Published in final edited form as:

J Viral Hepat. 2019 April ; 26(4): 476–484. doi:10.1111/jvh.13051.

Genomic variability of within-host hepatitis C variants in acute infection

DR. Chaturaka Rodrigo¹, Preston Leung², Andrew R. Lloyd², Rowena A. Bull^{1,2}, Fabio Luciani^{1,2}, DR. Jason Grebely², Gregory J. Dore², Tanya Applegate², Kimberly Page³, Julie Bruneau⁴, Andrea L. Cox⁵, William Osburn⁵, Arthur Y. Kim⁶, Naglaa H. Shoukry⁴, Georg M. Lauer⁶, Lisa Maher², Janke Schinkel^{7,8}, Maria Prins^{7,8}, PROF. Margaret Hellard^{9,10,11,12}, and Auda A. Eltahla^{1,3} on behalf of the InC Collaborative

¹School of Medical Sciences, UNSW, NSW, Australia ²The Kirby Institute, UNSW Sydney, NSW, Australia ³University of New Mexico, Albuquerque, New Mexico, USA ⁴Centre de Recherche du Centre hospitalier de l'Université de Montréal (CRCHUM), Montreal, QC, Canada ⁵Johns Hopkins Medical Institutions, Baltimore, MD, USA ⁶Harvard Medical School, Boston, MA, USA ⁷Academic Medical Center, Amsterdam, The Netherlands ⁸GGD Public Health Service of Amsterdam ⁹Burnet Institute, Melbourne, VIC, Australia ¹⁰Monash University, Australia ¹¹Alfred Hospital, Melbourne, Australia ¹²Doherty Institute and Melbourne School of Population and Global Health, University of Melbourne

Abstract

Interactions between the host immune system and the viral variants determine persistence of hepatitis C virus (HCV) infection after the acute phase of infection. This study describes the genetic variability of within-host HCV viral variants in acute infection and correlates it with host and virus-related traits and infection outcome. Next generation sequence data (Illumina, MiSeq platform) of viral genomes from 116 incident acute infections (within 180 days of infection) were analysed to determine all the single nucleotide polymorphism (SNP) frequencies above a threshold of 0.1%. The variability of the SNPs for the full open reading frame of the genome as well as for each protein coding region were compared using mean standardised Shannon Entropy (SEs) values calculated separately for synonymous and non-synonymous mutations. The envelope glycoproteins regions (E1 and E2) had the highest SE values (indicating greater variability) followed by the NS5B region. Non-synonymous mutations rather than synonymous mutations were the main contributors to genomic variability in acute infection. The mean difference of Shannon Entropy was also compared between subjects after categorizing the samples according to host and virus-related traits. Host *IFNL3* allele CC polymorphism at rs12979860 (vs. others) and viral genotype 1a (vs. 3a) were associated with higher genomic variability across the viral open reading frame. Time since infection, host gender or continent of origin were not associated with the viral genomic variability. Viral genomic variability did not predict spontaneous clearance.

Correspondence: Chaturaka Rodrigo MD PhD, Department of Pathology, School of Medical Sciences, UNSW Sydney, 2052, Sydney, NSW, Australia, c.rodrido@unsw.edu.au, Tel: +61 2 9385 8365.

Conflicts of interest: Authors have no conflicts of interest to declare regarding the content of this paper

Keywords

Hepatitis C virus; Shannon Entropy; IFNL3; Spontaneous clearance; InC3 study

Introduction

The highly error prone replicase of RNA viruses, results in dynamic populations of viral quasi-species within each infected host ¹⁻³. Advances in next generation sequencing (NGS) over the last decade have provided better understanding of host-pathogen relationships and within host viral variants⁴. NGS platforms now allow characterisation of the diversity of viral variants in a high throughput and cost-effective manner, with capacity to differentiate individual mutations with a frequency of occurrence as low as 0.1% ⁵.

HCV is a single stranded RNA virus with a genome length of 9.3kb⁶, which is translated as a single polyprotein encoding three structural proteins (Core, E1 and E2) and seven non-structural proteins (P7, NS2, NS3, NS4A, NS4B, NS5A, NS5B). The RNA dependent RNA polymerase (NS5B) has an estimated 2.5×10^{-5} mutations per nucleotide per genome replication, and the virion half-life is less than 3 hours ^{7,8}. The vast diversity generated by these high replication and mutation rates are proposed to be an adaptive mechanism employed by the virus to survive selection pressures, such as host immunity. It has been observed that in acute HCV infection, viral diversity undergoes two bottleneck events; one at the initiation of infection and another approximately 100 days post-infection associated with the peak of humoral and cellular antiviral immune responses ⁹. Following this second bottleneck, approximately 25% of the infected individuals spontaneously clear the virus ¹⁰. One of the largest datasets on HCV available to-date, pooled from nine prospective cohort studies on intravenous drug users at risk of HCV (InC3 study), suggested an association between gender, polymorphisms of the host interferon lambda 3 allele (IFNL3) and viral genotype with spontaneous clearance¹⁰. IFNL3, previously known as IL-28B is a cytokine of the type III interferon family that induces interferon stimulated genes (ISGs) and is important for protection against viral infections ¹¹. Multiple other studies have independently found associations between spontaneous clearance and different polymorphisms of this host gene ^{10,12,13}.

Viral genomic diversity during acute infection has not yet been studied in relation to spontaneous clearance. Better understanding of the pathogen and host related factors that influence viral genomic diversity in acute infection may be useful in designing preventive vaccines that mimic or enhance the natural process of spontaneous clearance ¹⁴.

Previous studies utilizing NGS to examine viral variant diversity have been limited to subsegments of the genome - probably due to the technical challenges in generating full length HCV amplicons in a high throughput and cost-effective manner¹⁵. Also, obtaining viraemic samples from acute infection time points is logistically difficult as the infection is largely asymptomatic. Finally, most studies to-date have not further characterized the observed genomic diversity as synonymous and non-synonymous mutations. This is important as the observed diversity may have a more functional significance if it changes the amino acid composition of the translated protein (non-synonymous mutations). However,

synonymous mutations are also known to affect the RNA structure and hence have a functional importance related to viral replication and survival.

This paper describes the genetic diversity observed across within-host viral variants in acute infection by quantifying the variability of single nucleotide polymorphisms (SNPs) across the near full-length, viral open reading frame (ORF). The observed variation was further characterised as synonymous and non-synonymous mutations. In addition, associations of viral genomic variability with host and virus-related traits as well as the infection outcome (persistent infection or spontaneous clearance) were also examined.

Methods

Sample selection

All samples for this study were acquired from an international collaboration of nine prospective cohorts (Australia – 4, USA – 3, Canada – 1 and the Netherlands – 1) of injecting drug users; the International Collaborative of Incident HIV and HCV in Injecting Cohorts (InC³ study). The InC³ cohort profiles¹⁶, the process of full length genome amplicon generation¹⁷ and sequencing¹⁸ has been described previously. In brief, these nine cohorts follow up people who inject drugs (PWID) prospectively to evaluate their risk of subsequent HCV and HIV infection. Seven of these cohorts identify acute HCV infections by follow up of high risk, HCV antibody negative PWIDs prospectively while the other two cohorts identify such cases via clinical referral networks. Two cohorts recruit participants from prisons in Australia and USA while others recruit from the community. The individual cohorts of the collaboration collect and share demographic, behavioural, virological and host genomic data with each other for collaborative studies. The incident case samples with an estimated date of infection less than 180 days from these cohorts were used to generate full-length HCV amplicons followed by next generation sequencing (Paired end sequencing, illumina, MiSeq platform). The estimated date of infection for each subject was calculated as published by Hajarizadeh et al. previously¹⁹. Of the full-length genomes sequenced, all samples with a coverage of at least 1000 nucleotides per position across the ORF were selected for this analysis. Each sample originated from a single patient.

Detection of single nucleotide polymorphisms

For every sample, a consensus sequence (a single representative sequence of all within-host viral variants) was generated using a bioinformatics pipeline described previously^{17,18}. The consensus sequence was annotated for the coding cDNA sequence (CDS) using previously annotated reference genomes for each genotype¹⁷. The Illumina short reads were imported into Geneious (version 8, Biomatters Inc, Auckland, New Zealand) software suite as FASTq files. After quality control and trimming to remove adapter sequences, the short reads were aligned against the annotated autologous consensus sequence using the Geneious aligner (settings: medium sensitivity/fast). The SNP frequency per position was determined with a minimum frequency variation of 0.1%. The calculations were limited to the protein coding region of the genome sequence. The SNP frequency of the final output was exported as a text file for further processing.

Within-host variability of SNPs

The variability of SNPs per genome position was quantified using Shannon entropy²⁰. It is a measure of uncertainty in a dataset. When applied to a sequence alignment, Shannon entropy per position depends on the number of possible alternatives (4 in nucleotide alignments; 20 in amino acid alignments) as well as the observed frequencies of these alternatives. A higher SE at a sequence position indicates more variability in terms of SNPs observed in that position. A SE of zero indicates an invariant column of nucleotides / amino acid residues for all variants. For each sample, SEs were calculated for each nucleotide position across the ORF using the output of the Geneious software as described in the previous section. A summary statistic termed the mean standardised SE was then derived by calculating the mean SE across a pre-defined length of the ORF and standardizing it according to the length of the selected segment. These mean standardized SE values were calculated for the near full-length ORF as well as for each viral protein coding region (Core, E1, E2, P7, NS2, NS3, NS4A, NS4B, partial NS5B) separately²¹. In the next step, these values were further split as SEs for synonymous (S) and non-synonymous (NS) mutations by arranging nucleotide triplets into codons in the reading frame. All SEs were calculated using customized bioinformatic scripts.

Statistical analysis

The combined, synonymous and non-synonymous Shannon entropy values were compared across each protein coding region by one-way ANOVA with post-hoc testing (Tukey's HSD test). The mean differences of combined, synonymous and non-synonymous Shannon entropies for near full-length ORF and individual protein coding regions were also compared after grouping subjects according to host and virus-related factors (independent sample T tests). The independent variables for this analysis were selected *a priori*, based on previous research^{10,18}, which had determined predictors of clearance, and phylogenetic variability of consensus (between-host) sequences. These variables were: estimated days since infection, viral genotype, host gender, host IFNL3 (rs12979860) allele polymorphism and continent of origin. The factors that were shown to be significant in unadjusted analysis were included in multiple linear regression with Shannon entropy as the dependent variable. Finally, to assess the impact of genomic variability on clearance, a logistic regression was carried out using two models. The first model used gender, genotype and host IFNL3 polymorphism to predict clearance following Grebely et al¹⁰, while the second model added Shannon entropy to the previous model. The second model was run separately using Shannon entropy calculated for the ORF and each protein coding region further subdivided as combined, synonymous and non-synonymous Shannon entropy. Furthermore, to avoid co-linearity, either viral genotype or host IFNL3 polymorphism were removed from some analyses if these factors were shown to be associated with the Shannon entropy in multiple regression analyses. Statistical significance was set at 0.05 and all analysis were carried out using SPSS statistical software package (Version 21, IBM, USA).

Ethical approval

Ethics approval for this study was granted by the University of New South Wales, Australia (No. 14201). In addition, each of the individual cohorts had ethical approval for data and specimen collections.

Results

The InC3 viral sequence repository had 190 full length HCV genomes sequenced by next generation sequencing. All these samples were from within 180 days since onset of infection. Of these, 116 samples met the inclusion criteria of having a coverage of at least 1000 nucleotides per position across the ORF. The included samples were predominantly from genotypes 1a (59, 51%) and 3a (37, 32%) and were a mix of Australian (35, 30%) and North American sequences (81, 70%). A significant proportion of samples (42%) were within 28 days since infection. The infection outcome was known for a total of 94 subjects, and of these 23 (20%) spontaneously cleared the infection (Table 1).

Variation of Shannon entropy across the HCV open reading frame

The mean standardized Shannon entropy values for the near full-length ORF and each individual protein coding region are summarized in supplementary table 1. Regarding the combined Shannon entropy (both synonymous and non-synonymous mutations), as expected, the E1 and E2 regions had the highest values followed by the NS5B region. The most conserved region was NS4A followed by Core (figure 1). Excluding the short P7 region, the only statistically significant difference in Shannon entropy between protein coding regions were between E1 and NS4A. Interestingly, a larger proportion (mean 63.9%, range 60.6–68.3%) of the total Shannon entropy of the ORF and individual protein coding regions were contributed by non-synonymous mutations rather than by synonymous mutations (figure 1). For non-synonymous mutations, the Shannon entropy of E1 region was significantly higher than that of core, NS3, NS4A, NS4B and NS5A regions (supplementary table 1). The Shannon entropy of synonymous mutations were less variable when compared across each protein coding region with no statistically significant differences between protein coding regions (figure 1, supplementary table 1).

Associations for Shannon entropy variation across the HCV open reading frame

In unadjusted analysis, both IFNL3 host polymorphism at rs12979860 (CC vs. non-CC) and viral genotype (1a vs. 3a) were found to be significant predictors of total Shannon entropy across the full ORF and multiple protein coding regions. For all significant associations, IFNL3 allele CC (vs. non-CC) at rs12979860 and viral genotype 1a (vs. 3a) were associated with a higher total Shannon entropy ($p < 0.05$). The following associations remained after a linear multiple regression analysis; For host IFNL3 CC genotype, significantly higher total Shannon entropy values were observed for the near full-length ORF, plus E1, E2, NS3, NS4A and NS4B protein coding regions. For viral genotype 1a, significantly higher total Shannon entropy values were observed for the near full-length ORF plus core, NS2, NS4B and partial NS5B protein coding regions. Time since infection, gender, infection outcome or continent of origin were not associated with genetic variability.

When the total Shannon entropy values were split as synonymous and non-synonymous two clear patterns emerged. Host IFNL3 polymorphism only influenced the synonymous Shannon entropy where significantly higher values were observed for the CC allele (vs. others) across the near full-length ORF and in core, E2, NS2, NS3, NS4B and NS5A protein coding segments. For non-synonymous mutations a similar relationship was not observed in the ORF or within individual protein coding regions (table 2 and supplementary table 3).

Conversely, the viral genotype mostly influenced the non-synonymous Shannon entropy where significantly higher values were observed for subtype 1a (vs. subtype 3a) across the near-full length ORF and in core, E1, E2, NS2, NS3, NS4B, NS5A and NS5B protein coding regions. For synonymous mutations a similar relationship was only observed within the NS5B region (table 2 and supplementary table 2).

Shannon entropy and spontaneous clearance of HCV infection

When viral genotype, host gender and IFNL3 polymorphism were combined in a logistic regression to predict spontaneous clearance, only host IFNL3 polymorphism emerged as a significant predictor. In the second model where Shannon entropy (synonymous or non-synonymous, for different regions of the genome) was considered with either host IFNL3 polymorphism or viral genotype (avoiding co-linearity), again, only host IFNL3 polymorphism emerged as a strong predictor of clearance (when it was included in the model, supplementary table 4).

Discussion

This genomic analysis of 116 HCV samples revealed that non-synonymous mutations were the major contributor to the observed viral genome variability in the ORF during acute infection. Viral genotype and host IFNL3 polymorphism at rs12979860 were both predictors of viral genome variability along the near full-length ORF as well as for some individual protein coding regions. Viral genotype 1a (vs. 3a) and host IFNL3 CC allele at rs12979860 were associated with a higher Shannon entropy. Moreover, viral genotype was mostly predictive of non-synonymous genomic variability while IFNL3 polymorphism was predictive of synonymous genomic variability. Host IFNL3 polymorphism at rs12979860, but not viral genomic variability, was predictive of spontaneous clearance.

The observation that non-synonymous mutations are the major driver of genomic variation during acute HCV infection is interesting as it likely reflects the process of adaptation of the virus to the new host which has a different profile of antiviral immunity compared to the previous host. In HCV and in many other rapidly mutating RNA virus infections (e.g. dengue and influenza), the incoming virus population in a new transmission event, rapidly undergoes a bottleneck event from which only a few transmitted founder variants emerge to establish the infection within the new host^{1,14,22–24}. The survival of these variants may be determined by the functionally useful non-synonymous mutations. As the immune system of the new host recognizes the infection, more escape mutations will occur within viral variants. Some of these will be transient if they impair the replication capacity of the virus²⁵. This may explain why within-host HCV viral variants undergo a second genetic bottleneck at approximately 100 days post-infection, mainly driven by the selection pressures applied

by the host immune system⁹. Non-synonymous mutations in the acute infection samples studied here were pronounced in the structural envelope proteins (E1 and E2) that directly interact with host hepatocytes for viral entry. It has been shown previously that non-synonymous mutations during acute infection in the E1 and E2 proteins were higher within cytotoxic T cell epitopes^{9,14}. In this analysis, the third highest genomic variability was observed in the NS5B region reflecting a key role of this protein in the acute infection time window. This region codes for the RNA polymerase which determines the replicative capacity of the virus. It also has numerous T cell epitopes where longitudinal within-host non-synonymous mutations in acute infection have been well characterized²⁶. Previous studies (mainly focussed on genotype 1) have shown that epitopes of NS5B protein targeted by cytotoxic T cells to have a higher mutation rate compared to non-epitopes (similar to the pattern in E1 and E2 proteins)^{27,28}. Some of these mutations were strongly associated with spontaneous clearance^{29,30}.

Once the virus establishes the best functional phenotype for survival during acute infection, it may be disadvantageous to diverge unless there is a significant alteration in the host immune selection pressure, such as the delayed appearance of neutralising antibodies. In HCV, within-host genetic diversity is less than between-host genetic diversity suggesting a strong within-host adaptation process. Proof of concept of this hypothesis requires longitudinal follow up of subjects which may show a dominance of synonymous mutations after the initial acute infection time window. Unfortunately, such samples were not available in this study. However, results of several other studies (both in humans and chimpanzees), even before the advent of next generation sequencing, indirectly support this hypothesis. In chimpanzees infected with HCV, variants with escape mutations in epitopes that impaired cytotoxic T lymphocyte (CTL) recognition remained with minimal diversification for years³¹. In humans, mutations in HCV epitopes targeted by CTL are driven by the prevalence of HLA alleles in a population. For example, some mutations acquired in a patient with an HLA allele is reversed when the virus is introduced to a person without the same HLA allele suggesting a strong within-host adaptation³². With the advent of NGS, the contribution of synonymous and non-synonymous mutations to viral evolution became more observable but there is very limited data for HCV in this regard. A study that compared the Shannon entropy of the NS5B region between acute (n=16) and chronic infections (n=78) with NGS concluded that genomic variability was significantly higher in the latter group and progressively increased over time¹⁵. Interestingly, synonymous mutations accounted for over 78% of the genetic diversity observed in the total sample with the distinction (synonymous vs. non-synonymous) being more visible for chronic than acute phase samples.

The host IFNL3 polymorphism rs12979860 was a determinant of viral genomic variability in acute infection, and more specifically of synonymous mutations across the genome. The traditional view was that IFNL family of cytokines were mainly concerned with epithelial immunity, but recent evidence show it to modulate or influence the function of several different immune cell types such as dendritic cells, neutrophils and natural killer cells^{33,34}. High affinity IFNL3 binding to the receptor induces a better antiviral effect against both HBV and HCV³⁵. Several polymorphisms of the host IFNL3 gene have been associated with HCV clearance, although the mechanisms remain unclear^{10,12,13}. Recently for one such variation at rs4803217, where an unfavourable T allele (instead of G) was associated with

viral persistence, the mechanism of action was discovered to be enhanced micro-RNA binding (induced by HCV infection) to the 3'-UTR of the IFNL3 leading to its degradation³⁶. Interestingly less favourable variations of IFNL3, at rs12979860 and rs4803217 which are both associated with HCV persistence, were also strongly linked to an IFNL4 variant (G, ss469415590) that is also associated with HCV persistence^{37,38}. Unlike IFNL3, expression of IFNL4 is positively associated with HCV persistence. The G variant at ss469415590 allows the expression of IFNL4 while an alternative TT variant (associated with spontaneous clearance) renders it a pseudogene by a frameshift mutation silencing the expression.

In chronic HCV infection, comparative genomic analysis between the host and the virus have demonstrated the host MHC region to influence the genetic polymorphisms in the NS3 and NS4 regions³⁹. Outside the MHC locus, the host IFNL4 genotype had a strong influence on viral NS5B polymorphisms, as well as the viral load⁴⁰. In the present study, for the first time, IFNL3 polymorphism CC at rs12979860 was found to be a significantly associated with increased viral genomic variation in acute HCV infection. However, this high genomic variability was mostly driven by synonymous mutations. This observation is difficult to explain by data from this study, but it is possible that the virus increases its mutation rate when confronted with the unfavourable host allele (for HCV persistence) and the initial increase in synonymous mutations may be a stepping stone to non-synonymous mutations leading to immune escape. However, in some cases, the host immune system may clear the virus before this happens.

In contrast to the host IFNL3 polymorphism, viral genotype was associated with genomic variability of non-synonymous mutations where a higher diversity was observed in subtype 1a compared to 3a. Currently seven genotypes of HCV have been identified and key features common to all include: genome structure, replication process and ability to cause chronic hepatitis⁴¹. Significant differences have been reported between genotypes with regard to time to spontaneous clearance¹⁰, treatment response⁴², host vulnerability to complications such as steatosis and hepatocellular cancer⁴³ and interactions with host immunity⁴⁴. Phylogenetically, genotypes 1 and 3 are distinct with the earlier being more related to genotypes 4, 5 and 6, while the latter is more closely related to genotypes 2 and 7⁴¹. Given the clinically relevant differences between these subtypes, it is plausible that these phylogenetic differences may occur at critical sites such as immune epitopes and may have different profiles in acute and chronic infections. It has previously been demonstrated that there is little overlap of CD4 and CD8 T cell epitopes between subjects with chronic and acute infections in subtypes 1a or 3a⁴⁴. This paper adds to this observation by showing a higher diversity observed across the genome (except the E1 and NS4A region) for subtype 1a compared to 3a in acute infection. This diversity may pose a challenge to develop a pan-genotypic preventative vaccine.

Host IFNL3 CC (at rs12979860) genotype and viral subtype 1a which were associated with a higher genomic variability in this study, have also been shown to be favourably associated with spontaneous clearance by previous authors¹⁰. It is also known that escape mutations of the virus may have epistatic interactions with other mutations to balance inherent fitness costs⁴⁵. The higher diversity of acute infection viruses under unfavourable circumstances

(to the virus) might reflect an attempt of the virus quasi-species populations to find fit escape mutations via epistatic interactions during the critical transition period from acute to chronic infection. However, viral genomic variability was not an independent predictor of spontaneous clearance in this analysis. While the overall pathogen genomic variation in acute infection may be influenced by both viral genotype and host genetic polymorphisms, such diversity does not seem to be a pre-requisite for clearance. In fact, the ability to clear the infection may be determined by fewer, specific mutations at critical sites such as immune epitopes⁴⁶, which may not be identified in a genome / ORF level analysis.

This study has several limitations. Lack of longitudinal samples prevented the analysis of Shannon entropy variability between acute and chronic infection. Also, it is difficult to estimate the optimum sample size for a study of this nature. Given the logistical difficulties in obtaining well defined samples from acute HCV infection, this by far is the largest sample size of deep sequenced full-length HCV genomes available to-date from acute infection time points contributed by nine prospective cohorts with samples collected over two decades. Yet, inadequate power may explain the inability to replicate some associations in previous literature such as the relationship between host gender on spontaneous clearance. On a positive note, the relationships observed in this study were consistent across multiple regions of the genome and some matched with previous literature such as the relationship between IFNL3 polymorphism and spontaneous clearance.

Conclusion

This data provides the most comprehensive analysis of genomic variability across the near full-length HCV ORF in acute infection conducted to-date, using cross continental sequences, showing that variation in acute infection was highest in Envelope and NS5B regions with non-synonymous mutations being the major contributor to the overall variation. Both host IFNL3 polymorphism at rs12979860 and viral genotype were significant determinants of synonymous and non-synonymous viral genomic variability respectively. High genomic variability was associated with traits known to favour spontaneous clearance, namely IFNL3 CC genotype and HCV 1a subtype. However, in this study, genomic variability itself was not an independent predictor of clearance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The infrastructure for sharing of data and specimens in InC³ was funded by the National Institute on Drug Abuse (NIDA) R01DA031056. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of NIDA or the National Institutes of Health (NIH). Research support for the InC³ cohorts includes: the Netherlands National Institute for Public Health and the Environment to the Amsterdam Cohort Study (ACS); Baltimore Before and After Study (BBA) - National Institutes of Health (NIH U19 AI088791); Boston Acute HCV Study: transmission, immunity, outcomes network (BAHSTION) is funded under NIH NIAID U19 AI066345; ATACHC – NIDA RO1 DA 15999-01; HITS-p - National Health and Medical Research Council of Australia (NHMRC) - Project No. 222887, Partnership No. 1016351, Program Nos. 510488 and 1053206; HITS-c – UNSW Hepatitis C Vaccine Initiative and NHMRC Project Grant No. 630483; Networks/MIX – NHMRC Project Grants Nos. 331312 and 545891 and the Victorian Operational Infrastructure Support Programme (Department of Health, Victoria, Australia); HepCo - the Canadian Institutes of Health Research

(MOP-103138 and MOP-106468; Réseau SIDA Maladies Infectieuses, Fonds de la Recherche du Québec-Santé); UFO – NIH R01 DA016017. The following InC3 investigators are funded by research fellowships: NHMRC, Australia – JG, MH, LM, GD, AL; Fonds de la Recherche du Québec, Canada – JB and NS.

References

1. Boeras DI, Hraber PT, Hurlston M, et al. Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(46):E1156–1163. [PubMed: 22065783]
2. Brown RJ, Juttla VS, Tarr AW, et al. Evolutionary dynamics of hepatitis C virus envelope genes during chronic infection. *The Journal of general virology*. 2005;86(Pt 7):1931–1942. [PubMed: 15958671]
3. Bukh J, Miller RH, Purcell RH. Genetic heterogeneity of hepatitis C virus: quasispecies and genotypes. *Seminars in liver disease*. 1995;15(1):41–63. [PubMed: 7597443]
4. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews Genetics*. 2016;17(6):333–351.
5. Eltahla AA, Rodrigo C, Betz-Stablein B, et al. Analysis of resistance-associated substitutions in acute hepatitis C virus infection by deep sequencing across six genotypes and three continents. *Journal of viral hepatitis*. 2017;24(1):37–42. [PubMed: 27666440]
6. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*. 1989;244(4902):359–362. [PubMed: 2523562]
7. Ribeiro RM, Li H, Wang S, et al. Quantifying the diversification of hepatitis C virus (HCV) during primary infection: estimates of the in vivo mutation rate. *PLoS pathogens*. 2012;8(8):e1002881. [PubMed: 22927817]
8. Neumann AU, Lam NP, Dahari H, et al. Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science*. 1998;282(5386):103–107. [PubMed: 9756471]
9. Bull RA, Luciani F, McElroy K, et al. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS pathogens*. 2011;7(9):e1002243. [PubMed: 21912520]
10. Grebely J, Page K, Sacks-Davis R, et al. The effects of female sex, viral genotype, and IL28B genotype on spontaneous clearance of acute hepatitis C virus infection. *Hepatology*. 2014;59(1):109–120. [PubMed: 23908124]
11. Zanon I, Granucci F, Broggi A. Interferon (IFN)-λ Takes the Helm: Immunomodulatory Roles of Type III IFNs. *Frontiers in Immunology*. 2017;8(1661).
12. Ge D, Fellay J, Thompson AJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature*. 2009;461(7262):399–401. [PubMed: 19684573]
13. Thomas DL, Thio CL, Martin MP, et al. Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature*. 2009;461(7265):798–801. [PubMed: 19759533]
14. Bull RA, Leung P, Gaudieri S, et al. Transmitted/Founder Viruses Rapidly Escape from CD8+ T Cell Responses in Acute Hepatitis C Virus Infection. *Journal of virology*. 2015;89(10):5478–5490. [PubMed: 25740982]
15. Montoya V, Olmstead AD, Janjua NZ, et al. Differentiation of acute from chronic hepatitis C virus infection by nonstructural 5B deep sequencing: a population-level tool for incidence estimation. *Hepatology*. 2015;61(6):1842–1850. [PubMed: 25645961]
16. Grebely J, Morris MD, Rice TM, et al. Cohort profile: the international collaboration of incident HIV and hepatitis C in injecting cohorts (InC3) study. *International journal of epidemiology*. 2013;42(6):1649–1659. [PubMed: 23203695]
17. Bull RA, Eltahla AA, Rodrigo C, et al. A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC genomics*. 2016;17(1):247. [PubMed: 26988550]
18. Rodrigo C, Eltahla AA, Bull RA, et al. Phylogenetic analysis of full-length, early infection, hepatitis C virus genomes among people with intravenous drug use: the InC3 Study. *J Viral Hepat*. 2017;24(1):43–52.
19. Hajarizadeh B, Grady B, Page K, et al. Patterns of hepatitis C virus RNA levels during acute infection: the InC3 study. *PloS one*. 2015;10(4):e0122232. [PubMed: 25837807]

20. Reza F An Introduction to Information Theory. New York, USA: Dover Publications, Inc.; 1994.
21. Kuiken C, Combet C, Bukh J, et al. A comprehensive system for consistent numbering of HCV sequences, proteins and epitopes. *Hepatology*. 2006;44(5):1355–1361. [PubMed: 17058236]
22. Sim S, Aw PP, Wilm A, et al. Tracking Dengue Virus Intra-host Genetic Diversity during Human-to-Mosquito Transmission. *PLoS neglected tropical diseases*. 2015;9(9):e0004052. [PubMed: 26325059]
23. Li H, Stoddard MB, Wang S, et al. Elucidation of Hepatitis C Virus Transmission and Early Diversification by Single Genome Sequencing. *PLoS pathogens*. 2012;8(8):e1002880. [PubMed: 22927816]
24. Sobel Leonard A, McClain MT, Smith GJ, et al. Deep Sequencing of Influenza A Virus from a Human Challenge Study Reveals a Selective Bottleneck and Only Limited Intrahost Genetic Diversification. *Journal of virology*. 2016;90(24):11247–11258. [PubMed: 27707932]
25. Uebelhoefer L, Han JH, Callendret B, et al. Stable cytotoxic T cell escape mutation in hepatitis C virus is linked to maintenance of viral fitness. *PLoS pathogens*. 2008;4(9):e1000143. [PubMed: 18773115]
26. Eltahla AA, Leung P, Pirozyan MR, et al. Dynamic evolution of hepatitis C virus resistance-associated substitutions in the absence of antiviral treatment. *Scientific reports*. 2017;7:41719. [PubMed: 28139734]
27. Kim DW, Lee SA, Kim H, Won YS, Kim BJ. Naturally occurring mutations in the nonstructural region 5B of hepatitis C virus (HCV) from treatment-naïve Korean patients chronically infected with HCV genotype 1b. *PloS one*. 2014;9(1):e87773. [PubMed: 24489961]
28. Ruhl M, Knuschke T, Schewior K, et al. CD8+ T-Cell Response Promotes Evolution of Hepatitis C Virus Nonstructural Proteins. *Gastroenterology*. 2011;140(7):2064–2073. [PubMed: 21376049]
29. Neumann-Haefelin C, McKiernan S, Ward S, et al. Dominant influence of an HLA-B27 restricted CD8+ T cell response in mediating HCV clearance and evolution. *Hepatology*. 2006;43(3):563–572. [PubMed: 16496339]
30. Neumann-Haefelin C, Thimme R. Adaptive immune responses in hepatitis C virus infection. *Current topics in microbiology and immunology*. 2013;369:243–262. [PubMed: 23463204]
31. Erickson AL, Kimura Y, Igarashi S, et al. The outcome of hepatitis C virus infection is predicted by escape mutations in epitopes targeted by cytotoxic T lymphocytes. *Immunity*. 2001;15(6):883–895. [PubMed: 11754811]
32. Timm J, Lauer GM, Kavanagh DG, et al. CD8 epitope escape and reversion in acute HCV infection. *The Journal of experimental medicine*. 2004;200(12):1593–1604. [PubMed: 15611288]
33. Lasfar A, de laTorre A, Abushahba W, et al. Concerted action of IFN-alpha and IFN-lambda induces local NK cell immunity and halts cancer growth. *Oncotarget*. 2016;7(31):49259–49267. [PubMed: 27363032]
34. Koltsida O, Hausding M, Stavropoulos A, et al. IL-28A (IFN-lambda2) modulates lung DC function to promote Th1 immune skewing and suppress allergic airway disease. *EMBO molecular medicine*. 2011;3(6):348–361. [PubMed: 21538995]
35. Mendoza JL, Schneider WM, Hoffmann HH, et al. The IFN-lambda-IFN-lambdaR1-IL-10Rbeta Complex Reveals Structural Features Underlying Type III IFN Functional Plasticity. *Immunity*. 2017;46(3):379–392. [PubMed: 28329704]
36. McFarland AP, Horner SM, Jarret A, et al. The favorable IFNL3 genotype escapes mRNA decay mediated by AU-rich elements and hepatitis C virus-induced microRNAs. *Nature immunology*. 2014;15(1):72–79. [PubMed: 24241692]
37. Hong M, Schwerk J, Lim C, et al. Interferon lambda 4 expression is suppressed by the host during viral infection. *The Journal of experimental medicine*. 2016;213(12):2539–2552. [PubMed: 27799623]
38. Lu YF, Mauger DM, Goldstein DB, Urban TJ, Weeks KM, Bradrick SS. IFNL3 mRNA structure is remodeled by a functional non-coding polymorphism associated with hepatitis C virus clearance. *Scientific reports*. 2015;5:16037. [PubMed: 26531896]
39. Ansari MA, Pedergnana V, L C Ip C, et al. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nature genetics*. 2017;49:666. [PubMed: 28394351]

40. Ansari MA, Pedergnana V, C LCI, et al. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nature genetics*. 2017;49(5): 666–673. [PubMed: 28394351]
41. Smith DB, Bukh J, Kuiken C, et al. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology*. 2014;59(1): 318–327. [PubMed: 24115039]
42. Poynard T, Marcellin P, Lee SS, et al. Randomised trial of interferon alpha2b plus ribavirin for 48 weeks or for 24 weeks versus interferon alpha2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. International Hepatitis Interventional Therapy Group (IHIT). *Lancet (London, England)*. 1998;352(9138):1426–1432.
43. Di Bisceglie AM. Hepatitis C and hepatocellular carcinoma. *Hepatology*. 1997;26(3 Suppl 1):34s–38s. [PubMed: 9305661]
44. von Delft A, Humphreys IS, Brown A, et al. The broad assessment of HCV genotypes 1 and 3 antigenic targets reveals limited cross-reactivity with implications for vaccine design. *Gut*. 2016;65(1):112–123. [PubMed: 26092843]
45. Gal-Tanamy M, Keck ZY, Yi M, et al. In vitro selection of a neutralization-resistant hepatitis C virus escape mutant. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(49):19450–19455. [PubMed: 19052239]
46. Cox AL, Mosbruger T, Mao Q, et al. Cellular immune selection with hepatitis C virus persistence in humans. *The Journal of experimental medicine*. 2005;201(11):1741–1752. [PubMed: 15939790]

Figure 1. Mean standardized genomic variability across different protein coding regions of the HCV genome in acute infection

Mean standardized genomic variability measured as Shannon Entropy (SE) are presented with standard deviations (whiskers) across 116 acute infection samples. Legends; a) red/dot: SE of all mutations, b) black/box: SE of synonymous mutations, c) blue/triangle: SE of non-synonymous mutations.

The statistically significant differences were as follows; a) for total SE: E1 vs. NS4A, b) for non-synonymous SE: E1 vs. core, NS3, NS4A, NS4B and NS5A, c) for synonymous SE: none

Table 1.

Characteristics of included samples for Shannon entropy analysis.

Characteristic	Number	Percentage (%)
<i>Estimated date of infection</i>		
Less than 28 days	49	42.2
More than 28 days	67	57.8
0 – 60 days	62	53.4
61–120 days	32	27.6
121 – 180 days	22	19.0
<i>Continent of origin</i>		
Australia	35	30.2
North America	81	69.8
<i>Gender</i>		
Male	79	68.1
Female	37	31.9
<i>Genotype</i>		
1a	59	50.9
3a	37	31.9
Other	20	17.2
<i>Incarceration at time of sampling</i>		
Yes	99	85.3
No	17	14.7
<i>Host IFNL3 phenotype (rs12979860)</i>		
CC	47	40.5
Others	44	37.9
Missing	25	21.6
<i>Infection outcome</i>		
Spontaneous clearance	23	19.8
Chronic infection	71	61.2
Missing	22	19.0

Table 2.

Significant predictors of synonymous and non-synonymous Shannon entropies (independent T test, $p < 0.05$).

Genome region	Predictors of SE of synonymous mutations	Predictors of SE of non-synonymous mutations
Near full-length open reading frame	IFNL3 genotype *	Viral genotype *
Core	IFNL3 genotype *	Viral genotype *
	Viral genotype	
E1	IFNL3 genotype *	Viral genotype
E2	IFNL3 genotype *	Viral genotype *
P7	-	-
NS2	IFNL3 genotype *	Viral genotype *
NS3	IFNL3 genotype *	Viral genotype *
NS4A	IFNL3 genotype *	Viral genotype
NS4B	IFNL3 genotype *	Viral genotype *
	Viral genotype	
NS5A	IFNL3 genotype *	Viral genotype *
NS5B	Viral genotype *	Viral genotype *

* predictors that remained significant on multiple regression analysis

Note: host IFNL3 genotype refers to rs12979860 polymorphism categorized as CC vs. others, viral genotype is categorized as subtype 1a vs. 3a. CC polymorphism of the host and subtype 1a of the virus are associated with higher SEs.