Published in final edited form as:
Infect Genet Evol. 2019 July ; 71: 36–41. doi:10.1016/j.meegid.2019.02.032.

Genomic characterisation of hepatitis C virus transmitted founder variants with deep sequencing

Arunasingam Abayasingam[a], Preston Leung[b], Auda Eltahla[a], Rowena A. Bull[a], Fabio Luciani[a], Jason Grebely[b], Gregory J. Dore[b], Tanya Applegate[b], Kimberly Page[c], Julie Bruneau[d], Andrea L. Cox[e], Arthur Y. Kim[f], Janke Schinkel[g], Naglaa H. Shoukry[d], Georg M. Lauer[f], Lisa Maher[b], Margaret Hellard[h,i,j,k], Maria Prins[g,l], Andrew Lloyd[b], and **Chaturaka Rodrigo[a] on behalf of the InC3 Study Group**

[a]School of Medical Sciences, Faculty of Medicine, UNSW Sydney, NSW, Australia, [b]The Kirby Institute, UNSW Sydney, NSW, Australia, [c]Division of Epidemiology, Biostatistics and Preventive Medicine, University of New Mexico, Albuquerque, New Mexico, USA, [d]CRCHUM, Université de Montréal, Montreal, QC, Canada, [e]Department of Medicine, Johns Hopkins Medical Institutions, Baltimore, MD, USA, [f]Harvard Medical School, Boston, MA, USA, [g]Department of Internal Medicine, Division of Infectious Diseases, Tropical Medicine and AIDS, Center for Infection and Immunity Amsterdam, Academic Medical Center, Meibergdreef 9, Amsterdam, The Netherlands, [h]Burnet Institute, Melbourne, VIC, Australia, [i]Monash University, Melbourne, Australia, [j]Alfred Hospital, Melbourne, Australia, [k]Doherty Institute and Melbourne School of Population and Global Health, University of Melbourne, [l]GGD Public Health Service of Amsterdam, Amsterdam, The Netherlands

Abstract

Transfer of hepatitis C virus (HCV) infection from a donor to a new recipient is associated with a bottleneck of genetic diversity in the transmitted viral variants. Existing data suggests that one, or very few, variants emerge from this bottleneck to establish the infection (transmitted founder [T/F] variants). In HCV, very few T/F variants have been characterized due to the challenges of obtaining early infection samples and of high throughput viral genome sequencing.

This study used a large, acute HCV, deep-sequenced dataset from first viremia samples collected in nine prospective cohorts across four countries, to estimate the prevalence of single T/F viruses,

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

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

[2] This position was highly variable with each founder sequence having a different amino acid

Special note: As suggested by reviewers, haplotypes were also reconstructed for two more windows of the genome. These included; a) last part of E2 – P7 – NS2 (1439 nts in length) and b) last part of NS3 – NS4A – NS4B (1544 nts in length). Together all reconstructed regions accounted for 80% of the full HCV genome. The existence of a single T/F was reconfirmed for all 14 samples according to the criteria specified in methods.

Disclosures
The authors have nothing to disclose in relation to this work.

and to identify host and virus-related factors associated with infections initiated by a single T/F variant. The short reads generated by Illumina sequencing were used to reconstruct viral haplotypes with two haplotype reconstruction algorithms. The haplotypes were examined for random mutations (Poisson distribution) and a star-like phylogeny to identify T/F viruses. The findings were cross-validated by haplotype reconstructions across three regions of the genome (Core-E2, NS3, NS5A) to minimize the possibility of spurious overestimation of single T/F variants.

Of 190 acute infection samples examined, 54 were very early acute infections (HCV antibody negative, RNA positive), and single transmitted founders were identified in 14 (26%, 95% CI: 16–39%) after cross validation across multiple regions of the genome with two haplotype reconstruction algorithms. The presence of a single T/F virus was not associated with any host or virus-related factors, notably viral genotype or spontaneous clearance.

In conclusion, approximately one in four new HCV infections originates from a single T/F virus. Resolution of genomic sequences of single T/F variants is the first step in exploring unique properties of these variants in the infection of host hepatocytes.

Keywords

Hepatitis C virus; transmitted-founder virus; people who inject drugs; InC3 study; Next generation sequencing

1. Introduction

Rapidly mutating RNA viruses such as human immunodeficiency virus (HIV), hepatitis C virus (HCV) and influenza virus exist as populations of genetically diverse viral variants in each infected host (Farci et al., 2000; Koelle et al., 2006; Shankarappa et al., 1999; Sheridan et al., 2004). Despite this diversity, observations in HIV and HCV have demonstrated that upon transmission from an infected donor to a new recipient, very few variants will successfully establish infection in the new host (Bull et al., 2011). These “transmitted founder” viruses (T/F viruses) have been under intense scrutiny in recent years as they are critical in understanding the early events leading to successful establishment of infection, and may provide an “Achilles’ heel” to be targeted by vaccine strategies (Bimber et al., 2009; Boeras et al., 2011; Bull et al., 2015; Bull et al., 2011). In addition, they are linked to disease characteristics that are important in treatment decision making. For example in HIV, transmission through injecting drug use (as opposed to per mucosal sexual transmission) is associated with a higher number of HIV T/F variants (Bar et al., 2010) and the presence of multiple T/F variants in turn, is associated with higher HIV set point viral loads (Janes et al., 2015).

In HCV infection, characterization of T/F sequences is potentially useful in understanding the relationship between viral traits and primary infection outcome (i.e., spontaneous clearance versus chronic infection) (Grebely et al., 2014). However, there are three major obstacles in accurate determination of HCV T/Fs.

Firstly, given that primary HCV infection is largely asymptomatic, identifying individuals with early acute HCV infection to study T/F populations requires is difficult. This requires periodic sampling from high risk recruits within prospective cohorts (e.g. intravenous drug users). Limited data from one such cohort showed that HCV transmitted founders dominate the quasi-species population for approximately 100 days following establishment of infection, after which a second bottleneck in viral diversity was observed, likely driven by adaptive immune responses of the host (Bull et al., 2011). Therefore, in HCV, studies of T/F variants should ideally sample subjects within 100 days of infection.

Secondly, there is no ideal method that is simultaneously cost effective, accurate and high throughput to characterize T/Fs. Single genome amplification and sequencing does well in terms of accuracy, but it is laborious and not cost effective. The advent of next generation sequencing technologies has made viral sequencing more affordable and high throughput (Goodwin et al., 2016). However, the currently popular paired end next generation sequencing technology (Illumina) fragments the full-length viral cDNA prior to sequencing. The short-read length output of such sequencing (approximately 150–300 nucleotides [nts]) is then used to reconstruct viral haplotypes with bioinformatics algorithms (Töpfer et al., 2013; Zagordi et al., 2010). Several such algorithms exist (e.g. haploclique, ShoRAH, QuRe, Predicthaplo), but the recovery rate of the true viral variants with each method is not optimal (Pandit and de Boer, 2014). In addition, the probability of reconstructing spurious variants increases with the length of the reconstructed haplotypes (Zagordi et al., 2012) and ideally such lengths are capped around 2000nts. Currently there is no gold standard against which the accuracy of reconstructed haplotypes can be compared, but *in silico* experiments with simulated datasets (mixes of known sequences in pre-determined proportions) have been helpful in benchmarking haplotype reconstruction algorithms. A comparison of ShoRAH, QuRe and Predicthaplo on reconstructing HIV haplotypes with simulated datasets showed ShoRAH and QuRe overestimated the number of haplotypes. Predicthaplo accurately reconstructed a subset of the most prevalent haplotypes, but not all of them (Pandit and de Boer, 2014). Quasirecomb is another algorithm that employs a hidden Markov model to reconstruct haplotypes assuming that the quasi-species population emerges from a few dominant variants. A comparison of Quasirecomb with Predicthaplo for reconstruction of longer fragments of the HIV genome sequenced on the Illumina platform showed that Quasirecomb performed better in accurate reconstruction of the original variants in the mix (Giallonardo et al., 2014).

Thirdly, there is no current consensus on how many regions of the viral genome should be sampled (as viral haplotype reconstruction is limited to lengths of 2000nts to reduce errors) to confidently identify transmitted founders. Previous studies that used next generation sequencing followed by haplotype reconstruction to identify T/Fs, have relied on reconstructing a single region of the genome.

The work described in this paper uses a large dataset of deep sequenced, full-length viruses obtained from subjects with acute HCV infections identified in nine prospective cohorts described previously (Rodrigo et al., 2017) to characterize HCV transmitted founders. Noting the above-mentioned difficulties, a very conservative approach of cross validating findings via two different haplotype reconstruction algorithms across multiple windows of the HCV genome was undertaken to identify transmitted founder variants. Once these variants were identified, the sequences were phylogenetically compared with chronic infection sequences to identify any unique characteristics.

2. Methods

2.1 Sample selection

The establishment of the InC3 Virus Sequence Repository (InC3 VRS) has been described previously (Grebely et al., 2013; Rodrigo et al., 2017). In brief, full genome amplification and deep sequencing (Illumina) was carried out on incident viraemic HCV samples (collected within 180 days of the estimated date of infection) which originated from nine prospective cohorts in four countries [North America (n=4), Australia (n=4) and Europe (n=1)]. The estimated date of infection for each sample was determined as previously described (Hajarizadeh et al., 2015). Among 4880 subjects, 543 incident infections were identified in all nine cohorts. Of the stored sample sets from these subjects, 368 plasma or serum samples were available and full length genomes were successfully sequenced from 190 of these samples using paired end short read technology (Illumina, MiSeq) with an average coverage of approximately 17,000 per base position (Rodrigo et al., 2017). For the analysis described here, sequence data for an “early acute” sample group was selected from this repository, when all the following criteria were met: a) HCV RNA positive and HCV antibody negative at the initial incident sampling time-point; b) a coverage of at least 100 per base position across the genome; and c) availability of a previously generated autologous reference consensus sequence with a quality score of Q40 (Illumina, Miseq) across the length of the genome. For comparison, a set of chronic HCV infection sequences from two previously published cohort studies, available in public sequence databases, were used (Foster et al., 2015; Kuntzen et al., 2008). One of these studies included treatment-naïve patients from North America and Europe who had been infected for at least one year prior to sequencing (Kuntzen et al., 2008), and the other recruited chronically infected patients from the USA, Europe, Australia and New Zealand (Foster et al., 2015). The dataset of the first study was enriched in genotypes 1a and 1b, while the latter was mostly genotypes 2 and 3a. As the number of chronic infection sequences available from these two studies greatly exceeded that of acute infection sequences, for phylogeny trait analysis described below, a subset of genotype-matched sequences was randomly selected.

The raw Illumina read outputs were pre-prepared with an established bioinformatics pipeline that trimmed and clipped reads (Trimmomatic, version 0.33; Illuminaclip, version 0.33) to remove adapter sequences and low-quality reads (Bolger et al., 2014). Short reads with a length less than 50 nucleotides and unpaired reads were excluded. The trimmed reads were aligned against an autologous reference sequence that was generated previously using the Burrows-Wheeler aligner (Rodrigo et al., 2017). The aligned sequences were compressed and sorted into a Binary Alignment/Map format using SAMtools, version 1.3.1. (Supplementary material).

2.2 Haplotype reconstruction

Haplotype reconstruction and frequency estimation was performed on three distinct regions of the HCV genome, namely: a) Core, E1 and E2 as a continuous segment, b) NS3, and c) NS5A, with two computational algorithms, ShoRAH and QuasiRecomb (Töpfer et al., 2013; Zagordi et al., 2010). Based on previous results, the genome length for haplotype reconstruction was kept at approximately 1500nts, as longer genome lengths are more likely to produce spurious variants (Bull et al., 2011). The Core-E2 region was selected as the E1E2 segment was expected to have high diversity and is important for viral entry into hepatocytes in early infection. NS3 and NS5A were selected as they are the largest individual protein coding regions matching a length of approximately 1500nts (sequencing of the NS5B region was incomplete in the original dataset). The specifications used in each pipeline are detailed in the supplementary material. Of all options available for haplotype reconstruction algorithms, ShoRAH and Quasirecomb were selected as they are on the opposite ends of the spectrum with regard to the recovery of true variants and the number of false variants reconstructed (based on the results of the *in silico* studies mentioned above) (Giallonardo et al., 2014; Pandit and de Boer, 2014). Both computational algorithms provide a list of reconstructed haplotypes and their frequency of occurrence as the final output. For the purposes of this analysis, haplotypes with a frequency of occurrence less than 1% were excluded, as at such low frequencies it is difficult to differentiate spurious variants (due to sequencing errors and artefacts) from true variants (Bull et al., 2011).

2.3 Estimation of T/F sequences

Based on the assumption that in early infection, prior to immune selection pressure, mutations in the HCV genome should be random occurrences, and hence described best by a Poisson distribution, a Poisson-Fitter test was used to test this null hypothesis. An online version of this tool is available at the Los Alamos HIV database (Giorgi et al., 2010). This tool performed two tests: a statistical test to see if the observed mutations between the variants fitted a Poisson distribution; and a phylogenetic tree topology test to verify a star-like phylogeny to confirm divergence from a common source (T/F variant) (Giorgi et al., 2010). For this analysis, the mutation rate of the HCV genome was set to 2.5×10^{-5} per replication, and this value was assumed to be uniform across the genome (Ribeiro et al., 2012). If both tests were positive (star-like phylogeny with the observed mutations fitting a Poisson distribution), the most frequently observed haplotype in the population was recorded as the T/F variant (a single T/F). For ShoRAH generated haplotypes, a star-like phylogeny was observed with a a divergence from Poisson distribution. These were accepted as single T/Fs provided that Quasirecomb generated haplotypes fulfilled criteria for both a star-like phylogeny and mutations fitting a Poisson distribution.

For each sample, this analysis was repeated for Core-E2, NS3 and NS5A regions to cross-validate the results. This study elected to use a conservative definition of a T/F virus based on the assumption that a true T/F variant should conform to above criteria regardless of the area of the genome examined and regardless of the algorithm used. However, given the inability of the algorithms to accurately predict low frequency variants, existence of a single T/F sequence was also accepted when for Core-E2 and one other region, both algorithms agreed on a single T/F variant, and for the remaining region, at least one algorithm predicted a single founder variant.

2.4 Phylogenetic characterization of T/F sequences

Given that in early infection, the structural proteins of the virus play a critical role in entry to host hepatocytes, single T/F variant Core-E2 sequences (2238nts, 746 amino acid residues) were compared to chronic infection sequences obtained from public databases at the consensus level, and as individual sequences (Ansari et al., 2017; Kuntzen et al., 2007). This analysis was conducted separately for the two most common subtypes, 1a and 3a. Consensus sequences of the Core-E2 region were compared in a pairwise alignment while individual sequences were assessed in a phylogeny trait analysis using BaTs (Bayesian analysis of tip significance) algorithm (Parker et al., 2008). The hypothesis for the latter analysis was that if single T/F variants were phylogenetically unique, they should cluster with like-sequences in a phylogenetic tree. The input for this analysis was a Posterior set of trees generated by a Bayesian Monte Carlo Markov Chain analysis. The posterior set of trees were generated with BEAST version 1.8 (Drummond and Rambaut, 2007) using parameters as described previously for 1a and 3a HCV subtypes (Rodrigo et al., 2016).

2.5 Demographic, host genotype and behavioural correlates of a single T/F variants

The demographic, behavioural, and infection outcome data from the early acute cases were summarized (with measures of central tendency and dispersion) and analysed to observe any associations with the existence of a single T/F variant versus multiple variants using the Chi square test ($p<0.05$). Variables included age, sex, ethnicity, history of incarceration, most frequently injecting drug, HCV genotype (1a vs 3a), HCV viral load at time of sampling, host IFNγ3 gene polymorphism (at position rs12979860, known to be associated with spontaneous clearance) and the outcome of the infection (Rodrigo et al., 2017). Statistical analysis was carried out using the SPSS statistical software (version 24 IBM, New York, USA).

2.6 Ethical considerations

Ethical approval for this study was granted by UNSW Sydney, Australia (No. 14201). In addition, each of the individual cohorts had ethical approval for original data, specimen collection and analysis.

3. Results

The InC3 VSR had 190 deep sequenced full-length genomes from samples collected during acute infection. Of these 54 (1a - 32, 2a - 2, 2b - 3, 4a- 1, 3a - 16) were early acute (preseroconversion) samples that met the inclusion criteria for this study. The subjects from which these initial viraemic samples originated were mostly male (33/54, 61%), young (mean age: 24, SD ± 7 years), hepatitis C antiviral treatment naïve (54/54, 100%), and Caucasian (41/54, 76%) (Table 1).

In the early acute samples, 26% (14/54, 95% CI: 16 – 39%) were identified as having a single T/F variant (genotype 1a - 5, 2a - 1. 3a – 7, 4a – 1). In 12 of these subjects, both bioinformatics algorithms were concordant across all three regions of the genome tested. Of the samples not fitting the pre-determined pattern of a single T/F, both algorithms agreed on high diversity that could not be explained by a single T/F for at least one region of the genome in 70% of samples (38/54, 95% CI: 57 – 81%) and both algorithms uniformly agreed on a high diversity across all three regions for 25% (13/54, 95% CI: 15 – 37%) of the samples. In two samples, both algorithms agreed for a single T/F pattern in both NS3 and NS5A regions, but one algorithm disagreed for Core-E2 region.

A detailed comparison of results between the two haplotype reconstruction algorithms is provided in Table 2. As expected, ShoRAH identified more haplotype variants that were above the 1% cut-off for frequency of occurrence in comparison to the output of Quasirecomb. Of the different genomic regions examined, Core-E2 was the most diverse and least likely to designate a single T/F, while the NS5A region was the most likely. The predictions made by each algorithm for genomic areas of interest for each sample are given in supplementary table 1.

Phylogenetic comparisons were focussed on the Core-E2 region, which codes for the structural proteins playing a crucial role in viral entry during acute infection. For subtype 1a, the consensus sequence of five single T/F viruses (Core-E2 only) were compared with that of 330 chronic infection sequences. Fifteen non-synonymous differences were observed (excluding those in HVR1 region) with all except two being located within the E2 region. Of these, seven differences were in previously identified epitope clusters for broadly neutralizing antibodies, and three sites (S424R, Y444H, E531[2]) were located immediately next to a glycosylation site or a CD81 binding site. Each single T/F polypeptide sequence had between 25 – 39 pairwise differences to the consensus polypeptide sequence from 1a chronic sequences in the Core-E2 region after excluding the HVR1 region (94.6 – 96.5% identity). For subtype 3a, the consensus sequence of seven single T/F viruses were compared with that of 240 chronic infection sequences. Seven non-synonymous differences were observed, with all except one being in the E2 protein (excluding those in HVR1 region). Two of these (A525V/N), T529N/S were within an epitope cluster for broadly neutralizing antibodies and one of them (529) may be a CD81 binding site. Each single T/F polypeptide sequence had between 16 – 40 pairwise differences to the consensus polypeptide sequence from 3a chronic sequences in the Core-E2 region after excluding the HVR1 region (94.4 – 97.8% identity). Both NS3 and NS5A areas had less variation when compared to the chronic consensus sequence (Table 3). However, given the small sample size of single founders, it was not possible to use statistical tests to determine the significance of these observations.

For the phylogeny trait analysis, treatment naïve, subtype 1a (n=61) and 3a chronic sequences (n=38) were randomly selected (one sequence per patient) from the publicly available sequence datasets to compare with the single T/F viruses. There was no preferential clustering to indicate relatedness of the single founders in either genotype.

Gender, ethnicity (White Caucasian vs. others), age, continent of origin (Australia vs. North America), incarceration at the time of sampling, viral genotype (1a vs. others), type of injecting drug (heroin vs. others), host IFNγ3 genotype (rs12979860, CC vs. non-CC) and outcome of infection were not significantly associated with the presence of a single T/F (versus non-single T/F in early acute infection) (Chi square test or independent sample T test, $p>0.05$). However, when genotype 1a was compared against 3a the latter was statistically significantly more likely to have a single T/F (5/32 vs. 7/16, $p<0.05$)

4. Discussion

This analysis of a unique deep-sequenced early acute infection HCV sequence dataset placed the frequency of occurrence of a single T/F virus at approximately 26% of new infections. It also highlights the potential pitfalls in assigning a single T/F status from reconstructed haplotypes without cross-validation across multiple regions of the genome.

From a biological perspective, a single T/F is a distinct entity defined by its entire genome. Therefore, ideally, the results should be concordant regardless of the algorithm used or the genome region tested. Yet, this study shows that given the limitations of currently available next generation sequencing methods and haplotype reconstruction algorithms, spurious assumptions of a single T/F status may be made when less diverse areas of the genome are tested with algorithms that reconstruct fewer variants. If a single region is to be used due to difficulties in sequencing full genomes (or hemigenomes) for T/F estimation, the findings here suggest that the more diverse Envelope region should be preferred as the risk of overestimating T/Fs is comparatively lower. In this consideration, it should be noted that in two samples studied here, both algorithms agreed on single T/F when the NS3 and NS5A regions were utilised but differed in the results for the Core-E2 region (with one agreeing for a single T/F). In such difficult situations, the absence of a gold standard makes it impossible to differentiate the technical bias of algorithms vs. the true existence of a single T/F. An alternative to solve this issue is to avoid haplotype reconstruction altogether and use single genome amplification (SGA) followed by sequencing (Brown et al., 2012; Li et al., 2012; Li et al., 2015; Mitchell et al., 2015). SGA can accurately sequence viral haplotypes to identify mutations free from potential recombination artefacts as demonstrated by Li et al. previously. Li et al. compared their method against a different NGS platform (454 pyrosequencing) which generated longer reads with relatively more errors compared to Illumina technology used in this paper. While the latter technology is more accurate in terms of mutations, it generates shorter reads making haplotype reconstruction biased to many assumptions. However, SGA is labour-intensive and costly. Hence alternative methods that use NGS data with haplotype reconstruction for identifying homogenous viral populations in early infection in a high throughput manner was successfully assessed in this work. Validating these findings against SGA is beyond the scope of this paper though a possibility. It is also possible to validate the haplotypes reconstructed via molecular cloning and sequencing. This method has already established the validity of ShoRAH generated haplotypes (Bull et al., 2011). Another alternative nowadays is to use third generation deep sequencing (single molecule long read sequencing) currently offered on two commercial platforms: Pacific Biosciences (PacBio – Single molecule real-time sequencing) and Oxford Nanopore Technology (Goodwin et al., 2016; Ip et al., 2015). Unfortunately, both these platforms have a relatively high error rate in base calling and are not yet optimized for viral variant sequencing. Therefore, for high throughput processing, Illumina sequencing followed by haplotype reconstruction remains the preferred choice.

Most early acute samples did not have a single T/F variant. This can be interpreted in two ways - the transmission and survival of many T/F variants, or early high diversity emerging from an extinct single T/F. Due to the unavailability of longitudinal samples from the same subjects to track the evolution of the mutation patterns (e.g. as distinct populations in the case of multiple founders) this issue could not be resolved for this dataset. However, previous work by Li et al. with SGA and sequencing is helpful in clarifying this issue. They observed a higher number of transmitted founders in acute-acute transmissions or when the donor had recently undergone a bottleneck event (Li et al., 2012). Having access to longitudinal samples, they demonstrated that the average number of lineages did not change from the initial sampling time point throughout the other early infection sampling time points making it more likely to be multiple infecting founders. The estimate for the prevalence of single T/F given here is a conservative one as it required the concordance between two haplotype reconstruction algorithms. Use of any one of these algorithms by itself would increase this estimate. However, given the limitations of each algorithm and in the absence of a gold standard to compare with, it is difficult to choose one algorithm over another. Therefore, the approach described here is likely to have good specificity in identifying true single transmitted founders with the potential compromise in sensitivity which cannot be further improved with currently available technology.

The phylogenetic comparisons with BaTs did not reveal any unique characteristics of T/F virus Core-E2 sequences compared to chronic sequences. Yet, we found that genotype 3a is more likely to present with a single T/F than genotype 1a. However, the total number of T/F viruses available per genotype was not large enough to draw valid conclusions from this analysis. Although comparison with chronic infection sequences is one method to detect unique features of single T/F viruses, this approach alone cannot resolve biological significance as there may be differences in the secondary or tertiary structure of the Envelope proteins induced by non-synonymous mutations that are difficult to infer by sequence analysis alone. Despite this study having access to a large set of well-defined acute infections collected in multiple countries over 25 years, the number of single T/F viruses identified per genotype were still too small to run a valid statistical comparison of mutations, demonstrating the difficulty of accurate characterization of T/F variants even with current best technology and resources. Nevertheless, the discovery of these sequences is a step forward in evaluating the phenotype of single HCV T/F variants using assays for host cell receptor binding and antibody-mediated neutralization.

5. Conclusions

A single T/F virus was observed in 26% of acute HCV infections. When reconstructed haplotypes are used for estimating T/F variants, there is a risk of overestimating their abundance if more conserved regions of the genome are used. Given that T/F viruses are a biological entity defined by the complete genome, and that reliable haplotype reconstruction is limited to up to 20% of the length of the HCV genome, it is wise to cross validate the findings in haplotype reconstruction using several windows along the genome before assigning the T/F status to a variant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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 U19 AI066345; ATAHC – 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); UFO – NIH R01 DA016017. The following InC3 investigators are funded by research fellowships: NHMRC, Australia – JG, MH, LM, GD, ARL; Fonds de la Recherche du Québec, Canada – JB and NHS.

Abbreviations

BaTs	Bayesian analysis of tip significance
cDNA	Complimentary deoxyribonucleic acid
HCV	Hepatitis C virus
HIV	Human immunodeficiency virus
HVR1	hypervariable region 1
IFNλ3	interferon lambda 3
InC3	the International Collaboration of Incident HIV and Hepatitis C in Injecting cohorts
InC3 VRS	InC3 viral sequence repository
NS3–5	nonstructural proteins 3–5
RNA	ribonucleic acid
T/F	transmitted / founder

References

Ansari MA, Pedergnana V, C LCI, Magri A, Von Delft A, Bonsall D, Chaturvedi N, Bartha I, Smith D, Nicholson G, McVean G, Trebes A, Piazza P, Fellay J, Cooke G, Foster GR, Hudson E, McLauchlan J, Simmonds P, Bowden R, Klenerman P, Barnes E, Spencer CCA, 2017 Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. Nature genetics 49, 666–673. [PubMed: 28394351]

Bar KJ, Li H, Chamberland A, Tremblay C, Routy JP, Grayson T, Sun C, Wang S, Learn GH, Morgan CJ, Schumacher JE, Haynes BF, Keele BF, Hahn BH, Shaw GM, 2010 Wide variation in the multiplicity of HIV-1 infection among injection drug users. Journal of virology 84, 6241–6247. [PubMed: 20375173]

Bimber BN, Chugh P, Giorgi EE, Kim B, Almudevar AL, Dewhurst S, O'Connor DH, Lee HY, 2009 Nef gene evolution from a single transmitted strain in acute SIV infection. Retrovirology 6, 57. [PubMed: 19505314]

Boeras DI, Hraber PT, Hurlston M, Evans-Strickfaden T, Bhattacharya T, Giorgi EE, Mulenga J, Karita E, Korber BT, Allen S, Hart CE, Derdeyn CA, Hunter E, 2011 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 108, E1156–1163. [PubMed: 22065783]

Bolger AM, Lohse M, Usadel B, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. [PubMed: 24695404]

Brown RJ, Hudson N, Wilson G, Rehman SU, Jabbari S, Hu K, Tarr AW, Borrow P, Joyce M, Lewis J, Zhu LF, Law M, Kneteman N, Tyrrell DL, McKeating JA, Ball JK, 2012 Hepatitis C virus envelope glycoprotein fitness defines virus population composition following transmission to a new host. Journal of virology 86, 11956–11966. [PubMed: 22855498]

Bull RA, Leung P, Gaudieri S, Deshpande P, Cameron B, Walker M, Chopra A, Lloyd AR, Luciani F, 2015 Transmitted/Founder Viruses Rapidly Escape from CD8+ T Cell Responses in Acute Hepatitis C Virus Infection. Journal of virology 89, 5478–5490. [PubMed: 25740982]

Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, Cameron B, Maher L, Dore GJ, White PA, Lloyd AR, 2011 Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS pathogens 7, e1002243. [PubMed: 21912520]

Drummond AJ, Rambaut A, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 7, 214. [PubMed: 17996036]

Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerenwinkel N, 2008 Viral population estimation using pyrosequencing. PLoS Comput Biol 4, e1000074. [PubMed: 18437230]

Farci P, Shimoda A, Coiana A, Diaz G, Peddis G, Melpolder JC, Strazzera A, Chien DY, Munoz SJ, Balestrieri A, Purcell RH, Alter HJ, 2000 The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. Science 288, 339–344. [PubMed: 10764648]

Foster GR, Pianko S, Brown A, Forton D, Nahass RG, George J, Barnes E, Brainard DM, Massetto B, Lin M, Han B, McHutchison JG, Subramanian GM, Cooper C, Agarwal K, 2015 Efficacy of sofosbuvir plus ribavirin with or without peginterferon-alfa in patients with hepatitis C virus genotype 3 infection and treatment-experienced patients with cirrhosis and hepatitis C virus genotype 2 infection. Gastroenterology 149, 1462–1470. [PubMed: 26248087]

Giallonardo FD, Töpfer A, Rey M, Prabhakaran S, Duport Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR, Patrignani A, Däumer M, Beisel C, Rusert P, Trkola A, Günthard HF, Roth V, Beerenwinkel N, Metzner KJ, 2014 Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Research 42, e115–e115. [PubMed: 24972832]

Giorgi EE, Funkhouser B, Athreya G, Perelson AS, Korber BT, Bhattacharya T, 2010 Estimating time since infection in early homogeneous HIV-1 samples using a poisson model. BMC bioinformatics 11, 532. [PubMed: 20973976]

Goodwin S, McPherson JD, McCombie WR, 2016 Coming of age: ten years of next-generation sequencing technologies. Nature reviews. Genetics 17, 333–351.

Grebely J, Morris MD, Rice TM, Bruneau J, Cox AL, Kim AY, McGovern BH, Shoukry NH, Lauer G, Maher L, Lloyd AR, Hellard M, Prins M, Dore GJ, Page K, In CSG, 2013 Cohort profile: the international collaboration of incident HIV and hepatitis C in injecting cohorts (InC3) study. Int J Epidemiol 42, 1649–1659. [PubMed: 23203695]

Grebely J, Page K, Sacks-Davis R, van der Loeff MS, Rice TM, Bruneau J, Morris MD, Hajarizadeh B, Amin J, Cox AL, Kim AY, McGovern BH, Schinkel J, George J, Shoukry NH, Lauer GM, Maher L, Lloyd AR, Hellard M, Dore GJ, Prins M, 2014 The effects of female sex, viral genotype, and IL28B genotype on spontaneous clearance of acute hepatitis C virus infection. Hepatology 59, 109–120. [PubMed: 23908124]

Hajarizadeh B, Grady B, Page K, Kim AY, McGovern BH, Cox AL, Rice TM, Sacks-Davis R, Bruneau J, Morris M, Amin J, Schinkel J, Applegate T, Maher L, Hellard M, Lloyd AR, Prins M, Dore GJ, Grebely J, 2015 Patterns of hepatitis C virus RNA levels during acute infection: the InC3 study. PloS one 10, e0122232. [PubMed: 25837807]

Ip CL, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, Leggett RM, Eccles DA, Zalunin V, Urban JM, Piazza P, Bowden RJ, Paten B, Mwaigwisya S, Batty EM, Simpson JT, Snutch TP, Birney E, Buck D, Goodwin S, Jansen HJ, O'Grady J, Olsen HE, 2015 MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research 4, 1075. [PubMed: 26834992]

Janes H, Herbeck JT, Tovanabutra S, Thomas R, Frahm N, Duerr A, Hural J, Corey L, Self SG, Buchbinder SP, McElrath MJ, O'Connell RJ, Paris RM, Rerks-Ngarm S, Nitayaphan S, Pitisuttihum P, Kaewkungwal J, Robb ML, Michael NL, Mullins JI, Kim JH, Gilbert PB, Rolland M, 2015 HIV-1 infections with multiple founders are associated with higher viral loads than infections with single founders. Nat Med 21, 1139–1141. [PubMed: 26322580]

Koelle K, Cobey S, Grenfell B, Pascual M, 2006 Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. Science 314, 1898–1903. [PubMed: 17185596]

Kuntzen T, Timm J, Berical A, Lennon N, Berlin AM, Young SK, Lee B, Heckerman D, Carlson J, Reyor LL, Kleyman M, McMahon CM, Birch C, Wiesch J.S.z., Ledlie T, Koehrsen M, Kodira C, Roberts AD, Lauer GM, Rosen HR, Bihl F, Cerny A, Spengler U, Liu Z, Kim AY, Xing Y, Schneidewind A, Madey MA, Fleckenstein JF, Park VM, Galagan JE, Nusbaum C, Walker BD, Lake-Bakaar GV, Daar ES, Jacobson IM, Gomperts ED, Edlin BR, Donfield SM, Chung RT, Talal AH, Marion T, Birren BW, Henn MR, Allen TM, 2008 Naturally occurring dominant resistance mutations to hepatitis C virus protease and polymerase inhibitors in treatment-naïve patients. Hepatology 48, 1769–1778. [PubMed: 19026009]

Kuntzen T, Timm J, Berical A, Lewis-Ximenez LL, Jones A, Nolan B, Schulze zur Wiesch J, Li B, Schneidewind A, Kim AY, Chung RT, Lauer GM, Allen TM, 2007 Viral sequence evolution in acute hepatitis C virus infection. Journal of virology 81, 11658–11668. [PubMed: 17699568]

Li H, Stoddard MB, Wang S, Blair LM, Giorgi EE, Parrish EH, Learn GH, Hraber P, Goepfert PA, Saag MS, Denny TN, Haynes BF, Hahn BH, Ribeiro RM, Perelson AS, Korber BT, Bhattacharya T, Shaw GM, 2012 Elucidation of hepatitis C virus transmission and early diversification by single genome sequencing. PLoS pathogens 8, e1002880. [PubMed: 22927816]

Li H, Stoddard MB, Wang S, Giorgi EE, Blair LM, Learn GH, Hahn BH, Alter HJ, Busch MP, Fierer DS, Ribeiro RM, Perelson AS, Bhattacharya T, Shaw GM, 2015 Single-Genome Sequencing of Hepatitis C Virus in Donor-Recipient Pairs Distinguishes Modes and Models of Virus Transmission and Early Diversification. Journal of virology 90, 152–166. [PubMed: 26468546]

Mitchell AM, Stone AE, Cheng L, Ballinger K, Edwards MG, Stoddard M, Li H, Golden-Mason L, Shaw GM, Khetani S, Rosen HR, 2015 Transmitted/founder hepatitis C viruses induce cell-type- and genotype-specific differences in innate signaling within the liver. mBio 6, e02510. [PubMed: 25714713]

Pandit A, de Boer RJ, 2014 Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. Retrovirology 11, 56–56. [PubMed: 24996694]

Parker J, Rambaut A, Pybus OG, 2008 Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 8, 239–246.

Ribeiro RM, Li H, Wang S, Stoddard MB, Learn GH, Korber BT, Bhattacharya T, Guedj J, Parrish EH, Hahn BH, Shaw GM, Perelson AS, 2012 Quantifying the diversification of hepatitis C virus (HCV) during primary infection: estimates of the in vivo mutation rate. PLoS pathogens 8, e1002881. [PubMed: 22927817]

Rodrigo C, Eltahla AA, Bull RA, Grebely J, Dore GJ, Applegate T, Page K, Bruneau J, Morris MD, Cox AL, Osburn W, Kim AY, Schinkel J, Shoukry NH, Lauer GM, Maher L, Hellard M, Prins M, Estes C, Razavi H, Lloyd AR, Luciani F, 2016 Historical Trends in the Hepatitis C Virus Epidemics in North America and Australia. The Journal of infectious diseases 214, 1383–1389. [PubMed: 27571901]

Rodrigo C, Eltahla AA, Bull RA, Luciani F, Grebely J, Dore GJ, Applegate T, Page K, Bruneau J, Morris MD, Cox AL, Osburn W, Kim AY, Shoukry NH, Lauer GM, Maher L, Schinkel J, Prins M, Hellard M, Lloyd AR, In CC, 2017 Phylogenetic analysis of full-length, early infection, hepatitis C virus genomes among people with intravenous drug use: the InC3 Study. J Viral Hepat 24, 43–52.

Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI, 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. Journal of virology 73, 10489–10502. [PubMed: 10559367]

Sheridan I, Pybus OG, Holmes EC, Klenerman P, 2004 High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. J Virol 78, 3447–3454. [PubMed: 15016867]

Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, Beerenwinkel N, 2013 Probabilistic Inference of Viral Quasispecies Subject to Recombination. Journal of Computational Biology 20, 113–123. [PubMed: 23383997]

Zagordi O, Däumer M, Beisel C, Beerenwinkel N, 2012 Read length versus Depth of Coverage for Viral Quasispecies Reconstruction. PloS one 7, e47046. [PubMed: 23056573]

Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N, 2010 Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. J Comput Biol 17, 417–428. [PubMed: 20377454]

Highlights

- Approximately one in four new HCV infections originates from a single T/F virus
- When using paired end short read NGS technology, cross validation of haplotype reconstruction across multiple windows of the genome with different algorithms reduce the risk of overestimating single T/F variants
- Phenotypic characterization of T/F variants in HCV is a priority for vaccine design

Table 1.

Characteristics of the early acute infection subjects[*] included in the analysis (n-54)

Characteristic	Number (%)/Median
Gender	
Male	33 (61)
Female	18 (34)
Missing	3 (6)
Median Age (years)	23.9
Genotype	
1a	32 (57)
2a/b	5 (9)
3a	16 (31)
4a	1 (2)
Ethnicity	
White Caucasian	41 (76)
Others	10 (19)
Missing	3 (6)
Past history of incarceration	
Yes	21 (39)
No	5 (9)
Missing	28 (52)
Most frequently injected drug	
Heroin	15 (28)
Other	8 (15)
Missing	31 (57)
Host IL-28B genotype	
CC	22 (41)
Others	15 (28)
Missing	17 (31)
Outcome of infection	
Spontaneous clearance	9 (17)
Chronic infection	33 (61)
Missing	12 (22)

[*] each subject contributed a single viraemic sample for the study

Table 2.

Comparison of results of different algorithms across three different regions of the genome for early acute samples

Genomic region	ShoRAH		Quasi-recomb	
	Single T/F*	High diversity	Single T/F	High diversity
Core - E2	17	37	26	28
1a*	22%	78%	47%	53%
3a*	50%	50%	44%	66%
NS3	21	33	25	29
1a	31%	69%	41%	59%
3a	56%	44%	50%	50%
NS5A	23	31	29	25
1a	44%	56%	56%	44%
3a	37%	63%	44%	56%

*1a (n-32), 3a (n-16)

Table 3.

Comparison of T/F polypeptide sequences with the consensus sequence of circulating chronic variants

	Genotype 1a comparison[*]		Genotype 3a comparison[*]	
Sequence region	**Consensus of single T/F variants[**]**	**Individual T/F single variants[**]**	**Consensus of single T/F variants[**]**	**Individual T/F single variants[**]**
Core-E2 (minus HVR1)	14 (2.0%)	25–39 (3.5 – 5.4 %)	7 (0.9%)	16–40 (2.2 – 5.6 %)
NS3	5 (0.8%)	6–15 (1.1 – 2.4%)	2 (0.3%)	8–13 (1.3 – 2.1%)
NS5A	5 (1.1%)	8–24 (1.8 – 5.4 %)	2 (0.4%)	10–19 (2.2 – 4.2%)

[*] Comparison with consensus of 330 and 240 chronic sequences for genotypes 1a and 3a respectively

[**] Quantified as the number of pairwise mismatches in an alignment using the Blosum62 cost matrix (% pairwise difference)