Infection, Genetics and Evolution 67 (2019) 136–144

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research Paper

Using high-throughput sequencing for investigating intra-host hepatitis C evolution over long retrospective periods

Caporossi A.[a,b,*,1], Kulkarni O.[a,1], Blum MGB[a], Leroy V.[c], Morand P.[d,e], Larrat S.[d,e], François O.[a]

[a] Université Grenoble Alpes, TIMC-IMAG, CNRS UMR, 5525 Grenoble, France
[b] Pôle Santé Publique, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France
[c] Clinique d'Hépatologie, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France
[d] Institut de Biologie Structurale, UMR2075 CEA-CNRS-UGA, Grenoble, France
[e] Laboratoire de Virologie, Pôle Biologie, Centre Hospitalier Universitaire Grenoble Alpes, Grenoble, France

ARTICLE INFO

Keywords:
Hepatitis C Virus
Multiple infections
High-throughput sequencing
Longitudinal study
Genetic heterogeneity
Selective sweeps

ABSTRACT

Collections of biological samples held by hospitals represent invaluable resources for conducting retrospective evolutionary studies of chronic infections. Using high-throughput sequencing, those collections permit analysis of within-host genetic diversity over long follow-up periods, and allow a better understanding of resistance to treatment regimes during disease evolution. Here, we studied the evolution of hepatitis C virus (HCV) populations in two patients with an absence of response to dual therapies. We implemented amplicon sequencing to survey genomic variation at the Core and NS5B regions of HCV over a period of 13 years from blood samples obtained at multiple time points. We observed mixed infection by multiple HCV genotypes in both patients. Genetic heterogeneity and sample composition analysis provided information about the changes in viral population over the course of clinical treatment, with NS5B experiencing an increase in diversity after treatment initiation. Secondary infections were estimated to predate treatment year, and our results pointed towards diversifying selection occurring post-treatment, acting on standing genomic variation and maintaining high genetic heterogeneity during infection. For these two patients infected with multiple HCV genotypes, the maintenance of viral diversity was explained with the hypothesis of soft selective sweep started at the same time as antiviral treatment was initiated.

1. Introduction

Patient sample collections held by public hospitals contain an invaluable source of data to answer questions about pathogen evolution and human health (Andersen et al., 2015; Crawford et al., 2015; Holmes et al., 2016). In recent years, high-throughput sequencing (HTS) technologies have enabled studies of viral populations over the course of infection by allowing the rapid acquisition of thousands of short RNA sequences from multiple time-point samples (Nelson and Hughes, 2015). HTS has become an essential tool in describing genetic heterogeneity and in understanding intra-patient evolutionary dynamics of viral species (Beerenwinkel et al., 2012). More specifically HTS has been applied to describe within-host diversity of HCV (Lauck et al., 2012; Li et al., 2011; Ramachandran et al., 2011), to detect transmission bottlenecks in infection (Wang et al., 2010), to predict response to triple therapy in chronic HCV patients and evaluate resistance variants of protease inhibitors (Larrat et al., 2015a), and to study the response of HCV to selection imposed by antiviral drugs (Ahmed and Felmlee, 2015). In spite of the importance of understanding within-host evolutionary dynamics of HCV, surveys of intra-host evolution of genotypes of chronically infected patients along extended periods of time have, however, remained relatively infrequent (Culasso et al., 2014; Raghwani et al., 2016). In this study, HTS were used to conduct a retrospective analysis of virus evolution from blood samples of two chronically infected HCV patients.

During the course of infection, HCV evolves rapidly and results in quasi-species (Eigen et al., 1988; Gregori et al., 2013). The quasi-species composition of HCV populations allows the virus to escape the immune response and develop resistance to treatment leading to chronic infection (Bull et al., 2011; Larrat et al., 2015b; Lauring et al., 2013; Pellerin et al., 2004; Vermehren and Sarrazin, 2012). HCV has been classified into 7 genotypes, which differ from each other from 30% to 33% at the nucleotide level, and 65 subtypes, which differ from each other from 20% to 25% (Simmonds et al., 2005; Smith et al., 2014). The

* Corresponding author at: Laboratoire TIMC-IMAG, Faculté de Médecine, Domaine de la Merci, 38706, La Tronche cedex, France.
E-mail address: acaporossi@chu-grenoble.fr (A. Caporossi).
[1] These authors contributed equally to the study.

https://doi.org/10.1016/j.meegid.2018.11.004
Received 26 February 2018; Received in revised form 11 September 2018; Accepted 2 November 2018
Available online 03 November 2018
1567-1348/

evolution of drug resistance occurs by the fixation of specific drug-resistance mutations, and HCV genotypes and subtypes confer differential resistance to multiple therapies. Fifteen years ago, HCV treatments were based on dual therapies with interferon-alpha and ribavirin. For these treatments, genotypes 1 and 4 were known to be more resistant than genotypes 2 and 3 to the antiviral action of interferon-alpha and ribavirin (Pawlotsky, 2009). However the mechanisms by which resistance occurred in patients with mixed infections from multiple genotypes remained unknown (Schröter et al., 2003). It had been hypothesised that resistance to treatment in RNA viruses could involve either hard or soft selective sweeps (Feder et al., 2016; Pennings et al., 2014). In a classic or 'hard' sweep, a resistance mutation arises in a single viral particle, and ultimately reaches fixation in the entire intra-patient population (Burke, 2012; Smith and Haigh, 1974). As this mutation spreads to the whole population, background mutations linked to it also increase in frequency, and reduce genetic diversity in the population. A 'soft' selective sweep, by contrast, describes the dynamics of selection acting on mutations present on many particles in a population with standing genetic variation (Hermisson and Pennings, 2005). Resistance mutations originate multiple times on different genetic backgrounds, and selection can increase overall genetic diversity. Soft sweeps have been identified in HIV (Messer and Petrov, 2013; Pennings, 2012), but this mode of selection has not been described in longitudinal studies of HCV patients.

To understand chronic infection in non-responder patients treated with dual therapy in the 2000's, we performed deep sequencing on two regions of the HCV genome, Core (464 bp) and NS5B (381 bp). Multiple genotypes and subtypes were detected in both patients. By using RNA sequences from multiple time points, the evolution of genetic diversity was reconstructed over a period of time greater than thirteen years. The evolutionary dynamics of viral populations were studied by using measures of nucleotide diversity, exploratory data analysis, coalescent modelling and phylogenetic methods. Phylogenetic reconstruction on temporal samples pinpointed population structure and complex evolutionary dynamics of viral populations during infection. For these two patients, our results supported the hypothesis of soft selective sweeps underpinning intra-host evolution of multiple genotypes of HCV over long time scales.

2. Materials & methods

2.1. Patients under study & sample preparation and sequencing

Two patients followed at Grenoble-Alpes University Hospital for a chronic HCV infection were included in the study. The patients had a known date of infection because of an identified transmission event due to transfusion or professional exposure. The year of primary infection was 1979 for patient 1, and 1970 for patient 2. Both patients were treated with dual therapies based on pegylated interferon and ribavirin. Patient 1's treatment had been administered for six months from January to June 2003. Patient 2's treatment had been administered for seven months from March to September 2008. The number of temporal samples per patient was equal to 5 and 8, and a total of thirteen serum samples were available over follow-up periods of 10 and 13 years (Table 1). Informed consents for participating in research were obtained from the patients and study was retrospectively carried out using a biobank numbered DC 2008–680.

Table 1
Patient and treatment information.

Patient id	Infection year[a]	Treatment duration[b]	Number of samples	Range of samples
1 DEHE	1979	Jan 2003 June 2003	8	2002–2014
2 AMCA	1970	Mar 2008 Oct 2008	5	2005–2014

[a] Infection by genotype 1 from blood transfusion or professional exposure.
[b] Dual therapy: pegylated interferon and ribavirin.

The samples were stored at −80 °C and were retrospectively analysed. HCV RNA was extracted from 1 ml of plasma using EasyMAG (bioMérieux, Marcy l'Étoile, France) with an elution volume of 25 µl. Extracted RNA (15 µl) was purified using Turbo DNase Ambion (Life Technologies, Cergy Pontoise, France), and cDNA was synthesised with the AccuScript high-fidelity kit (Agilent, Garches, France) using random primers. For nucleotide sequencing, the 454 GS Junior platform was used (Roche Diagnostics). The Core region (nucleotides [nt] 288 to 751 according to reference HCV strain H77) (Pham et al., 2009) and the NS5B (nt 8256–8636)(Sandres-Saune et al., 2003) region were amplified using Phusion Hot Start II (Finnzyme, Illkirch, France) and primers described in (Pham et al., 2009; Sandres-Saune et al., 2003) which were added universal M13 tails. The Polymerase Chain Reaction (PCR) cycling conditions were initial denaturation at 95 °C for 15 min followed by 50 cycles of denaturation at 94 °C for 30 s, annealing at 64 °C for Core and 57 °C for NS5B 30s and elongation at 72 °C for 70s with a final step of extension at 72 °C for 10 min. Sample-specific multiplex identifier (MID) sequences associated with universal M13 sequence were added with 25 cycles of PCR (56 °C 15 s, 72 °C 20s) using the same Taq polymerase on PCR products 1:50 diluted.

The amplification efficacy was assessed using the Agilent DNA 1000 reagent kit and the Agilent 2100 expert bioanalyzer (Agilent, Les Ulis, France). The expected amplicons were obtained for all the samples. The amplicons were pooled equimolarly and purified using Agencourt AMPure XP reagents (Beckman Coulter, Roissy, France). A library on the amplicon pool was prepared using the GS Junior rapid library prep kit. An emulsion PCR was run with the GS Junior emPCR (Lib-L) kit. Sequencing was performed with a 454 GS Junior PicoTiterPlate to a target depth reading of 1000×. For Core and NS5B regions respectively, a mean of 22,071 +/− 3336 and 9209 +/− 969 reads per nucleotide was obtained with a median length of 579 nt and 392 nt.

Sequence data preparation

The SFF files obtained from the 454 GS Junior were converted to the standard FASTQ format using the Roche supplied sffinfo tool, a part of the GS Data Analysis Package (http://www.454.com/products/analysis-software/). Cutadapt (Martin, 2011) was used for adapter removal. Corrections of 454 error modes such as homopolymer indels and carry forward/incomplete extension (CAFIE) were performed using RC454 (Henn et al., 2012) with a consensus assembly for each region (Core/NS5B) generated by VICUNA (Yang et al., 2012). Additional filtering by read sequence length (300 bp) was performed. Duplicate sequences were removed from the final alignments, and their multiplicity values per patient and per time point were recorded. The Core region sequences, which have been previously reported as being highly conserved ($\mu = 0.28–0.43 \times 10^{-3}$ substitutions/site/year (Gray et al., 2011)), were used as control and compared to GENBANK in order to estimate a rate of sequencing error.

2.2. Sample composition and genetic heterogeneity analysis

Sampled genotypes and subtypes were identified by using the HCV BLAST align tool on the HCV database web site (Kuiken et al., 2005). For all samples, the frequencies of identified genotypes or subtypes were computed from the BLAST outputs. Average identity scores, computed as the percentage of each identified sequence having the same residues at the same positions as the closest BLAST output sequence, were reported for all phylogenetic groups identified. Nucleotide Diversity Estimates (NDEs) were obtained for each region and each time point by using the R package pegas (Paradis, 2010; R Core Team, 2017). In addition, we estimated the median percentage of nucleotides that differ between pairs of sequences at each time point. This measure is similar to NDE, but more robust to multiple alignment errors. NDEs and median estimates were computed before and after separating

identified genotypes or subtypes in patient samples. Divergence rates relative to the baseline were computed as the percentage of nucleotides that differ between pairs of sequences in the baseline sample and at each time point.

Within-patient viral population genetic structure and temporal evolution were also investigated by using principal component analysis (PCA) on polymorphic RNA sites. PCA were performed with the R program 'prcomp' (R Core Team, 2017). Principal components were computed before and after separating identified genotypes or subtypes in patient samples. Because PCA is known to be affected by unbalanced designs (McVean, 2009), the patient data were subsampled in order to retain around 100 to 200 reads at each time point.

2.3. Phylogenetic trees

Global phylogenetic trees including all unique reads were obtained using the neighbour-joining (NJ) algorithm (Saitou and Nei, 1987). Figtree was used to obtain graphical tree representations (http://tree.bio.ed.ac.uk/software/figtree). The NJ trees were used for checking the clock-likeness of each data set using Path-O-Gen/TempEst, which performs a regression of root-to-tip genetic distance on sampling times (Rambaut et al., 2016). Only the trees that passed the criterion of having a positive correlation between divergence and sampling time points were considered for assessing maximum clade credibility trees from subsequent coalescent analyses. For patient 1 who was infected by multiple genotypes, separate trees for each genotype were used for assessing the clock-likeness criterion. Phylogenetic trees were also obtained using the maximum likelihood (ML) method. PhyML 3.1 was used to compute the trees with default parameters (Guindon et al., 2010). To minimise the effect of having large differences in read numbers, 100 unique reads per time point were analysed after multiple sequence alignment using MAFFT 7 with the option "adjustdirection" to comply with direction of nucleotide sequences (Katoh and Standley, 2013).

2.4. Coalescent analysis

The BEAST software (v1.8.1) (Drummond et al., 2012) was used for a Bayesian Markov Chain Monte Carlo (MCMC) analysis of the demographic parameters of HCV samples for patients 1 and 2. Each replicate included 20 sequences from each time point. These sequences were obtained by subsampling unique reads for each time point. Subsampling was performed using the 'subsample_fasta.py' script in the QIIME package (Caporaso et al., 2010). The final sequence alignment for BEAST was obtained using MAFFT (Katoh and Standley, 2013) on default settings. Molecular substitution model selection was performed using jModelTest (Darriba et al., 2012) on default parameters. The models with the optimal values of the Bayesian information criterion were chosen as parameters in BEAST runs. The HKY + G substitution model (Hasegawa et al., 1985) was the most suitable model across all data sets. Regarding the molecular clock model, relaxed models did not converge, and a strict model was used for all runs. For computing demographic parameters (TMRCA, Time since the most recent common ancestor), the tips were calibrated using the 'Guess dates' option (Drummond et al., 2005). All priors of the Bayesian analysis were set at their default values except for the clock rate parameter, which was assigned a weakly informative gamma distribution with shape 0.001 and scale 1000. Each replicate was assigned a chain length of 20 million steps. When MCMC runs did not achieve convergence, the chain lengths were doubled. Replicates achieving no convergence were discarded. Tracer v1.5 was used for post-run analysis. Convergence was checked *via* posterior effective sample size values being > 200.

2.5. Molecular-clock analysis

In addition to the coalescent analysis, LSD (To et al., 2015) a simple and fast molecular-clock model based on a Gaussian model was applied to date ancestral events by using least squares criterion. We used the ML trees as input of LSD as well as the sampling dates of the sequences for temporal constraints on the tips. We performed the analysis with the temporal precedence constraint version of the algorithm to impose the constraint that the date of every node is equal or more ancient than the dates of its descendants. We computed confidence intervals for mutation rates and for dates by replicating the analysis on 100 trees.

3. Results

3.1. Sample composition and genetic diversity analysis

Using high-throughput sequencing, a total number of 15,234 RNA sequences for the Core and NS5B regions were obtained at distinct time points of HCV infection for two patients. For each patient sample, the allele frequency spectrum exhibited a large excess of uniquely represented sequences (90%), and a total number of 14,494 reads (95%) occurred less than three times in each sample. For each sample, identity with previously classified genotypes was computed for the NS5B and Core regions by using the BLAST align tool on the HCV database (Kuiken et al., 2005). Identity scores ranged between 93.3% and 99.1% (Tables 2 and 3). For the control region (Core region), the average identity score of genotype 1b sequences was 98.2%. All scores were above thresholds considered in subtype classification (Simmonds et al., 2005), showing that genotype or subtype assignment was unambiguous for all identified sequences.

For patient 1, the sample composition at NS5B indicated a co-occurrence of two genotypes, 1b and 4/4 k (genotype 4 and subtype 4 k) (Fig. 1C, Table 2). The frequency of genotype 4/4 k increased from 0% in 2002 (before treatment) to 68.89% during the year of treatment. After treatment failure, genotype 4 increased in frequency to reach 99.1% in 2014, but those genotypes were not found in the 2010 sample. The results for the Core region showed a pattern similar to NS5B (Fig. 1A, Table 2). Consistently with the fact that Core is less divergent in terms of genotype separation than is NS5B, variation was less pronounced for the Core region. The frequency of genotype 4/4 k increased from 0% in 2002 (before treatment) to 4.97% during the year of treatment. After treatment failure, genotype 4 and subtype 4 k also increased in frequency at the Core region, and reached 55.8% in the

Table 2
Genotype composition for the NS5B and Core regions (Patient 1).

Year	Number of reads	% Genotype		% Identity (Var)[b]			
		1b	4/4 k	1b		4/4 k	
NS5B							
2002[a]	622	**99.9**	–	95.64	(0.96)	–	
2003[a]	1031	31.10	**68.89**	95.88	(1.31)	93.80	(0.22)
2005[a]	927	51.82	48.01	96.06	(0.98)	93.69	(0.25)
2006[a]	626	**82.58**	17.25	95.85	(0.70)	93.85	(0.16)
2007	257	15.95	**84.05**	95.73	(0.70)	93.38	(0.51)
2008	239	8.78	**91.22**	95.61	(1.94)	93.30	(0.41)
2010	240	**100**	–	96.10	(0.40)	–	
2014	242	0.08	**99.10**	95.50	(0.50)	92.77	(0.22)
Core							
2002[a]	542	**99.44**	–	98.87	(0.50)	–	
2003	181	**95.02**	4.97	98.32	(0.70)	97.33	(1.00)
2005	610	**100**	–	99.06	(0.58)	–	
2006	510	**100**	–	98.46	(0.47)	–	
2007	595	**97.82**	2.18	98.89	(0.66)	96.30	(1.56)
2008	396	**86.62**	13.38	98.16	(1.17)	97.24	(0.61)
2010[a]	432	**99.76**	–	97.89	(0.64)	–	
2014	523	44.16	**55.83**	97.80	(2.39)	96.94	(0.78)

[a] This sample had unclassified genotypes (< 2%) not shown.
[b] Identity scores obtained from BLAST in HCV align tool. Sample variances are indicated.

Table 3
Genotype composition for the NS5B and Core regions (patient 2).

Year	Number of reads	% Subtype		% Identity (Var)[b]			
		1a	1b	1a		1b	
NS5B							
2005	194	–	100	–	–	96.63	(0.30)
2007[a]	152	–	98.02	–	–	95.80	(1.05)
2008	93	–	100	–	–	96.11	(0.32)
2010[a]	683	18.44	81.25	96.20	(2.70)	94.96	(1.63)
2014[a]	420	18.33	81.4	96.22	(2.85)	95.46	(1.22)
Core							
2005	525	–	100	–	–	98.84	(0.41)
2007	494	–	100	–	–	99.01	(0.18)
2008	116	–	100	–	–	98.45	(0.56)
2010*	699	18.16	81.54	96.22	(2.36)	95.17	(1.76)
2014	397	–	100	–	–	98.04	(0.46)

[a] This sample had unclassified genotypes (< 5%) not shown.
[b] Identity score obtained from BLAST in HCV align tool. Samples variances are indicated.

2014 sample. Overall, the results supported a multiple infection hypothesis in which patient 1 was infected from two genetically distinct genotypes.

For patient 1, nucleotide diversity estimates (NDEs) for the Core region remained at a constant level during infection (NDE = 0.029, sd = 0.014, Fig. 1B). In contrast, the NDE values for the NS5B region exhibited a sharp increase at the onset of treatment with dual therapy during the year 2003 (NDE = 0.14, sd = 0.0009, Fig. 1D). This sudden increase coincided with the occurrence of genotype 4/4 k during the period of treatment, and it was followed by a gradual decrease after treatment (NDE = 0.023, sd = 0.006 in 2014). Considered separately, genotypes 1b experienced a decrease in diversity, whereas NDEs remained at more stable values for genotype 4/4 k (Fig. 1D). To better evaluate genetic heterogeneity for each genotype, we computed the distribution of the percentages of nucleotide differences in each time sample (Fig. S1). For genotype 1b, the median percentage was around 3% at baseline, and dropped to around 1% after treatment (Fig. S1A). For genotype 4/4 k, the median percentage of nucleotide differences fluctuated in the range (0.75%, 1.5%), and provided no evidence for variation through time (Fig. S1C).

For patient 2, the results at NS5B detected the presence of 2 subtypes of genotype 1, subtypes 1a and 1b (Fig. 2C, Table 3). The frequency of subtype 1a increased from 0% in 2005 (before treatment) to 18.4% in 2010/2014 after treatment. The results for Core indicated that this region was similar to subtype 1b, except for the 2010 sample for which the frequency of subtype 1a was around 18.2% (Fig. 2A, Table 3). The results supported a multiple infection hypothesis in which patient 2 was infected from subtype 1a and subtype 1b HCV strains. Genetic heterogeneity at the Core region remained at a constant level during infection (NDE = 0.011, sd = 0.003, Fig. 2B), except for the sample from year 2010. For the 2010 sample, substantial levels of NDE were observed (Fig. 2B, NDE = 0.093, sd = 0.001). For the NS5B region, genetic heterogeneity increased after treatment with dual therapy from NDE = 0.026 (2008, sd = 0.001) to NDE = 0.10 (2010, sd = 0.02), and it decreased to a lower level in 2014 (Fig. 2D). For genotype 1b, the median percentage of nucleotide differences was around 2.8% before treatment. It dropped to 1.8% during treatment and increased to > 4% when treatment was stopped (Fig. S1B).

3.2. Population structure analysis

RNA sequences from the NS5B region were used to produce principal component (PC) plots for all time samples after correction for uneven sampling (Fig. S2). For patient 1 and patient 2, the PC plots reproduced the temporal evolution of genotype and subtype frequencies obtained in Figs. 1 and 2 respectively. For patient 1, the first axis clearly separated genotype 1 strains from genotype 4 strains. Axis position in each time sample strongly agreed with sample composition. The second

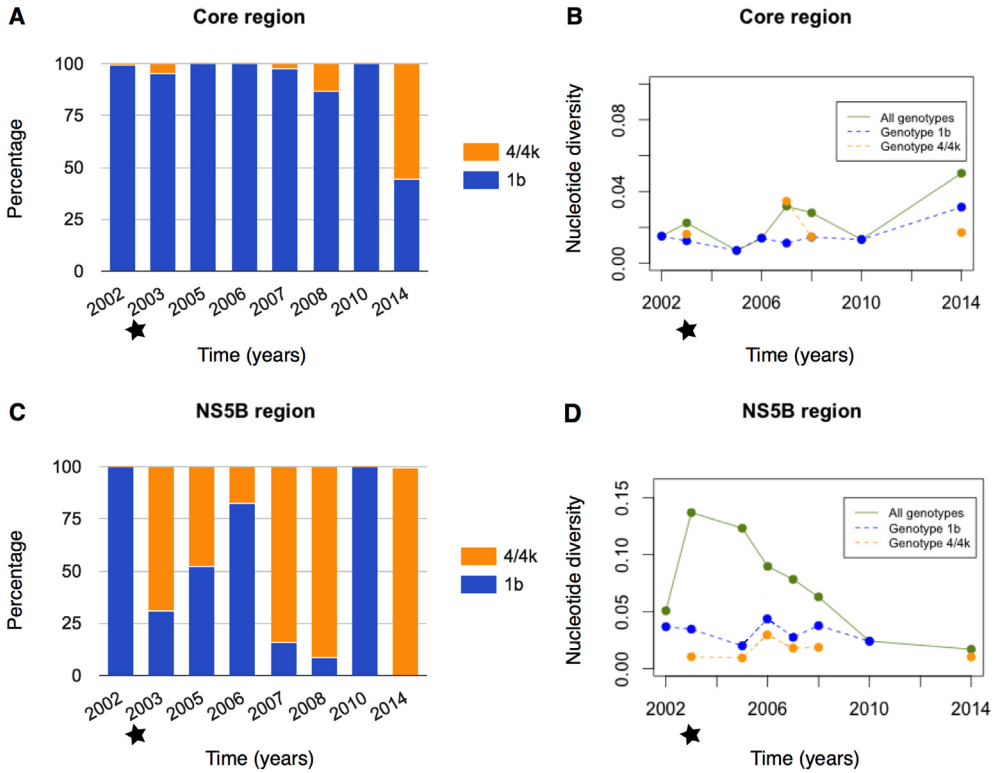

Fig. 1. Sample composition and genetic heterogeneity for patient 1. (A–B) Core region: Percentage of genotypes 4 and 4 k and nucleotide diversity estimates for each time sample. (C–D) NS5B region: Percentage of genotypes 4 and 4 k and nucleotide diversity estimates for each time sample. Treatment years are indicated by a star.

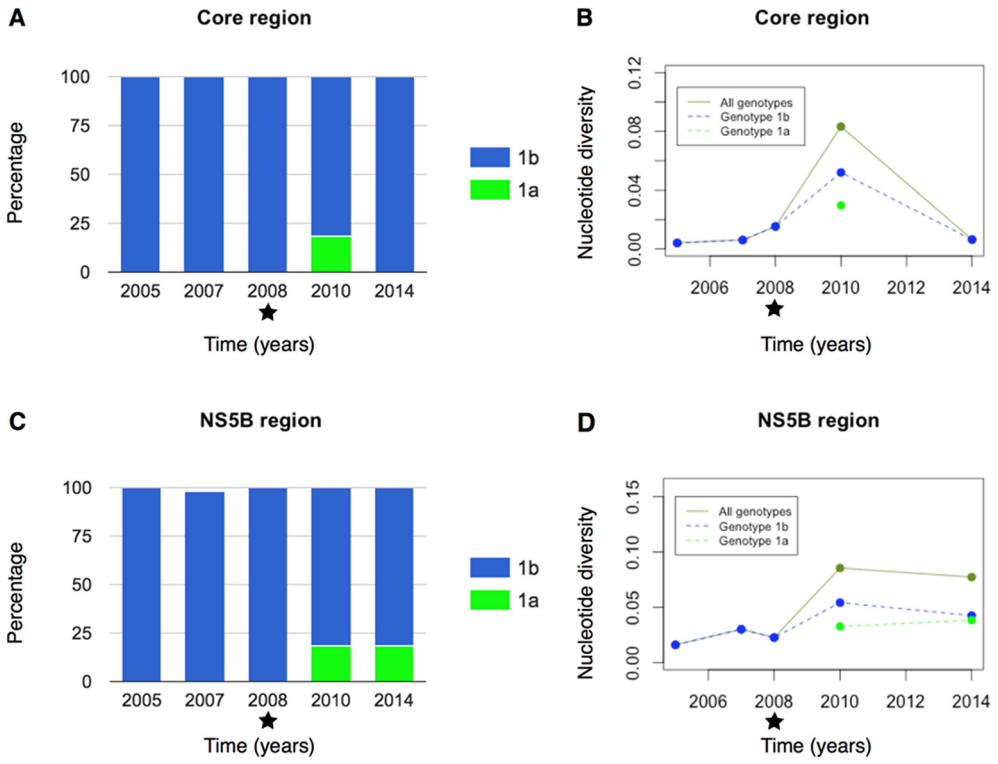

Fig. 2. Sample composition and genetic heterogeneity for patient 2. (A–B) Core region: Percentage of genotype 1b and nucleotide diversity estimates for each time sample. (C–D) NS5B region: Percentage of genotype 1b and nucleotide diversity estimates for each time sample. Treatment years are indicated by a star.

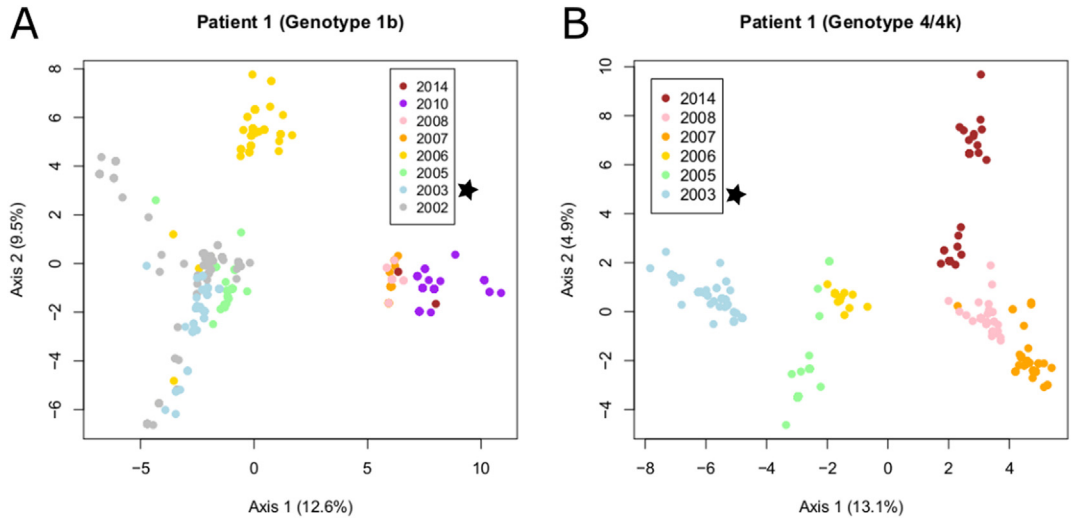

Fig. 3. Principal component plots for temporal samples from the NS5B region (patient 1). (A) PC plot for genotype 1b (B) PC plot for genotype 4/4 k. Treatment year is indicated by a star.

axis separated samples according to their sampling dates. When we carried out PCA for each genotype separately, the projections of genotype 1b samples before and during treatment (years 2002 and 2003) exhibited more dispersion than the samples after treatment (years 2007 and later, Fig. 3A). For genotype 4/4 k, this result was not observed, and all temporal samples exhibited similar levels of dispersion in the projections (Fig. 3B). For patient 2, the first axis of the PCA separated the strains from subtype 1a and subtype 1b. Strains obtained prior to treatment were grouped together, and clustered separately from sequences obtained post-treatment. When we carried out PCA for genotype 1b only, the projections of samples before and during treatment (years 2005–2008) exhibited less dispersion than the samples after treatment (years 2010–2014, Fig. 4). All PCA results were consistent with those obtained with the analysis of NDEs and percentages of nucleotide difference.

3.3. Phylogenetic analysis

Global NJ trees and coalescent trees for two patients were obtained from the NS5B and Core sequences sampled during the course of infection. Results for patients 1 and 2 are reported in Fig. 5 (NS5B region) and in Fig. S3 (Core region). Maximum-likelihood trees for a sample consisting of 100 unique reads per time point were also computed for the NS5B and Core regions.

For patient 1, the global phylogenetic NJ tree for NS5B exhibited two major lineages corresponding to the co-existence of genotype 1b

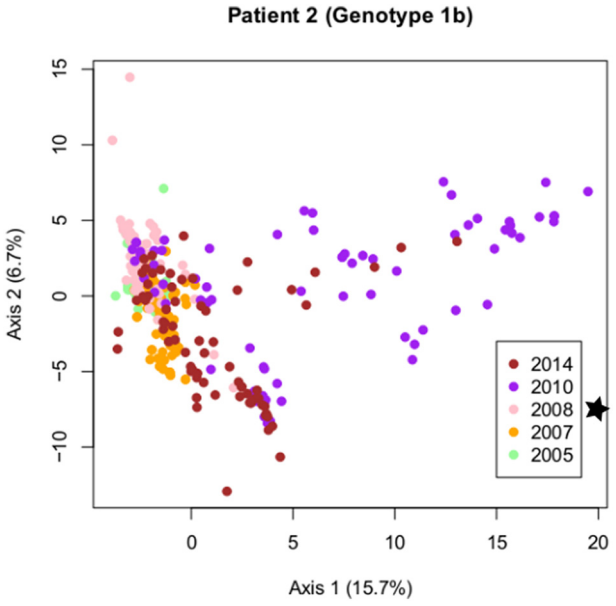

Fig. 4. Principal component plot for temporal samples from the NS5B region (patient 2). PC plot for genotype 1b. Treatment year is indicated by a star.

and genotype 4/4 k (Fig. 5A). Trees of genotype 4/4 k sequence alignments exhibited a divergent clade that contained sequences sampled in year 2007 and all sequences sampled in years 2008 and 2014.

The average divergence between baseline genotype 1b strains and the 2010 strains was about 16%, representing a relative divergence of 400% compared to baseline (Fig. S4A). The average divergence between genotype 4/4 k strains sampled in 2003 and the strains sampled in 2008–2014 was about 15%, and represented a relative divergence of 1500% compared to the first occurrence of this genotype (Fig. S4C). A coalescent analysis for genotype 1b alignments estimated the date of the most recent ancestor for these strains during the year 1983. The 95% highest posterior density interval for this date contained the primary infection date (I = (1979.1, 1985.3)). The analysis of genotype 4/4 k alignments estimated the date of the most recent ancestor for this genotype during the year 1999 (Fig. 5C, Fig. S5A). This date had a 95% highest posterior density interval equal to I = (1998.1, 2001.6) anterior to the treatment date. The molecular-clock analysis using LSD estimated the date of the most recent ancestor of the genotype 4 strains during the year 1995 with a 95% confidence interval (CI) equal to (1989.9, 1997.2) (Fig. S5A). The results of the coalescent analysis and the molecular-clock analysis indicated that secondary infection by genotype 4/4 k after the initial infection by genotype 1b predated the date of treatment.

For patient 2, the BEAST coalescent tree provided results consistent with a global phylogenetic NJ tree reconstructed from all sequences from the NS5B region (Fig. 5B–D). Following treatment initiation, two divergent lineages were observed in the years 2010 and 2014 corresponding to the HCV subtypes 1a and 1b (Fig. 5D, Fig. S5B). The average difference between the baseline strains and subtype 1a strains sampled in 2014 was about 16.5%, and were of the same order of divergence as in patient 1 genotype 4 strains. The divergence between baseline genotype 1b strains and the genotype 1b 2010 strains was about 4–5%, representing a relative divergence of 300% compared to

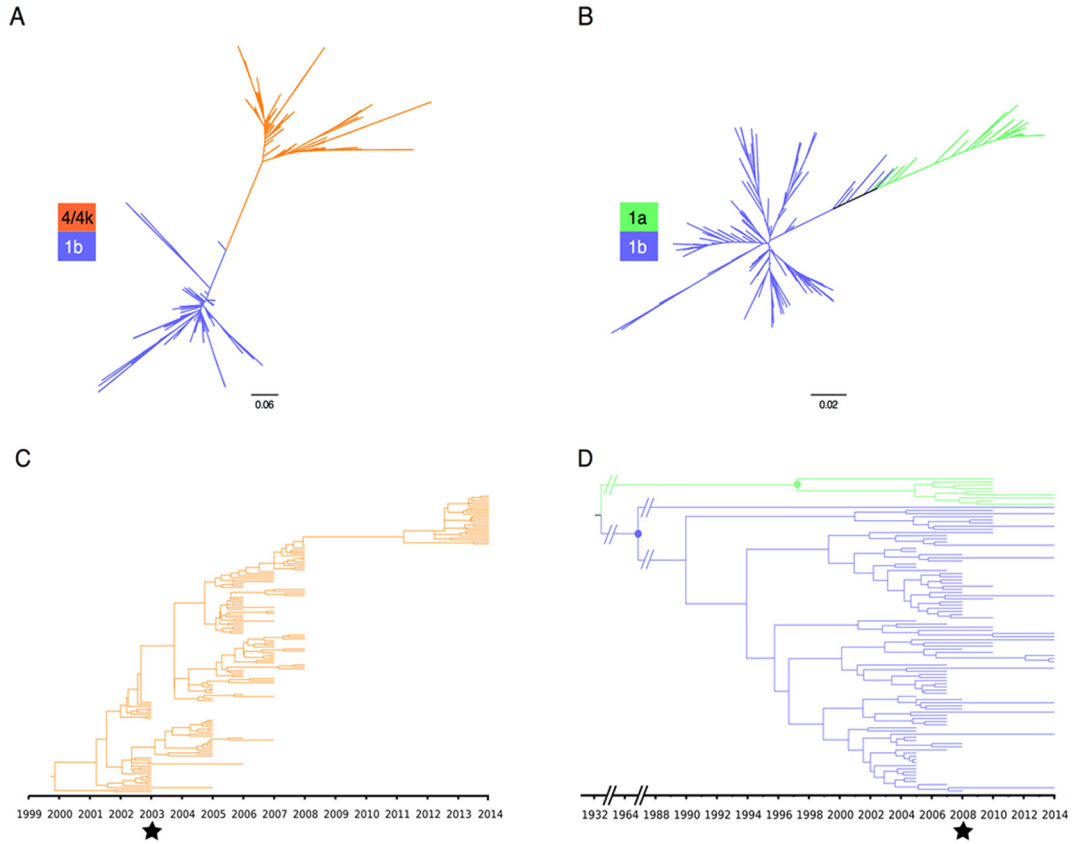

Fig. 5. Phylogenetic trees for the NS5B region. (A) Unrooted NJ tree reconstructed from all sequences from patient 1. (B) Unrooted NJ tree reconstructed from all sequences from patient 2. (C) Maximum clade credibility trees obtained from BEAST by randomly subsampling twenty genotype 4 sequences from 8 time points (patient 1). (D) Maximum clade credibility trees obtained from BEAST by randomly subsampling 25 sequences from 5 time points (patient 2). Dates for the most recent ancestor of subtype 1a (green) and 1b (blue) were shown as dots. Long internal tree branches are represented by '//' symbols. Treatment years are indicated by a star. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

baseline (Fig. S4B). The date of the most recent ancestor for genotype 1b sequences was inferred during the year 1964. The large 95% highest posterior density interval for the date estimate, I = (1940, 1984), included the known infection date (year 1970). The date of the most recent ancestor for subtype 1a sequences was found during the year 1997 with a 95% highest posterior density interval equal to I = (1994, 2003), indicating that secondary infection by subtype 1a was anterior to treatment (Fig. 5D). The molecular-clock analysis using LSD estimated the date of the most recent ancestor for genotype 1b during the year 1966 with a 95% CI equal to I = (1954.7, 1976.2). The date of the most recent ancestor for subtype 1a sequences was found during the year 1996 with a 95% CI equal to (1992.5, 1998.9) (Fig. S5B). For patient 2, the results of the coalescent analysis and the molecular-clock analysis indicated that secondary infection by genotype 1a after the initial infection by genotype 1b also predated the date of treatment.

4. Discussion

Viral populations consist of 'quasi-species' representing a collection of multiple strains of the same virus with small changes in their genomic sequence (Pellerin et al., 2004). Quantifying viral population diversity within a host requires HTS technology capable of sampling a large number of individual sequences accurately (Bernini et al., 2011; Culasso et al., 2014; Nelson and Hughes, 2015). Our study described the evolution of the Core and NS5B regions of the HCV genome in two non-responder patients before and after their treatment with dual therapy in the 2000's. Based on multiple time samples and high-throughput amplicon sequencing, we conducted a retrospective analysis of within-host HCV genetic heterogeneity associated with treatment. For both patients, sample composition, population structure analysis, and nucleotide diversity estimates provided consistent descriptions of the viral population dynamics. Phylogenetic analyses revealed the existence of multiple lineages of HCV corresponding to distinct genotypes (patient 1) or subtypes (patient 2). Both patients were infected by multiple HCV genotypes. Patient 1 was infected by genotype 1b and 4/4 k HCV strains, and patient 2 was infected by genotype 1b and 1a HCV strains. In each patient, genetic heterogeneity at the NS5B region first exhibited a sharp increase, and reached their maximal level after treatment. The diversity peak was followed by a long period during which genetic heterogeneity decreased.

The coalescent (BEAST) and molecular-clock (LSD) analyses of the NS5B region provided consistent estimates of the date of secondary infection for both patients, with higher precision for LSD due to a larger sample size. The most important finding is that secondary infection occurred before treatment administration, and was associated with substantial changes in frequency of viral genotypes after treatment administration. The inferred dates of secondary infection indicated that the two genotypes were present at the onset of the treatment period. For patient 1, the genetic diversity of genotype 1b populations decreased during treatment, as expected with selection against that genotype associated with treatment. In contrast, genotype 4/4 k population genetic diversity remained at constant levels and their relative frequency was substantially increased during treatment. Similar observations were made for patient 2, but the relative frequency of the first genotype (1b) was less impacted by the presence of a second subtype.

An explanation for the results could be soft selective sweeps occurring at the NS5B or at a linked region in the genome, maintaining mixed infection by balancing selection during viral evolution (Hermisson and Pennings, 2005). Soft sweeps differ from classic sweeps as they occur when one or many resistant strains are present in the standing genetic variation prior to treatment, possibly at low frequency (Fig. 6, Pennings, 2012). Classic selective sweeps assume that resistance mutations occur in the viral genome *de novo*, and evolve to fixation in the entire population (Messer and Petrov, 2013). Evidence for standing variation was given by the dates of secondary infection occurring before treatment. It is acknowledged that dual therapies have deleterious effect on genotype 1b during infection and that genotype 4/4 k HCV infection is difficult to cure with pegylated interferon and ribavirin (Yu and Chuang, 2009). Following treatment, genotypes 1b were slowly replaced by genotypes with relatively higher adaptive rates. For patient 2, genotype 1a stayed in minority after treatment. The phylogenetic analyses on the NS5B region confirmed our hypothesis and revealed complex evolutionary dynamics in viral populations.

An obvious limitation of our study is the number of patients in the study ($n = 2$). The patterns of genetic diversity observed are thus highly specific to the studied patients, and the extent to which soft sweeps could occur in general mixed infections remains unknown. Another restriction of our study is the limited genomic coverage of the amplicon sequencing approach. The Core and NS5B amplicon sequences represented only 8–9% of the virus genome, which complicates the identification of molecular targets of selection when soft sweeps occur at linked genes. More extensive sequencing efforts would be necessary to identify resistance mutations or confirm signatures of selective sweeps in regions other than Core or NS5B.

Distinct genomic compartments of HCV can exhibit highly different evolutionary trajectories (Culasso et al., 2014). In agreement with this observation, the trajectories of the NS5B region differed from those of the Core region substantially. The Core region of the HCV genome is known to correspond to a highly conserved structural gene (Cuypers et al., 2015). To test it, we evaluated the genetic divergence for the Core and NS5B regions and for the whole genomes of two genotyped virus (KC248195:1b and EU392171:4 k). Concerning Core, we found a divergence of 13.4% whereas for NS5B, the divergence was evaluated to 26.5%. At the whole genome level, the divergence was evaluated to 29.4%. We therefore concluded that the evolutionary rate of Core is lower than the genome-wide rate, and this could explain why genetic heterogeneity remained at low level during infection in both patients. PCR stochasticity may contribute to skew sequence representation after amplification which could explain the different patterns observed for Core and NS5B (Kebschull and Zador, 2015). Another explanation would be intra-host recombinants. Clinically described recombinants 2 k/1b (Galli and Bukh, 2014), 1b/2b (Uribe-Noguez et al., 2018) and 1a/4o (Sridhar et al., 2018) have been identified recently. Since the Core and NS5B regions lie at both ends of the HCV genomes and the NS2 region is a recombination point, intra-host recombination (1b/4 k for patient 1) is a potential explanation for the differences observed in Core and NS5B frequencies (González-Candelas et al., 2011; Pérez-Losada et al., 2015). To confirm this hypothesis, whole genome sequencing would be necessary to identify a recombination breakpoint.

Evolution of the Core and NS5B regions from HCV populations in two patients with mixed infection was studied using amplicon sequencing of hospital patient samples. Data analysis was not highly specific to the sequencing technology used, and it extends to other recent technologies with high quality and extensive coverage. While it is important in the evolution of drug resistance for HIV (Pennings et al., 2014), evolution through soft sweeps has not been reported for HCV so far. According to Feder et al. (2016), the evidence for soft sweeps occurring in the HCV genome after treatment might be consistent with a low efficiency of the medical treatment. Dual therapy was based on treatments for which the occurrence of non-responder patients has not been clearly explained by viral resistance (Larrat et al., 2015b). In contrast, more efficient direct anti-viral treatments have many naturally occurring resistance mutations, and these treatments could generate harder selective sweeps (Feder et al., 2016). We interpret the observed patterns of genetic diversity and the shape of phylogenetic trees as evidence for evolution through soft selective sweeps in patients with chronic multiple HCV infection. Our results contribute to our understanding of the evolution of HCV infections, and it demonstrates the interest of using hospital patient samples for research in pathogen evolution.

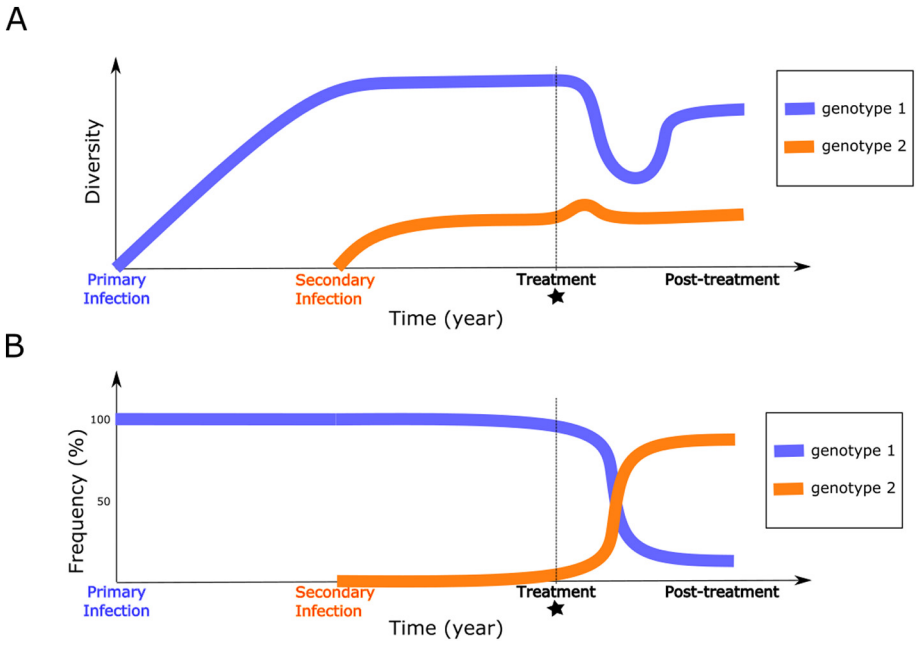

Fig. 6. Soft sweep. (A) Diversity profiles for two genotypes in a mixed infection under weak therapeutic pressure. (B) Frequencies of two genotypes in a mixed infection. The genotype frequency of the secondary infection increases after treatment acting on genotype 1 and is maintained during viral evolution. The secondary infection, present in the standing genetic variation prior to treatment, is associated with resistant strains which slowly replace genotypes of the primary infection with relatively lower adaptive rates.

Acknowledgements, funding

Om Kulkarni was funded by the Marie-Curie Initial Training Network INTERCROSSING (European Commission). The authors acknowledge support from Grenoble Alpes Data Institute (ANR 15-IDEX-0002, France).

Author contributions

A.C. and O.K. wrote the main manuscript text and prepared tables and figures. O.F. designed the analysis. S.L. managed the sample preparation and the sequencing part. M.B., V.L., P.M., S.L. and O.F. reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.11.004.

References

Ahmed, A., Felmlee, D.J., 2015. Mechanisms of hepatitis C viral resistance to direct acting antivirals. Viruses 7, 6716–6729.
Andersen, K.G., Shapiro, B.J., Matranga, C.B., Sealfon, R., Lin, A.E., Moses, L.M., Folarin, O.A., Goba, A., Odia, I., Ehiane, P.E., 2015. Clinical sequencing uncovers origins and evolution of Lassa virus. Cell 162, 738–750.
Beerenwinkel, N., Günthard, H.F., Roth, V., Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3.
Bernini, F., Ebranati, E., De Maddalena, C., Shkjezi, R., Milazzo, L., Presti, A.L., Ciccozzi, M., Galli, M., Zehender, G., 2011. Within-host dynamics of the hepatitis C virus quasi species population in HIV-1/HCV coinfected patients. PLoS One 6, e16551.
Bull, R.A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S.T., Chopra, A., Cameron, B., Maher, L., Dore, G.J., White, P.A., 2011. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog. 7, e1002243.
Burke, M.K., 2012. How does adaptation sweep through the genome? Insights from long-term selection experiments. In: Proc. R. Soc. B. The Royal Society, pp. 5029–5038.
Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335.
Crawford, D.C., Goodloe, R., Farber-Eger, E., Boston, J., Pendergrass, S.A., Haines, J.L., Ritchie, M.D., Bush, W.S., 2015. Leveraging epidemiologic and clinical collections for genomic studies of complex traits. Hum. Hered. 79, 137–146.
Culasso, A.C.A., Baré, P., Aloisi, N., Monzani, M.C., Corti, M., Campos, R.H., 2014. Intra-host evolution of multiple genotypes of hepatitis C virus in a chronically infected patient with HIV along a 13-year follow-up period. Virology 449, 317–327.
Cuypers, L., Li, G., Libin, P., Piampongsant, S., Vandamme, A.-M., Theys, K., 2015. Genetic diversity and selective pressure in hepatitis C virus genotypes 1–6: significance for direct-acting antiviral treatment and drug resistance. Viruses 7, 5018–5039.
Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772.
Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192.
Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.
Eigen, M., McCaskill, J., Schuster, P., 1988. Molecular quasi-species. J. Phys. Chem. 92, 6881–6891.
Feder, A.F., Rhee, S.-Y., Holmes, S.P., Shafer, R.W., Petrov, D.A., Pennings, P.S., 2016. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. elife 5.
Galli, A., Bukh, J., 2014. Comparative analysis of the molecular mechanisms of recombination in hepatitis C virus. Trends Microbiol. 22, 354–364.
González-Candelas, F., López-Labrador, F.X., Bracho, M.A., 2011. Recombination in Hepatitis C Virus. Viruses 3, 2006–2024.
Gray, R.R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A., Pybus, O.G., 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol. Biol. 11, 131.
Gregori, J., Esteban, J.I., Cubero, M., Garcia-Cehic, D., Perales, C., Casillas, R., Alvarez-Tejado, M., Rodríguez-Frías, F., Guardia, J., Domingo, E., 2013. Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One 8, e83361.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.
Henn, M.R., Boutwell, C.L., Charlebois, P., Lennon, N.J., Power, K.A., Macalalad, A.R., Berlin, A.M., Malboeuf, C.M., Ryan, E.M., Gnerre, S., 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8, e1002529.
Hermisson, J., Pennings, P.S., 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169, 2335–2352.
Holmes, E.C., Dudas, G., Rambaut, A., Andersen, K.G., 2016. The evolution of Ebola virus: Insights from the 2013–2016 epidemic. Nature 538, 193.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Kebschull, J.M., Zador, A.M., 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. gkv717.
Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics 21, 379–384.
Larrat, S., Kulkarni, O., Claude, J.-B., Beugnot, R., Blum, M.G., Fusillier, K., Lupo, J., Tremeaux, P., Plages, A., Marlu, A., 2015a. Ultradeep pyrosequencing of NS3 to predict response to triple therapy with protease inhibitors in previously treated chronic hepatitis C patients. J. Clin. Microbiol. 53, 389–397.

Larrat, S., Vallet, S., David-Tchouda, S., Caporossi, A., Margier, J., Ramière, C., Scholtes, C., Haïm-Boukobza, S., Roque-Afonso, A.-M., Besse, B., 2015b. Naturally occurring resistance-associated variants of hepatitis C virus protease inhibitors in poor responders to pegylated interferon-ribavirin. J. Clin. Microbiol. 53, 2195–2202.
Lauck, M., Alvarado-Mora, M.V., Becker, E.A., Bhattacharya, D., Striker, R., Hughes, A.L., Carrilho, F.J., O'Connor, D.H., Pinho, J.R.R., 2012. Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. J. Virol. 86, 3952–3960.
Lauring, A.S., Frydman, J., Andino, R., 2013. The role of mutational robustness in RNA virus evolution. Nat. Rev. Microbiol. 11, 327.
Li, H., Hughes, A.L., Bano, N., McArdle, S., Livingston, S., Deubner, H., McMahon, B.J., Townshend-Bulson, L., McMahan, R., Rosen, H.R., 2011. Genetic diversity of near genome-wide hepatitis C virus sequences during chronic infection: evidence for protein structural conservation over time. PLoS One 6, e19562.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17, 10.
McVean, G., 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5, e1000686.
Messer, P.W., Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28, 659–669.
Nelson, C.W., Hughes, A.L., 2015. Within-host nucleotide diversity of virus populations: insights from next-generation sequencing. Infect. Genet. Evol. 30, 1–7.
Paradis, E., 2010. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 26, 419–420.
Pawlotsky, J.-M., 2009. Therapeutic implications of hepatitis C virus resistance to antiviral drugs. Ther. Adv. Gastroenterol. 2, 205–219.
Pellerin, M., Lopez-Aguirre, Y., Penin, F., Dhumeaux, D., Pawlotsky, J.-M., 2004. Hepatitis C virus quasispecies variability modulates nonstructural protein 5A transcriptional activation, pointing to cellular compartmentalization of virus-host interactions. J. Virol. 78, 4617–4627.
Pennings, P.S., 2012. Standing genetic variation and the evolution of drug resistance in HIV. PLoS Comput. Biol. 8, e1002527.
Pennings, P.S., Kryazhimskiy, S., Wakeley, J., 2014. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. 10, e1004000.
Pérez-Losada, M., Arenas, M., Galán, J.C., Palero, F., González-Candelas, F., 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. Infect. Genet. Evol. 30, 296–307.
Pham, D.A., Leuangwutiwong, P., Jittmittraphap, A., Luplertlop, N., Bach, H.K., Akkarathamrongsin, S., Theamboonlers, A., Poovorawan, Y., 2009. High prevalence of Hepatitis C virus genotype 6 in Vietnam. Asian Pac. J. Allergy Immunol. 27, 153.
Raghwani, J., Rose, R., Sheridan, I., Lemey, P., Suchard, M.A., Santantonio, T., Farci, P., Klenerman, P., Pybus, O.G., 2016. Exceptional heterogeneity in viral evolutionary dynamics characterises chronic hepatitis C virus infection. PLoS Pathog. 12, e1005894.
Ramachandran, S., Campo, D.S., Dimitrova, Z.E., Xia, G., Purdy, M.A., Khudyakov, Y.E., 2011. Temporal variations in the hepatitis C virus intrahost population during chronic infection. J. Virol. 85, 6369–6380.
Rambaut, A., Lam, T.T., Max Carvalho, L., Pybus, O.G., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Sandres-Saune, K., Deny, P., Pasquier, C., Thibaut, V., Duverlie, G., Izopet, J., 2003. Determining hepatitis C genotype by analyzing the sequence of the NS5b region. J. Virol. Methods 109, 187–193.
Schröter, M., Feucht, H.-H., Zöllner, B., Schäfer, P., Laufs, R., 2003. Multiple infections with different HCV genotypes: prevalence and clinical impact. J. Clin. Virol. 27, 200–204.
Simmonds, P., Bukh, J., Combet, C., Deléage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspé, G., Kuiken, C., Maertens, G., 2005. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. Hepatology 42, 962–973.
Smith, J.M., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23–35.
Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., Simmonds, P., 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. Hepatology 59, 318–327.
Sridhar, S., Yip, C.C.Y., Chan, J.F.W., To, K.K.W., Cheng, V.C.C., Yuen, K.-Y., 2018. Impact of inter-genotypic recombination and probe cross-reactivity on the performance of the Abbott RealTime HCV Genotype II assay for hepatitis C genotyping. Diagn. Microbiol. Infect. Dis. 91, 34–37.
Team, R.C., 2017. R: A Language and Environment for Statistical Computing.
To, T.-H., Jung, M., Lycett, S., Gascuel, O., 2015. Fast dating using least-squares criteria and algorithms. Syst. Biol. 65, 82–97.
Uribe-Noguez, L.A., Ocaña-Mondragón, A., Mata-Marín, J.A., Cázares-Cortázar, A., Ribas-Aparicio, R.M., Gómez-Torres, M.E., Gaytán-Martínez, J., Martínez-Rodríguez, M. de la L., 2018. Case report: Identification of recombinant HCV genotype 1b-2b by viral sequencing in two patients with treatment failure, who responded to re-treatment with sofosbuvir and daclatasvir. J. Infect. Chemother. Off. J. Jpn. Soc. Chemother. 24, 928–931.
Vermehren, J., Sarrazin, C., 2012. The role of resistance in HCV treatment. Best Pract. Res. Clin. Gastroenterol. 26, 487–503.
Wang, G.P., Sherrill-Mix, S.A., Chang, K.-M., Quince, C., Bushman, F.D., 2010. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. J. Virol. 84, 6218–6228.
Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M., Zody, M.C., Henn, M.R., 2012. De novo assembly of highly diverse viral populations. BMC Genomics 13, 475.
Yu, M.-L., Chuang, W.-L., 2009. Treatment of chronic hepatitis C in Asia: when East meets West. J. Gastroenterol. Hepatol. 24, 336–345.