983	62	0.0002	2.03E-05	0.0519	0.8417	1	58214448	INTRONIC	T	*DAB1*	0	-
rs12717805	63	0.0002	0.0001	0.9173	0.5238	1	97145528	DOWNSTREAM	A	*AL592205.4*	4592	-
rs12403993	64	0.0002	0.0006	0.0257	0.7566	1	101446586	INTERGENIC	G	*AL109741.19-1*	20224	-
rs5008684	65	0.0002	0.0002	0.3207	0.0505	1	106846997	INTERGENIC	T	*RP11-233E12.4*	299903	-
rs16860185	66	0.0002	0.0009	0.0339	0.6229	1	160560694	INTRONIC	G	*NOS1AP*	0	-
rs12467276	67	0.0002	0.0004	0.4617	0.1745	2	98681499	INTRONIC	C	*MGAT4A*	0	-
rs12468086	68	0.0002	0.0004	0.4617	0.1745	2	98687006	INTRONIC	C	*MGAT4A*	0	-
rs6721186	69	0.0002	6.96E-05	0.3537	0.6843	2	131380573	INTRONIC	A	*ARHGEF4*	0	-
rs1449480	70	0.0002	0.0001	0.9834	0.3295	2	142552249	INTRONIC	G	*LRP1B*	0	-
rs7563853	71	0.0002	0.0002	0.8329	0.503	2	207558101	INTERGENIC	C	*CPO*	15658	-
rs4673070	72	0.0002	0.0002	0.8912	0.9259	2	224253062	INTERGENIC	A	*AP1S3*	72723	-
rs4679954	73	0.0002	4.44E-05	0.0732	0.1415	3	153135719	INTERGENIC	T	*SUCNR1*	53693	-
rs10513658	74	0.0002	0.0003	0.6286	0.3217	3	170272541	INTERGENIC	T	*EVI1*	11440	-
rs7632657	75	0.0002	9.34E-05	0.6077	0.4776	3	170386667	INTERGENIC	T	*EVI1*	-39613	-

rs10936625	76	0.0002	0.0006	0.1021	0.304	3	171577602	INTRONIC	T	*SKIL*	0	-
rs10025139	77	0.0002	0.0001	0.7495	0.2091	4	35989633	INTRONIC	T	*AC104078.3*	0	-
rs1036253	78	0.0002	8.85E-05	0.9501	0.1732	4	124987090	INTERGENIC	T	*AC108075.3*	-216013	-
rs432268	79	0.0002	0.001	0.0188	0.6844	5	36672573	INTRONIC	A	*SLC1A3*	0	-
rs11745510	80	0.0002	0.0005	0.1947	0.2188	5	64824137	INTERGENIC	A	*ADAMTS6*	-10677	-
rs9325132	81	0.0002	0.0006	0.3463	0.043	5	148528200	INTRONIC	C	*ABLIM3*	0	-
rs10053468	82	0.0002	0.0007	0.1906	0.0213	5	159339742	INTERGENIC	C	*ADRA1B*	7147	-
rs267174	83	0.0002	0.0005	0.1439	0.3473	6	7768578	INTRONIC	G	*BMP6*	0	-
rs1573134	84	0.0002	0.0004	0.3522	0.6319	6	16983447	INTERGENIC	T	*AL137003.12*	-113747	-
rs13193633	85	0.0002	0.0001	0.6128	0.2458	6	66679067	INTERGENIC	C	*AL391500.13*	-74399	-
rs10276168	86	0.0002	0.0001	0.7491	0.3618	7	53227825	INTERGENIC	N/A	*AC074397.7*	156076	-
rs2279648	87	0.0002	0.0003	0.9598	0.1282	10	75537199	INTRONIC	G	*VCL*	0	-
rs10824090	88	0.0002	0.0004	0.5404	0.1629	10	75593452	INTRONIC	T	*ADK*	0	-
rs7909915	89	0.0002	0.0003	0.7156	0.1442	10	75630948	INTRONIC	T	*ADK*	0	-
rs10824128	90	0.0002	0.0003	0.7339	0.1442	10	75677909	INTRONIC	G	*ADK*	0	-
rs10824151	91	0.0002	0.0005	0.2453	0.462	10	75762185	INTRONIC	A	*ADK*	0	-
rs10886161	92	0.0002	0.0003	0.6419	0.0662	10	119614983	INTERGENIC	T	*RAB11FIP2*	139436	-
rs10857712	93	0.0002	6.55E-05	0.7919	0.4692	10	135075656	INTRONIC	T	*MTG1*	0	-
rs9603439	94	0.0002	1.54E-05	0.1022	0.0299	13	38300920	INTRONIC	G	*FREM2*	0	-
rs4885488	95	0.0002	0.0002	0.9499	0.4265	13	77338577	INTERGENIC	A	*EDNRB*	29048	-
rs4884075	96	0.0002	0.0008	0.1093	0.3091	13	77439622	INTERGENIC	A	*AL139002.18-1*	-19714	-
rs2391352	97	0.0002	0.0005	0.2153	0.5361	13	106064529	INTERGENIC	N/A	*ARGLU1*	-46016	-
rs17114224	98	0.0002	7.02E-05	0.5653	0.5915	14	28285129	INTERGENIC	N/A	*FOXG1;FOXG1B*	-19672	-
rs2043120	99	0.0002	0.0012	0.1475	0.0097	14	81675555	INTERGENIC	A	*AL357095.4*	-290413	-
rs12432601	100	0.0002	0.0008	0.1764	0.0161	14	81709202	INTERGENIC	C	*AL357095.4*	-256766	-

* P values were combined using the Stouffer's weight Z-method [2].

Table S8. Top SNPs revealed in the GWAS for baseline HCV viral load

SNP	Rank	Combined* P	P (Caucasians)	P (AAs)	P(Hispanics)	Chr.	Coordinate	Type	Ancestral allele	Closest gene	Distance to gene	Distance to exon boundary
rs12979860	1	1.21E-10	3.48E-08	8.46E-05	0.0184	19	44430627	UPSTREAM	T	*IL28B*	-3018	-
rs8099917	2	2.68E-09	2.51E-08	0.0983	0.041	19	44435005	DOWNSTREAM	T	*IL28B /AC011445.6*	2391	-
rs8109886	3	1.97E-08	5.44E-07	0.003	0.3429	19	44434602	DOWNSTREAM	A	*IL28B /AC011445.6*	2794	-
rs12980275	4	6.25E-08	3.05E-06	0.0021	0.0208	19	44423623	DOWNSTREAM	G	*IL28B*	2410	-
rs12972991	5	1.10E-06	6.58E-06	0.1614	0.0256	19	44423587	DOWNSTREAM	A	*IL28B*	2446	-
rs2032182	6	1.50E-06	8.49E-07	0.9682	0.7538	18	28649371	INTERGENIC	G	*KLHL14*	-42399	-
rs13044588	7	1.92E-06	2.07E-06	0.795	0.2303	20	47513437	INTRONIC	T	*KCNB1*	0	-
rs1871437	8	5.39E-06	1.97E-05	0.4717	0.0028	15	90536919	INTERGENIC	C	*SLCO3A1*	29136	-

rs12980602	9	7.21E-06	7.24E-05	0.0336	0.0609	19	44444660	INTERGENIC	C	*IL28A*	-6337	-
rs345800	10	8.14E-06	8.21E-06	0.5272	0.9064	15	31169949	INTERGENIC	A	*FMN1*	-22572	-
rs4803223	11	9.42E-06	7.05E-05	0.0726	0.0433	19	44438059	UPSTREAM	A	*AC011445.6*	-310	-
rs1924989	12	1.66E-05	6.90E-06	0.8096	0.9758	6	49438764	UPSTREAM	C	*RP1-142O9.1*	-2712	-
rs12656675	13	1.86E-05	1.45E-05	0.713	0.9074	5	1.05E+08	INTERGENIC	A	*AC106777.2*	-146207	-
rs10186029	14	1.92E-05	1.00E-05	0.7042	0.3501	2	2.14E+08	INTRONIC	N/A	*IKZF2*	0	-
rs811526	15	1.93E-05	3.31E-05	0.8902	0.0083	2	79080607	INTERGENIC	A	*REG3G*	-25727	-
rs1916642	16	2.28E-05	6.02E-06	0.1432	0.0279	5	72488303	INTERGENIC	C	*TMEM174*	-16501	-
rs7501817	17	2.47E-05	1.02E-05	0.9612	0.5119	17	29253650	INTRONIC	G	*ACCN1*	0	-
rs6879012	18	2.50E-05	6.03E-06	0.1708	0.0921	5	72498637	INTERGENIC	G	*TMEM174*	-6167	-
rs7811230	19	2.72E-05	2.39E-05	0.846	0.5202	7	1.1E+08	INTERGENIC	G	*IMMP2L*	158695	-
rs1349964	20	2.74E-05	2.06E-05	0.693	0.7285	15	31140281	INTRONIC	T	*FMN1*	0	-
rs9618567	21	2.79E-05	2.62E-05	0.3604	0.2286	22	17793706	INTRONIC	C	*HIRA*	0	-
rs2058742	22	2.81E-05	3.25E-05	0.3837	0.6035	17	67552142	INTERGENIC	T	*AC007461.8*	5167	-
rs331645	23	3.06E-05	1.49E-05	0.7395	0.1823	1	59409975	INTERGENIC	G	*FGGY*	-125353	-
rs2057368	24	3.17E-05	3.99E-05	0.4374	0.9812	14	54373759	DOWNSTREAM	G	*GCH1*	4717	-
rs13063052	25	3.25E-05	3.78E-05	0.3572	0.5199	3	1.49E+08	INTERGENIC	T	*ZIC1*	205270	-
rs7567984	26	3.54E-05	4.04E-05	0.7744	0.0073	2	2.28E+08	INTRONIC	C	*SLC19A3*	0	-
rs1871946	27	3.62E-05	5.94E-06	0.3062	0.7162	2	2.14E+08	INTRONIC	T	*IKZF2*	0	-
rs11858145	28	3.71E-05	1.76E-05	0.9563	0.7285	15	31146866	NON_SYNONYMOUS_CODING	A	*FMN1*	0	-
rs807459	29	3.81E-05	4.04E-05	0.4313	0.5897	22	17603352	NON_SYNONYMOUS_CODING	T	*CLTCL1*	0	40
rs1633399	30	3.91E-05	4.04E-05	0.4313	0.5406	22	17563787	NON_SYNONYMOUS_CODING	A	*CLTCL1*	0	-10
rs807547	31	3.91E-05	4.04E-05	0.4313	0.5406	22	17575680	NON_SYNONYMOUS_CODING	T	*CLTCL1*	0	-16
rs1474664	32	4.02E-05	8.24E-05	0.3571	0.3501	6	1.39E+08	UPSTREAM	C	*KIAA1244*	-1633	-
rs712952	33	4.04E-05	4.04E-05	0.4313	0.4762	22	17577949	NON_SYNONYMOUS_CODING	G	*CLTCL1*	0	70
rs13164411	34	4.45E-05	4.67E-05	0.4301	0.5564	5	1.05E+08	INTERGENIC	C	*AC106777.2*	-118829	-
rs9941446	35	4.50E-05	3.08E-05	0.6619	0.4356	18	63632033	INTERGENIC	G	*C18orf4;DSEL*	-296836	-
rs12005744	36	4.70E-05	6.97E-05	0.5475	0.3594	9	1.04E+08	INTERGENIC	G	*AL391867.5*	19654	-
rs870992	37	4.70E-05	5.48E-05	0.7507	0.3051	5	52228994	INTRONIC	A	*ITGA1*	0	-49
rs6440401	38	4.70E-05	1.95E-06	0.0417	0.3151	3	1.47E+08	INTERGENIC	G	*AC055758.23*	8641	-
rs1913293	39	4.89E-05	2.23E-06	0.0495	0.3151	3	1.47E+08	INTERGENIC	A	*AC055758.23*	11208	-
rs8084688	40	5.01E-05	4.23E-05	0.9606	0.3784	18	28647852	INTERGENIC	C	*KLHL14*	-40880	-
rs8069265	41	5.09E-05	0.0001	0.1686	0.6302	17	76774020	DOWNSTREAM	T	*AZI1*	3968	-
rs1156404	42	5.37E-05	4.11E-05	0.7267	0.8073	6	70797027	INTRONIC	C	*COL19A1*	0	-
rs10491846	43	5.66E-05	0.0002	0.1175	0.6775	9	88352001	INTERGENIC	C	*AL158828.14*	-124138	-
rs7207863	44	5.96E-05	0.0004	0.0305	0.4724	17	54892557	INTERGENIC	G	*YPEL2*	58680	-
rs13162459	45	6.13E-05	5.58E-05	0.6885	0.939	5	1.05E+08	INTERGENIC	G	*AC106777.2*	-145288	-
rs474103	46	6.42E-05	2.94E-05	0.8112	0.1705	1	59392159	INTERGENIC	T	*FGGY*	-143169	-
rs1918245	47	6.56E-05	0.0002	0.1945	0.5504	X	4724367	INTERGENIC	G	*AC074035.22*	-77412	-
rs10049503	48	6.87E-05	0.0004	0.0284	0.6503	4	6667849	INTRONIC	C	*MAN2B2*	0	-

rs6717546	49	7.14E-05	8.73E-05	0.321	0.4513	2	2.34E+08	UPSTREAM	A	*UGT1A9;UGT1A8; UGT1A7;UGT1A6; UGT1A5;UGT1A4; UGT1A10;UGT1A3; UGT1A1*	170	-
rs13440284	50	7.17E-05	0.0001	0.4142	0.4045	9	1.04E+08	DOWNSTREAM	C	*AL391867.5*	4580	-
rs1360400	51	7.32E-05	0.0001	0.5835	0.0653	20	22909890	INTERGENIC	G	*RP11-189G24.2*	-15190	-
rs4148329	52	7.49E-05	6.34E-05	0.6128	0.6564	2	2.34E+08	UPSTREAM	C	*UGT1A9;UGT1A8; UGT1A7;UGT1A6; UGT1A5;UGT1A4; UGT1A10;UGT1A3; UGT1A1*	113	-
rs4906434	53	7.94E-05	0.0001	0.2804	0.4575	14	1.04E+08	INTERGENIC	T	*AL512357.4*	102426	-
rs10270301	54	8.60E-05	0.0001	0.3469	0.7609	7	1.1E+08	INTERGENIC	C	*IMMP2L*	169552	-
rs2645444	55	8.77E-05	0.0002	0.3338	0.475	8	11688962	INTERGENIC	T	*NEIL2*	6699	-
rs6799767	56	8.80E-05	0.0001	0.2877	0.7872	3	1.78E+08	INTERGENIC	T	*AC092920.7*	-450921	-
rs8095514	57	9.02E-05	0.0005	0.0526	0.266	18	8724995	INTRONIC	C	*KIAA0802*	0	-
rs4290724	58	9.14E-05	0.0003	0.2159	0.0587	20	22914362	INTERGENIC	A	*RP11-189G24.2*	-19662	-
rs498612	59	9.31E-05	0.0001	0.4328	0.9023	11	72585334	INTERGENIC	A	*P2RY2*	-21658	-
rs7758604	60	9.56E-05	0.0002	0.2745	0.5925	6	91999464	INTERGENIC	C	*AL132766.13-1*	-494011	-
rs4667375	61	0.0001	7.04E-05	0.8633	0.2254	2	1.51E+08	INTERGENIC	G	*AC093738.2*	-189299	-
rs6797570	62	0.0001	1.88E-05	0.1454	0.8903	3	63647792	INTERGENIC	C	*SNTN*	21874	-
rs746977	63	0.0001	0.0003	0.1181	0.712	4	31085425	INTERGENIC	G	*PCDH7*	331856	-
rs4454039	64	0.0001	0.0003	0.0751	0.6424	5	72445143	INTERGENIC	G	*TMEM171*	-7015	-
rs12517408	65	0.0001	5.58E-05	0.7896	0.9909	5	1.05E+08	INTERGENIC	G	*AC106777.2*	-154289	-
rs4711926	66	0.0001	3.79E-05	0.6338	0.6129	6	49333763	INTERGENIC	T	*RP1-142O9.2*	-15437	-
rs4235847	67	0.0001	0.0001	0.7427	0.6488	6	92024902	INTERGENIC	C	*AL132766.13-1*	-468573	-
rs4526163	68	0.0001	8.59E-05	0.8954	0.6444	6	92029093	INTERGENIC	T	*AL132766.13-1*	-464382	-
rs6454848	69	0.0001	0.0001	0.7427	0.6488	6	92033983	INTERGENIC	A	*AL132766.13-1*	-459492	-
rs4473838	70	0.0001	0.0001	0.7427	0.6488	6	92034376	INTERGENIC	A	*AL132766.13-1*	-459099	-
rs6961681	71	0.0001	0.0007	0.2334	0.0004	7	35970941	INTERGENIC	N/A	*AC087072.2*	-34542	-
rs9791486	72	0.0001	0.0014	0.0028	0.759	7	1.09E+08	INTERGENIC	G	*N/A*	-9	-
rs7022749	73	0.0001	0.001	0.0063	0.8336	9	89595315	INTRONIC	C	*AL160279.21-1*	0	-
rs7317197	74	0.0001	0.0006	0.0126	0.9096	13	1.09E+08	INTERGENIC	C	*IRS2*	199541	-
rs17677211	75	0.0001	7.08E-05	0.6945	0.4949	16	7534065	INTRONIC	A	*A2BP1*	0	-
rs9675424	76	0.0001	9.43E-05	0.9353	0.3784	18	28642306	INTERGENIC	C	*KLHL14*	-35334	-
rs11588152	77	0.0002	0.0003	0.3182	0.7066	1	58618271	INTERGENIC	N/A	*OMA1*	93862	-
rs854286	78	0.0002	0.0003	0.7578	0.1325	1	68956802	INTERGENIC	G	*RP4-694A7.3*	-203862	-
rs9988609	79	0.0002	0.0005	0.3133	0.1654	1	69087816	INTERGENIC	A	*RP4-694A7.3*	-334876	-
rs17489732	80	0.0002	0.001	0.0135	0.5737	2	36737588	INTERGENIC	C	*VIT*	-39823	-
rs7561980	81	0.0002	0.0002	0.4492	0.444	2	56439278	INTRONIC	G	*CCDC85A*	0	-
rs671772	82	0.0002	0.0003	0.7314	0.4256	2	80050973	INTRONIC	T	*CTNNA2*	0	-
rs749008	83	0.0002	6.66E-05	0.7139	0.2614	2	1.28E+08	INTRONIC	G	*BIN1*	0	-
rs13007371	84	0.0002	0.0005	0.0561	0.4827	2	1.42E+08	INTRONIC	A	*LRP1B*	0	-
rs1688416	85	0.0002	0.0001	0.4071	0.0544	3	4377649	INTRONIC	C	*SUMF1*	0	-181

rs602763	86	0.0002	0.0002	0.7046	0.5783	3	1.11E+08	INTERGENIC	A	*AC068781.19*	-25772	-
rs523993	87	0.0002	8.37E-06	0.0216	0.4025	3	1.47E+08	UPSTREAM	C	*AC055758.23*	-757	-
rs655679	88	0.0002	9.48E-06	0.0247	0.4025	3	1.47E+08	DOWNSTREAM	C	*AC055758.23*	3262	-
rs1024758	89	0.0002	0.0009	0.046	0.4728	4	15036596	INTRONIC	G	*C1QTNF7*	0	-
rs722235	90	0.0002	0.0005	0.1376	0.3189	4	15069432	INTERGENIC	C	*C1QTNF7*	12545	-
rs1509241	91	0.0002	0.0011	0.0832	0.1623	4	22973591	INTERGENIC	T	*PPARGC1A*	429151	-
rs1605813	92	0.0002	0.0017	0.0092	0.1073	4	23039866	INTERGENIC	A	*PPARGC1A*	362876	-
rs3934117	93	0.0002	0.0002	0.6329	0.8315	4	31582018	INTERGENIC	G	*N/A*	-9	-
rs6553978	94	0.0002	0.0002	0.7505	0.8315	4	31600519	INTERGENIC	C	*N/A*	-9	-
rs926163	95	0.0002	0.0002	0.714	0.7603	4	1.04E+08	INTERGENIC	C	*NFKB1*	-57733	-
rs1567193	96	0.0002	0.0004	0.2015	0.9188	4	1.31E+08	INTERGENIC	T	*AC079857.5-2*	-180608	-
rs2884832	97	0.0002	0.0008	0.0328	0.4941	4	1.51E+08	DOWNSTREAM	G	*AC096756.3-2*	4906	-
rs1054724	98	0.0002	4.61E-05	0.2626	0.4183	5	1.12E+08	3PRIME_UTR	C	*EPB41L4A*	0	-
rs10477734	99	0.0002	0.0004	0.2229	0.4673	5	1.31E+08	INTRONIC	A	*RAPGEF6*	0	-
rs2496363	100	0.0002	8.61E-05	0.8583	0.6524	6	49292729	INTERGENIC	C	*RP1-142O9.2*	-56471	-

* P values were combined using the Stouffer's weight Z-method [2].

XV. Results of the CNV analysis

To detect copy number variants that associate with HCV viral load or SVR, we evaluated the frequency of deletions (copy number <2) and duplications (copy number=>2) separately in subjects with high and low viral load and patients with and without a sustained viral response. CNV frequencies were assessed by evaluating the number of copies at each unique start and stop site of the CNVs that met all the quality control measures defined. For this analysis, viral load was dichotomized into high and low viral load with the divide defined by the combined mean. Each site was assessed for a difference in frequency in subjects of European ancestry using the permutation-based Fisher's exact test in plink [6]. This assessment resulted in no genome-wide significant associations for copy number variation.

We also assessed the impact of CNVs on HCV viral load and SVR using a set of SNPs validated as tags for 285 common copy number variants in samples of European ancestry (Steve McCarroll and David Altshuler, personal communication)[16]. A total of 185 of these were either present on the Illumina 610-Quad chip directly or tagged by an associated variant with an r^2 greater than 0.8. Linear regression incorporating EIGENSTRAT covariates to correct for population stratification within the subjects of European ancestry was performed using the PLINK genome-wide association analysis toolkit as described previously[17]. None of the tagging SNPs were significantly associated with HCV viral load or SVR after a Bonferroni correction for multiple testing.

XVI. Evaluating the possibility that rs12979860 associates with natural clearance of hepatitis C

We also tested whether the polymorphism associates with natural clearance of hepatitis C by comparing allele frequencies in the chronically infected patients included in this study to a random population sample with unknown hepatitis C status (supplementary information, part II). If the polymorphism influences natural clearance we would expect a frequency difference in this comparison, since all individuals who naturally clear the virus will be excluded from the chronic infection cohort, thereby reducing the frequency of the allele that increases the likelihood of natural clearance. We found that the frequency of the C allele was significantly reduced in the chronically infected cohort, with a frequency of 0.63 in individuals of European ancestry in the HCV cohort compared with a frequency of 0.73 in the ethnically matched controls that were corrected for any cryptic stratification ($P=2.48 \times 10^{-6}$), indicating that individuals with the C allele are preferentially excluded from the HCV cohort. This comparison shows that the rs12979860 C allele, associated with better response to treatment, is also potentially associated with a greater likelihood of natural clearance of hepatitis C. We note, however, that the magnitude of the effect is difficult to estimate absent a more direct comparison between a cohort known to naturally clear the virus and a matched chronically infected one.

XVII. Efforts to identify the causal variant or variants responsible for these associations

We tested for association between rs12979860 and *IL28B* RNA expression in peripheral blood mononuclear cells from 80 individuals (HCV-negative population controls) in the SNPExpress database [18] which relates a genome-wide set of polymorphisms to genome-wide expression patterns [18]. We found no correlation with expression levels of *IL28B* with the best proxy for rs12979860 available in this database (rs12980275, r^2=0.88 with rs12979860), although the expression levels in the absence of infection were low.

XVIII. The generalization of the observations from multiple ethnic populations in this study

In this study we cannot as yet estimate the contribution of rs12979860 to average differences amongst populations globally, since we do not for example have estimates of the effects of the polymorphism in East Asian patients. This will need to be prospectively tested.

We can make the following observations – a) the effect of the polymorphism is statistically consistent in three groups studied: Hispanics, individuals of European ancestry, and African Americans. Moreover, b) there is a striking correlation between estimated population average rates of SVR and the allele frequency in all groups including East Asians. Based on these facts, we hypothesize that the *IL28B* polymorphism is likely to contribute substantively to global variation in SVR rates.

References

1. Ghany, M. G., Strader, D. B., Thomas, D. L. & Seeff, L. B. American Association for the Study of Liver Diseases: Practice Guidelines; Diagnosis, Management, and Treatment of Hepatitis C: An Update. (2009).
2. Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol* **18**, 1368-73 (2005).
3. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. The French METAVIR Cooperative Study Group. *Hepatology* **20**, 15-20 (1994).
4. McHutchison, J. G. et al. Peginterferon alfa-2b or peginterferon alfa-2a in combination with ribavirin for the treatment of chronic hepatitis C genotype 1 infection. *N Engl J Med.* In press (2009).
5. Muir, A. J., Bornstein, J. D. & Killenberg, P. G. Peginterferon alfa-2b and ribavirin for the treatment of chronic hepatitis C in blacks and non-Hispanic whites. *N Engl J Med* **350**, 2265-71 (2004).
6. Purcell, S. et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet.* **81** (2007).
7. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
8. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).

9. Shiffman, M. L. et al. Peginterferon alfa-2a and ribavirin in patients with chronic hepatitis C who have failed prior treatment. *Gastroenterology* **126**, 1015-23; discussion 947 (2004).

10. Pillai, S. G. et al. A Genome-Wide Association Study in Chronic Obstructive Pulmonary Disease (COPD): Identification of Two Major Susceptibility Loci. *PLoS Genet.* **5**, e1000421. doi:10.1371/journal.pgen.1000421 (2009).

11. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381 - 385 (2008).

12. Asselah, T. et al. Genetics, genomics, and proteomics: implications for the diagnosis and the treatment of chronic hepatitis C. *Semin Liver Dis* **27**, 13-27 (2007).

13. Walters, K. A. & Katze, M. G. Using high-throughput genomics to study hepatitis C: what determines the outcome of infection? *Antiviral Res* **81**, 198-208 (2009).

14. Meyers, L. S., Gamst, G. & Guarino, A. J. Applied multivariate research: design and interpretation (SAGE Publications, 2006).

15. Sarasin-Filipowicz, M. et al. Interferon signaling and treatment outcome in chronic hepatitis C. *Proc Natl Acad Sci U S A.* **105**, 7034-9 (2008).

16. McCarroll, S. A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74 (2008).

17. Fellay, J. et al. A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944-7 (2007).

18. Heinzen, E. L. et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol* **6**, e1 (2008).