Virology 449 (2014) 317–327

Contents lists available at ScienceDirect

Virology

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

Intra-host evolution of multiple genotypes of hepatitis C virus in a chronically infected patient with HIV along a 13-year follow-up period

A.C.A. Culasso [a], P. Baré [b], N. Aloisi [b], M.C. Monzani [b], M. Corti [c,d,e], R.H. Campos [a,*]

[a] Cátedra de Virología, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Argentina
[b] Sección Virología, Academia Nacional de Medicina, Buenos Aires, Argentina
[c] Departamento de Medicina Interna, Orientación Enfermedades Infecciosas, Facultad de Medicina, Universidad de Buenos Aires, Buenos Aires, Argentina
[d] División VIH/sida, Hospital de Infecciosas F.J. Muñiz, Ciudad Autónoma de Buenos Aires, Argentina
[e] Jefe de Infectología, Fundación Argentina de la Hemofilia, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 2 July 2013
Returned to author for revisions
15 August 2013
Accepted 21 November 2013

Keywords:
Hepatits C virus
Intra-host
Evolution
Co-infection
Coalescene

ABSTRACT

The intra-host evolutionary process of hepatitis C virus (HCV) was analyzed by phylogenetic and coalescent methodologies in a patient co-infected with HCV-1a, HCV-2a, HCV-3a and human immunodeficiency virus (HIV) along a 13-year period.

Direct sequence analysis of the E2 and NS5A regions showed diverse evolutionary dynamics, in agreement with different relationships between these regions and the host factors.

The Bayesian Skyline Plot analyses of the E2 sequences (cloned) yielded different intra-host evolutionary patterns for each genotype: a steady state of a "consensus" sequence for HCV-1a; a pattern of lineage splitting and extinction for HCV-2a; and a two-phase (drift/diversification) process for HCV-3a.

Each genotype evolving in the same patient and at the same time presents a different pattern apparently modulated by the immune pressure of the host.

This study provides useful information for the management of co-infected patients and provides insights into the mechanisms behind the intra-host evolution of HCV.

Introduction

In 80% of the cases, Hepatitis C Virus (HCV) causes chronic infections, which may, in turn, lead to the development of severe hepatic complications, including cirrhosis and hepato-cellular carcinoma (Lauer and Walker, 2001). It has been estimated that about 2% of the world population may be chronically infected with this virus (Lavanchi, 2009). Understanding the mechanisms by which HCV causes chronic infections and regulates illness progression may improve the development of new vaccines and therapeutic techniques.

HCV, together with other RNA-genome viruses, is a rapid evolving pathogen (Kühnert et al., 2011). The crucial features of the infection (e.g. immune evasion, therapy resistance and pathogenesis) may be linked to the genetic diversity and in turn evolutionary processes that have shaped the virus (Kuntzen et al., 2007; von Hahn et al., 2007). Several studies have attempted to correlate different diversity indexes with the host immune response (Bull et al., 2011), disease progression (Farci et al., 2006) and response to treatment (Saludes et al., 2013). Currently, the use of coalescence-based methods has improved the possibility to recognize the intra-host diversification process of HCV by analyzing the dynamics of viral lineages in the host over the time (Gray et al., 2012).

Most HCV intra-host evolutionary studies have been carried out in different cases of chronic infection but evolutionary studies of several viral strains in the same host are less common. However, these researches require the analysis of co-infected patients in a long follow-up period. These conditions can be fulfilled by studying elderly patients with hemophilia (Qin et al., 2005) who became infected by the use of non-inactivated human-derived clotting factors. Unfortunately, these patients are often infected with other blood-borne pathogens like the HIV (Lee, 2009).

In this study, we describe the evolution of two different regions of the viral genomes of three HCV genotypes in a single HIV-co-infected patient. Since the three genotypes evolved in a same host, the evolution of each genotype may represent the different mechanisms used by HCV to sustain the chronic infection. The study of multiple variants in a single host may provide useful information for the management of co-infected patients and provide insights into the mechanisms behind the intra-host evolution of HCV.

* Correspondence to: Junín 954, 4th floor, (1113) Buenos Aires, Argentina. Tel.: +54 11 4964 8264.
E-mail address: rcampos@ffyb.uba.ar (R.H. Campos).

0042-6822/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2013.11.034

Results

The patient was co-infected with HCV-1a, 2a and 3a

To describe the genotype composition of each sample, 5 UTR-RFLP and NS5B sequencing analyses were performed. The patient was monitored throughout a period of 13 years. Samples were analyzed at the start of the follow-up in 1993 (time code 4188) and at 1995 (5168), 1997 (5679), 1998 (5974 and 6301), 1999 (6524), 2000 (6772), 2001 (7356), 2002 (7490), 2004 (8429) and 2006 (9107). All samples except sample 8429 (end of treatment) produced the expected 250-bp amplicon from the 5′UTR which allowed the detection of genotypes 1, 2 and 3, alone or in combinations (Table 1). Further subtyping was carried out with the NCBI's Viral Genotyping Tool using NS5B sequences (http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi). The NS5B sequences obtained from samples 5679 and 5974 were classified as subtype 1a, that from sample 9107 as genotype 2a, and that from sample 4188 as genotype 3a. Sample 8429 did not produce any amplification for NS5B. The remaining sequences were not analyzed because their chromatograms showed several mixed positions as may be expected in the presence of multiple genotypes (Table 1). To resolve the mixed viral populations, genotype specific primers were designed covering two regions that showed different evolutionary patterns: the variable (first half of) E2 region and the relatively more conserved (second half of) NS5A one (Supplementary Table 1). The results showed that in most of the samples the patient was co-infected with at least two genotypes (Table 1).

Phylogenetic analysis of direct sequences

To perform a genetic characterization of the different viral strains, phylogenetic analyses were carried out. The sequences obtained with genotype specific primers showed no evidence of inter-genotype recombination (as assessed by the sliding window methodology implemented in the NCBI's Viral genotyping tool).

The phylogenetic analysis using both the E2 and NS5A regions grouped all samples sequenced, according to its primer design, in genotypes 1a, 2a or 3a. Within each genotype, the patient's samples formed clusters with strong statistical support. In addition, the tree topologies suggested a relationship between ancestry and time of sampling (Fig. 1). The clustering and statistical support of the patient's samples was maintained when the analysis was carried out with more related sequences in the GenBank (Supplementary Fig. 2).

The phylogeny of the samples was compatible with its time

To describe in detail the phylogenetic relationship between the sequences of each genotype in both regions (E2 and NS5A), the datasets of direct sequences were analyzed separately, since the experimental design does not allow ruling out inter-genotype recombination between these genomic regions. The model of the nucleotide substitution used for all datasets was HKY (Hasegawa et al., 1985), except for the E2 region of genotype 3a, which was TrN (Tamura and Nei, 1993) with equal base frequencies. For the E2 region of HCV-2a, the rate heterogeneity among sites was modeled as a proportion of invariant sites. As a result of exhaustive tree searches, all datasets showed comb-like trees where the ancestry–descendant relationship was compatible with the sampling time (each sequence was close to the ancestor of the next one), except for the E2 dataset of HCV-1a (Fig. 2).

The branch lengths and the time were correlated

To analyze the pattern observed in the previous analysis quantitatively (i.e. if the sequences carried temporal information), a linear regression analysis was performed. The correlation coefficients for the regressions (r^2) and the slope (a rough estimation of the substitution rate) for the E2 region were 0.0446 (substitution rate not significantly different from 0) for HCV-1a, 0.8516 (substitution rate of 9.90×10^{-3} s/s/y) for HCV-2a, and 0.8478 (substitution rate of 3.53×10^{-3} s/s/y) for HCV-3a (Table 2). In the NS5A regions, the r^2 values were 0.9356 for HCV-2a (substitution rate of 1.17×10^{-3} s/s/y) and 0.8497 for HCV-3a (estimated substitution rate of 1.14×10^{-3} s/s/y) (Table 2). All datasets, except E2 HCV-1a, resulted in high correlation coefficients and slopes significantly different from 0 ($p < 0.05$), indicating temporal structure in these datasets (Fig. 2).

To confirm the slopes obtained by linear regression, a Bayesian coalesence analysis was carried out, yielding almost identical estimations for the substitution rates (data not shown).

Intra-host viral population analysis

To analyze the diversification process, a genetic population study of molecular clones was performed using Coalescence analyses in a Bayesian framework.

The intra-host HCV evolution was analyzed using clones from the E2 region

A total of 307 clones, including at least 14 clones per sample and genotype, were sequenced (Supplementary Table 2).

Since nested PCR conditions are prone to artifactual recombination phenomena, the alignments of cloned sequences were analyzed

Table 1
Summary of the results for the different analyses carried out in each sample. Sample Code: Four-digit time code of the sample (measured as days from an arbitrary date in the past). UTR (RFLP): Genotyping by RFLP analysis of the PCR product for the 5'UTR region. 1, 2 or 3 are the assigned genotype; 1/2: RFLP compatible with a mix of genotypes 1 and 2; NG: complex RFLP that did not allow the assignment of genotype. NS5B: result of the sequence analysis of the NS5B region. 1a, 2a or 3a: are the assigned subtype; Mix: The electropherogram showed many nucleotide positions with multiple base possibilities. E2 and NS5A: Result of the genotype-specific PCR reaction for the E2 and NS5A regions (the genotype specificity is indicated between parentheses). Seq: The PCR product was sequenced; +; the PCR product was not enough to be sequenced; -: the PCR produced no amplicons.

Region	4188	5168	5679	5974	6301	6524	6772	7356	7490	8429	9107
UTR (RFLP)	3	NG	1	1/2	1/2	1/2	2	2	2	–	2
NS5B	3a	Mix	1a	1a	Mix	Mix	Mix	Mix	Mix	–	2a
E2 (1a)	–	Seq	Seq	Seq	Seq	+	–	–	–		–
NS5A (1a)	–	–	Seq	Seq	+	–	–	–	–		–
E2 (2a)	–	–	–	+	Seq	Seq	Seq	Seq	Seq		Seq
NS5A (2a)	Seq	+	Seq	Seq	Seq	Seq	Seq	Seq	Seq		Seq
E2 (3a)	Seq	Seq	–	Seq	Seq	Seq	Seq	Seq	Seq		–
NS5A (3a)	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq	Seq		–

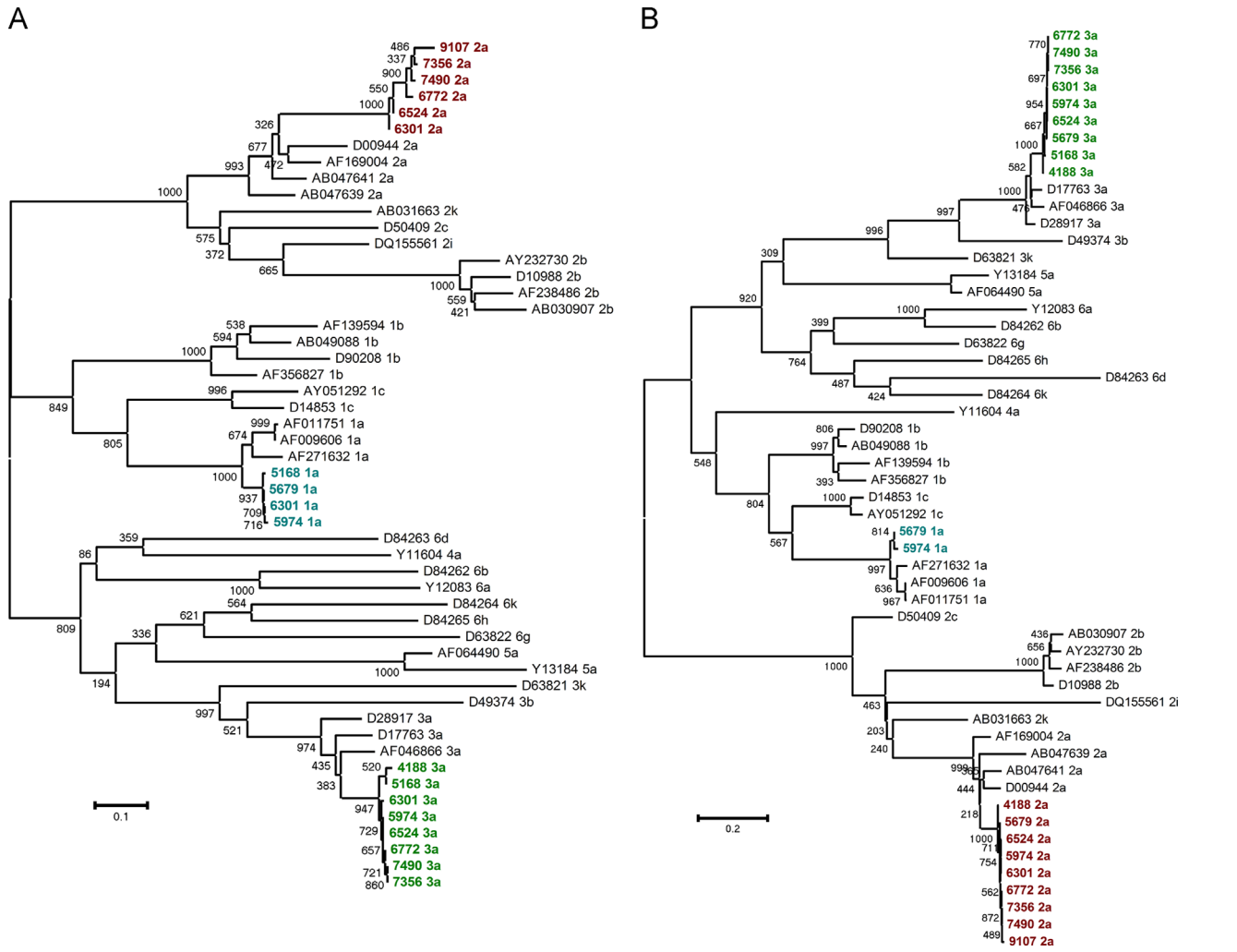

Fig. 1. Maximum Likelihood Tree for the E2 (A) and NS5A (B) regions obtained with PhyML 3.0 using GTR+Γ+I as a model of nucleotide substitution as recommended by ModelTest 3.07 (Akaike Information Criterion) for both datasets. The sequences with name in bold were obtained from samples of the patient studied and were named after the sample code and the genotype for which the primers were designed. The rest of the sequences were named after their GenBank accession number and genotype. The numbers above or below the branches represent the bootstrap support value (% over 1000 pseudoreplica).

with Bellerophon (Huber et al., 2004) and also phylogenetically and visually inspected to exclude recombinant sequences (chimeras) between already sequenced "parental" clones. Four clones (5, 10, 21 and 22) of HCV-1a from sample 5168 were removed because they showed evidence of recombination.

The phylogenetic analyses of the E2 sequences of HCV-2a and 3a cloned showed one highly supported monophyletic clade each, whereas the HCV-1a cloned sequences showed two highly supported clades (bootstrap value of 100%), each related to a different set of 1a sequences from GenBank (Supplementary Fig. 2G).

The diversity analysis using the mean Hamming distance between the cloned sequences of the same time for each genotype ranged between 0.54% and 1.83% ($<2\%$ expected as within host diversity). Strikingly, the HCV-1a dataset showed a mean Hamming distance of 7.15% for sample 5168 ($>2\%$ but $<15\%$ as expected for different lineages of the same subtype), which supports the occurrence of two different lineages at the beginning of the follow-up.

The mean Hamming distance provides rough information about the true diversity of the viral population (Gray et al., 2012). To overcome this issue, Bayesian Skyline Plot Analyses were used to assess the changes in the sequence diversity through the time (population dynamics). These plots showed that each genotype exhibited particular profiles of population dynamics that were not described by a simple (parametric) model. The Bayes Factor analysis of the different demographic models also confirmed that the best fit population model was, for the three genotypes, the Bayesian Skyline Plot.

Supplementary Fig. 3 represents the evolutionary follow-up of HCV-1a (I), 2a (II) and 3a (III), presented by the time-annotated tree of the clones (a), the Bayesian Skyline Plot (b) and the mean Hamming distance between the nucleotide sequences of the clones (c) plotted over the same time scale (as suggested by Gray et al. (2012)).

Genotype 1a: sequence stasis

The time-annotated tree for HCV-1a clones from variant 2 showed four clones in the first sample (5168) (Fig. 3). In the next time point (5679, after 511 days of evolution), the clones were classified into two sub-lineages. Interestingly, the sequences of the clones of the next time point (5974 after further 295 days of evolution) were all related to one of the sub-lineages of the previous sample, but in the last sample (6301 after another 326 days of evolution) the clones were related again to both sub-lineages in 5679. At amino acid sequence level, the HCV-1a clones

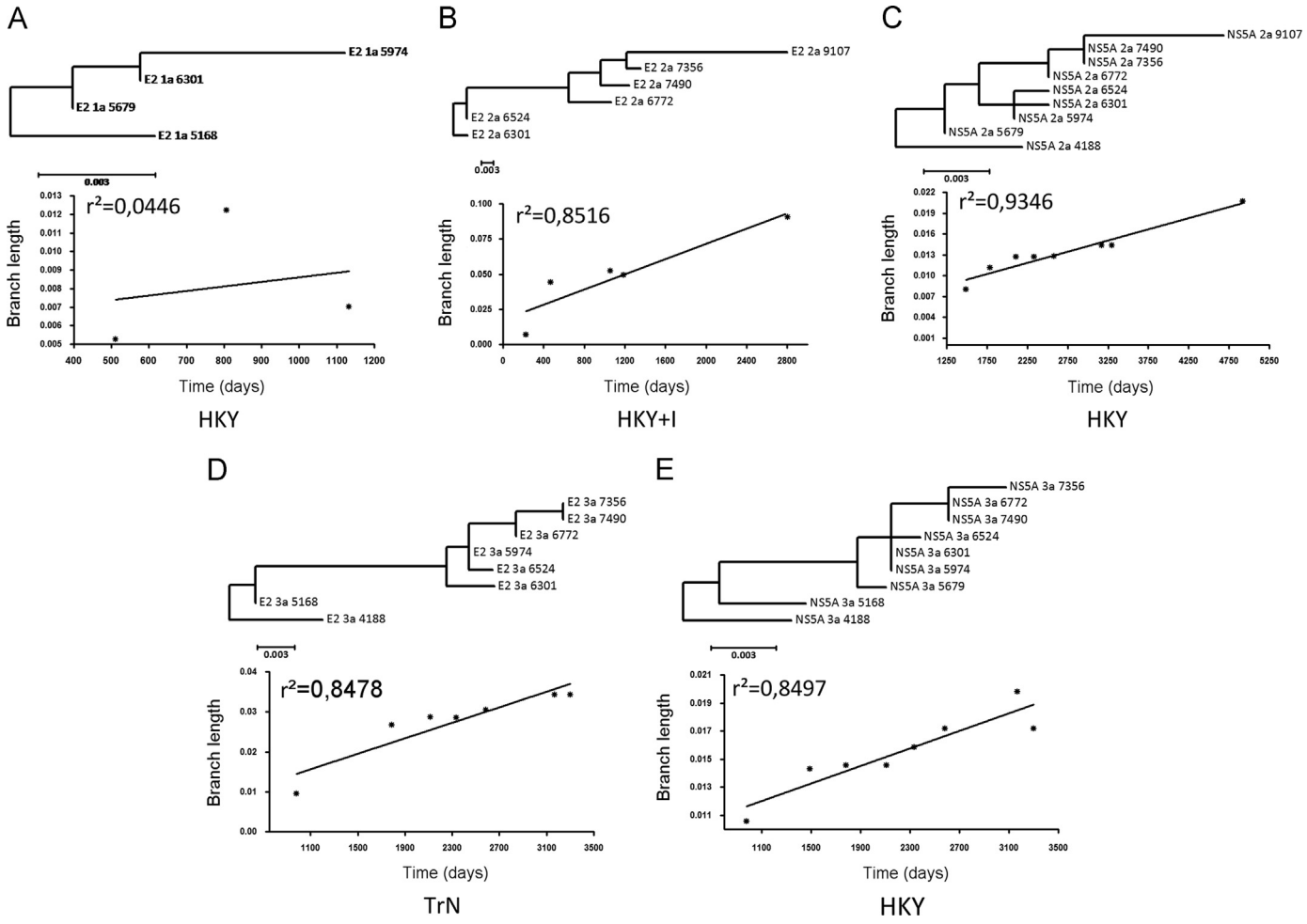

Fig. 2. Molecular Clock Analyses of Direct Sequences. Each frame contains the information of each dataset of direct sequences. The following data are indicated from the top to the bottom: Frame code, HCV genotype and HCV region; Maximum Likelihood Phylogenetic Tree obtained by exhaustive topology search; Branch length vs. Time regression; Model of nucleotide substitution used for the phylogenetic analysis. HKY: Haseagua, Kishino and Yano (1985) model; +I: model with a proportion of invariable sites; TrN: Tamura and Nei (1993) model.

Table 2
Summary of the results for the substitution rate estimation. Genotype: Dataset sequence genotype. Region: Dataset sequence genomic region. Dataset type: type of sequences. Direct: sequences from PCR product. Cloned: sequences from clones obtained from the PCR product. Method: method used. Regression: the rates were estimated from the slope of the linear regression of the branch length against the time; in Remarks, it is stated if the null hypothesis of a slope equal to zero was rejected. BSP: the rates were jointly estimated with the genealogy and the demographic history of the sequences in a Bayesian framework by the use of the Bayesian Skyline Plot; the model of molecular clock that fits with the data is stated in Remarks: Strict Clock or Uncorrelated Log-Normal (ULN) Relaxed Clock. Rate: nucleotide substitution rate. Lower and Upper C.I.: Value of the lower and upper limits of the confidence intervals for regression or the values of the upper and lower high posterior probability densities (95%) for the Bayesian Skyline Plot analysis. *: the substitution rate of the NS5A region for HCV-1a was estimated as the Hamming distance (p-dist) between the two direct sequences divided by the time span between them (time). n/a: not available.

Genotype	Region	Dataset type	Method	Rate (s/s/y)	Lower C.I. (s/s/y)	Upper C.I. (s/s/y)	Remarks
1a	E2	Direct	Regression	8.99×10^{-4}	-5.20×10^{-2}	5.38×10^{-2}	Slope = 0
1a	E2	Cloned	Regression	2.37×10^{-3}	1.29×10^{-3}	3.45×10^{-3}	Slope ≠ 0
1a	E2	Cloned	BSP	3.59×10^{-3}	2.34×10^{-3}	4.91×10^{-3}	ULN Relaxed Clock
1a	NS5A	Direct	p-dist/time	7.30×10^{-4}	n/a	n/a	Only 2 sequences*
2a	E2	Direct	Regression	9.90×10^{-3}	2.31×10^{-3}	1.76×10^{-2}	Slope ≠ 0
2a	E2	Cloned	Regression	1.24×10^{-2}	1.16×10^{-2}	1.31×10^{-2}	Slope ≠ 0
2a	E2	Cloned	BSP	1.40×10^{-2}	1.03×10^{-2}	1.80×10^{-2}	Strict Clock
2a	NS5A	Direct	Regression	1.17×10^{-3}	8.60×10^{-4}	1.48×10^{-3}	Slope ≠ 0
3a	E2	Direct	Regression	3.53×10^{-3}	1.81×10^{-3}	5.26×10^{-2}	Slope ≠ 0
3a	E2	Cloned	Regression	4.69×10^{-3}	4.38×10^{-3}	5.00×10^{-3}	Slope ≠ 0
3a	E2	Cloned	BSP	7.26×10^{-3}	5.45×10^{-3}	9.13×10^{-3}	ULN Relaxed Clock
3a	NS5A	Direct	Regression	1.14×10^{-3}	6.60×10^{-4}	1.62×10^{-3}	Slope ≠ 0

of all time points showed a high degree of similarity and only a few mutations were detected sporadically. Noteworthy, no amino acid mutations were fixed at any position of the E2 region including hyper-variable region 1 (HVR-1) along 3 years and 38 days of evolution. This apparently slow evolutionary process resulted in the estimation of a nucleotide substitution rate of

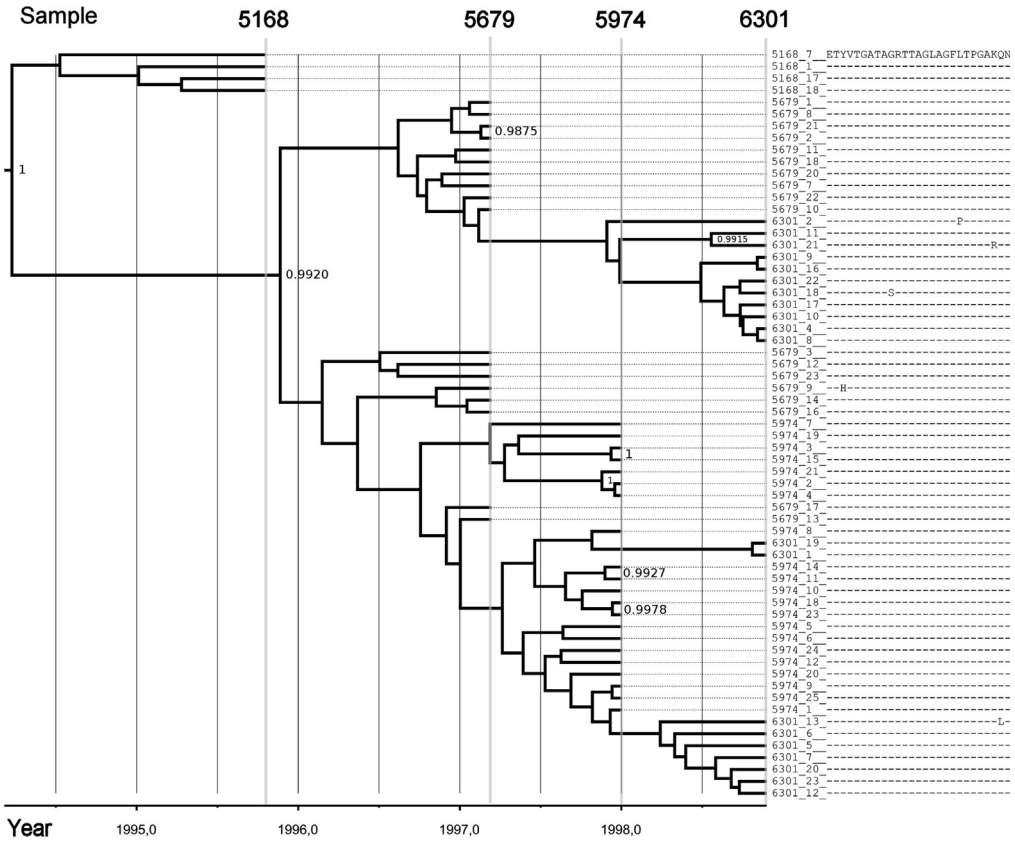

Fig. 3. Time-annotated tree and amino acid alignment of the HVR-1 for HCV-1a E2 Clones. Maximum clade credibility tree of the E2 sequences of HCV-1a cloned. The number on the right side of each clade is the posterior probability (the rest of the clades had a posterior probability lower than 0.95). Samples: time-point code of the cloned samples. The sequences are named after their sample time-code and clone number. The *X* axis represents the time in calendar years. The alignment of HVR-1 is shown on the right of the sequence names. In the alignment, each letter corresponds to one amino acid following the one-letter IUPAC code. The "-" character represents that this position is occupied by the amino acid of the same position in the top sequence.

3.59×10^{-3} s/s/y (between 2.34 and 4.91×10^{-3} s/s/y) for the E2 region of HCV-1a (Table 2).

The Bayesian Skyline Plot obtained for HCV-1a showed a period of constant population diversity until a time situated between samples 5168 and 5679. Diversity then increased about 10 times up to sample 5679, decreased until 5974, and finally remained constant until the last HCV-1a sample. The mean Hamming distance (nucleotides) of the clones ranged between 0.57% and 0.93% and showed no correlation with the skyline profile (Supplementary Fig. 3A).

Genotype 2a: lineage exchange

The time-annotated tree for HCV-2a clones showed the existence of two groups of sequences in each time-point except for the last sample (9107) (Fig. 4). The complexity of the evolutionary process carried out by HCV-2a in this patient could be roughly described as follows: The sequences cloned from the first sample (6301) formed two defined groups: α and β, which remained with few changes 223 days later in the next sample (6524) as groups γ and δ respectively. Interestingly, all the sequences cloned from the further samples were related only to group δ, indicating the "extinction" (or at least the lack of detection) of the lineage leading to group γ. A similar behavior was observed throughout all the evolutionary process, suggesting a sub-population shift characterized by the expansion of divergent sequences that formed new lineages and led to the extinction of others. This process is clearly observed at amino acid level, where, at each time, two lineages with many amino acid differences at HVR-1 co-existed in the same sample. As a result of this quick turnover of viral variants, the estimation of the substitution rate behind this process was relatively high: 1.40×10^{-2} s/s/y (between 1.03 and 1.80×10^{-2} s/s/y) (Table 2).

The Bayesian Skyline Plot obtained for HCV-2a showed an overall constant profile with peaks in the diversity by the time of samples 6301, 7356 and 9107. The mean Hamming distance of the HCV-2a clones ranged between 0.56% and 1.65% and showed some degree of correlation with the skyline profile, since there were peaks in samples 6524 and 7356. However, for sample 9107, which showed a peak in the skyline, the Hamming distance was the lowest (Supplementary Fig. 3B).

Genotype 3a: mixed pattern

The time-annotated tree for HCV-3a clones showed a complex pattern of evolution (Fig. 5). In the first time-point (sample 4188), the clones formed a cluster that was related to the common ancestor of the two clusters of the next time-point (sample 5168, after 980 days). Only one of those clusters was related to the ancestor of all clones of the following samples. Beyond the third time-point (sample 6301, another 1133 days), the diversity of the clones increased and then the clustering of the ancestral sequences was difficult to assign, resulting in an intermingling of sequences of one time-point with the ancestor of the further ones. At the amino acid sequence level, all clones of the first time-point showed the same HVR-1 sequence. Similarly, all the clones for the second time-point showed the same HVR-1, which had four amino acid differences with that of the previous sample. From the third time-point onwards, there were different lineages with a common HVR-1 amino acid sequence but each lineage differing from each other by one or two amino acid positions. In this case, the

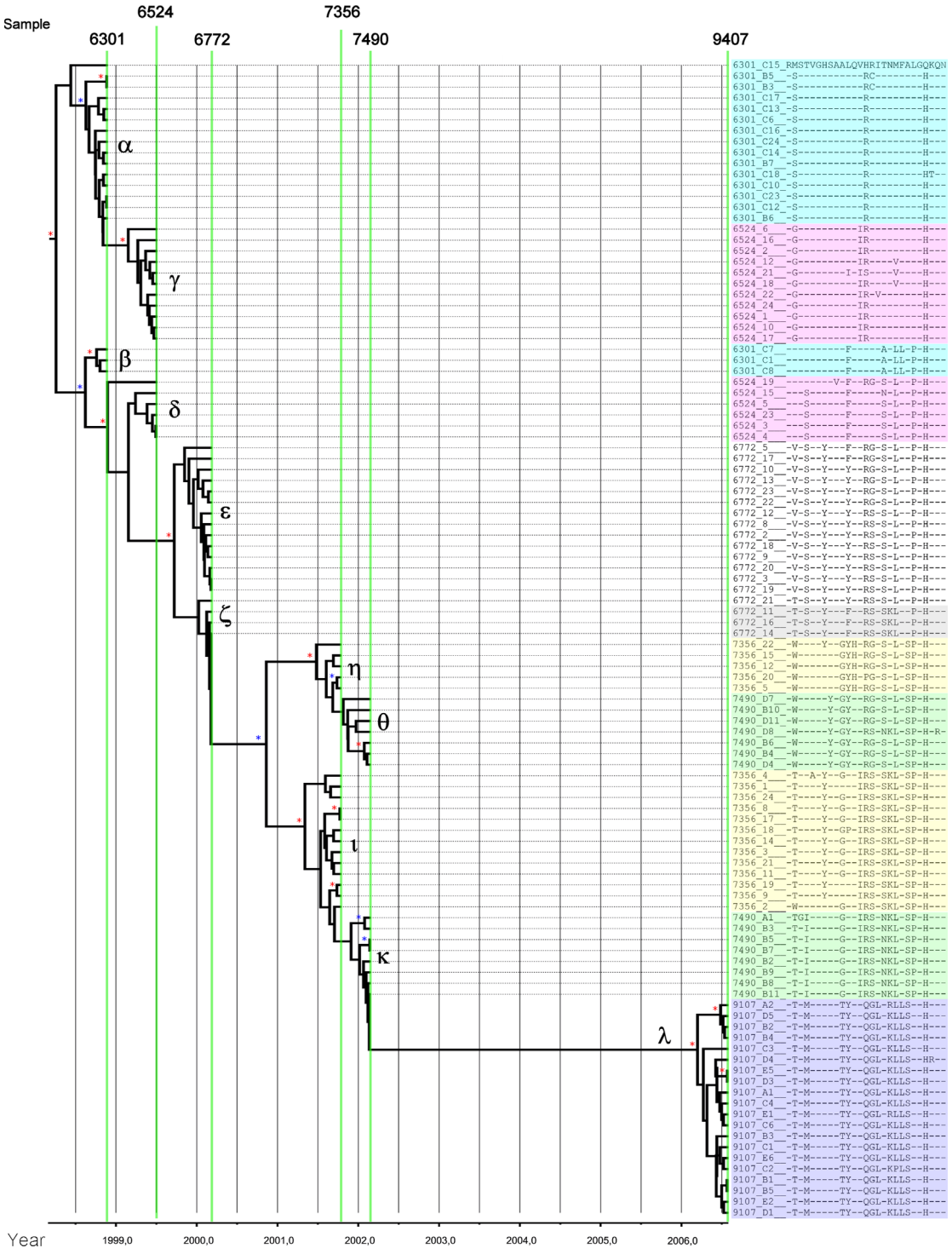

Fig. 4. Time-annotated tree and amino acid alignment of HVR-1 for HCV-2a E2 Clones. Maximum clade credibility tree of the E2 sequences of HCV-2a cloned. The posterior probabilities (PP) of the supported clades are depicted with stars above the supported node: red stars for PP ≥ 0.99; blue stars for 0.95 ≥ PP < 0.99. Samples: time-point code of the cloned samples. The sequences are named after their sample time-code and clone number. The *X* axis represents the time in calendar years. The alignment of HVR-1 is shown on the right of the sequence names. In the alignment, each letter corresponds to one amino acid following the one-letter IUPAC code. The "-" character represents that this position is occupied by the amino acid of the same position in the top sequence. The Greek letters are the names of the groups of sequences. The sequences from the same time are shaded with the same color.

estimation of the substitution rate for the whole process resulted in an intermediate substitution rate of 7.26×10^{-3} s/s/y (between 5.45 and 9.13×10^{-3} s/s/y) (Table 2).

The Bayesian Skyline Plot obtained for HCV-3a also showed two periods with different level of diversity: one with low diversity, which spans from the time of the first sample (4188) until the third one (6301), with a peak in the time of sample 4188; and another one with high diversity, which spans from sample 6301 to the last HCV-3a sample (7490), with a groove in the time of sample 6524.

The mean Hamming distance ranged from 0.64% to 1.83%, and although the two periods cannot be distinguished by this index, the mean distance increased with the time (Supplementary Fig. 3C).

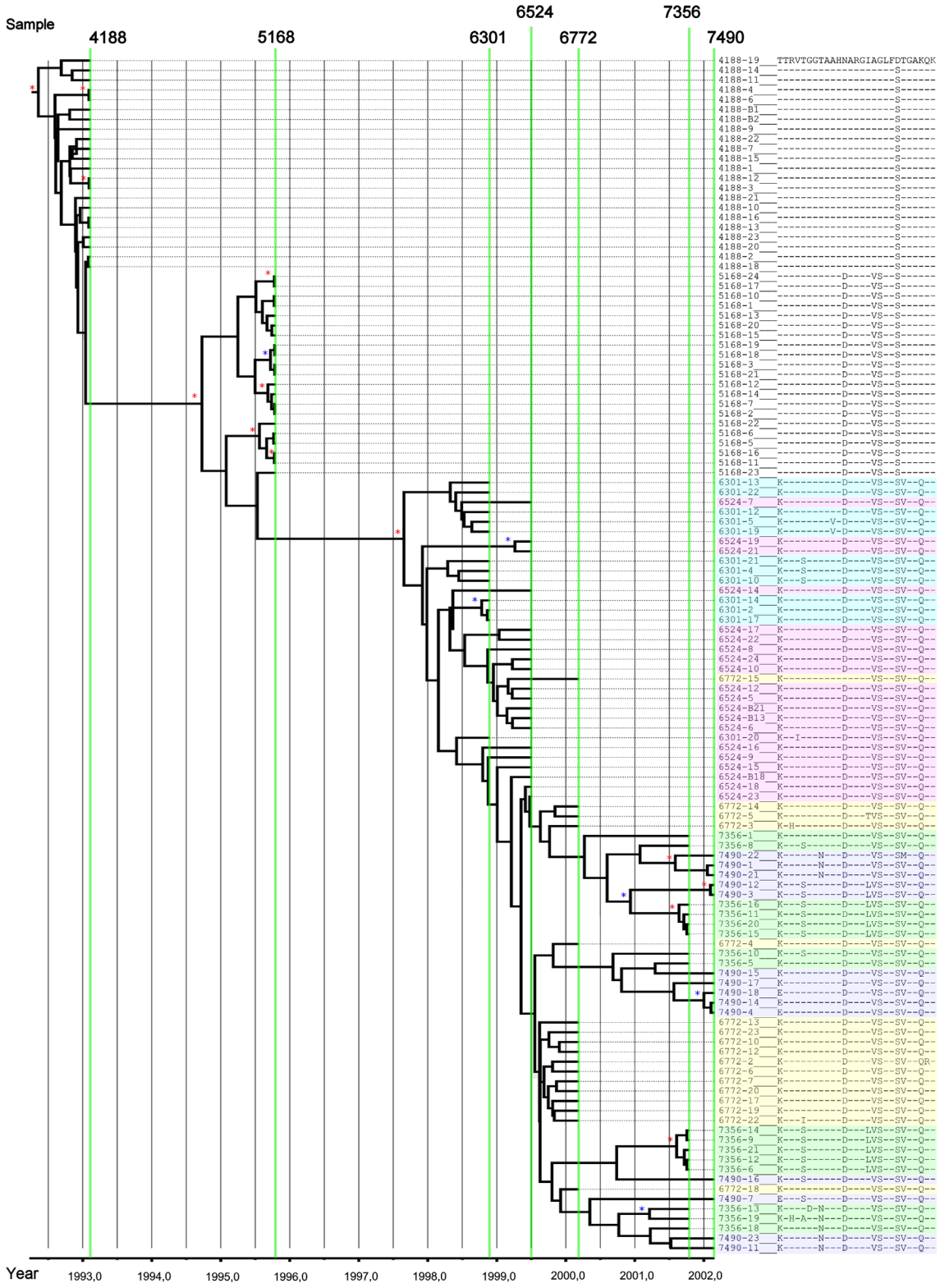

Fig. 5. Time-annotated tree and amino acid alignment of HVR-1 for HCV-3a E2 Clones. Maximum clade credibility tree of the E2 sequences of HCV-3a cloned. Samples: time-point code of the cloned samples. The sequences are named after their sample time-code and clone number. The *X* axis represents the time in calendar years. The alignment of HVR-1 is shown on the right of the sequence names. In the alignment, each letter corresponds to one amino acid following the one-letter IUPAC code. The "-" character represents that this position is occupied by the amino acid of the same position in the top sequence. The sequences from the same time are shaded with the same color.

Discussion

The patient analyzed in this work exhibited a co-infection with three HCV genotypes (1a, 2a and 3a), which is a known issue among patients with hemophilia treated with human plasma-derived clotting factors (Lee, 2009). In this case, the analyses of multiple samples and genotype-specific primers were required to unequivocally identify the co-infecting HCV genotypes. This

observation highlights that, to address the actual genotypes involved in the infection, it is important to carry out a careful genotype evaluation, especially in the presence of risk factors for a multiple infection, such as clotting factor treatment.

The evolution of each genotype was followed using direct sequences as well as cloned sequences by phylogenetic and coalescence methods. The study of direct sequences showed differences in the evolutionary patterns exhibited by the E2 region in each genotype, but no differences in the NS5A region. In addition, the analyses of cloned sequences confirmed the differences in the evolutionary dynamics of the E2 region.

Direct sequences

In the phylogenetic analyses, the tree topology for HCV-2a and 3a strongly suggests a relationship between ancestry and sampling time. Moreover, as a result of exhaustive tree searches, the E2 and NS5A direct sequence datasets for HCV-2a and 3a yielded comb-like trees where ancestry was compatible with the sampling time. That was not the case for the E2 sequences of HCV-1a, in which "new" samples were located as ancestors of "older" ones.

These observations were further supported by the correlation between the calendar distance and the branch lengths estimated in the best trees for HCV-2a and 3a datasets but not for HCV-1a. This scheme is compatible with the action of either a strict or relaxed molecular clock, but further analysis was limited by the low number of taxa in the direct sequence dataset.

In addition, the rough estimations of the substitution rates estimated for the E2 regions of HCV-2a and -3a were different while those estimated for the NS5A regions were undistinguishable (Table 2), suggesting the action of different mechanisms of evolution in these genomic regions. Although the regression analysis is useful to explore the temporal structure of the datasets, the assumptions made for the regression are not fulfilled by a phylogeny, making the estimation of the substitution rate by this means unreliable (Drummond et al., 2003). These rates (slopes) were further confirmed by a Bayesian coalescence analysis which may have limitations due to the low number of sequences in the direct datasets. Despite the different limitation of these analyses, the results were compatible with the known characteristics of the E2 and NS5A regions. The HCV E2 region is characterized as the most variable region as a result of its interaction with the immune system (von Hahn et al., 2007; Smith et al., 2010; Guan et al., 2012). In contrast, the NS5A region is a conserved region thought to be mainly under negative selection due to its regulatory functions (He et al., 2006). Each infection (in this case represented by different genotypes), evolving in the same host and at the same time, showed dissimilar substitution rates at the HCV E2 region, which suggests a different interplay with host's factors. Since the host's genetic background is constant, the source of this diversity in the interactions could be the adaptive immune system. In the E2 region of HCV-1a, the amino acid sequence remained constant throughout the period analyzed, evidencing some degree of immune tolerance, whereas for HCV-3a, and more clearly for HCV-2a, the region appears as extremely variable with a higher substitution rate and broad modifications at the amino acid sequences.

The results of the direct sequences are in accordance with previous differences observed for the E2 and NS5A regions and with the fact that each genotype showed differences in the E2 region as a result of different interplays of each genotype with the host immune system.

Cloned sequences

For a deeper study of the evolutionary process, a population genetic analysis was performed using cloned E2 sequences.

The diversification process has been associated with several features of clinical impact (Farci et al., 2006; Bull et al., 2011; Saludes et al., 2013). The use of coalescence-based methods allows analyzing the dynamics of viral lineages in the host over the time (Gray et al., 2012). The analysis of the time-annotated trees and the Bayesian Skyline Plots allowed the recognition of complex interactions between the lineages that compose the whole viral population that circulate in the patient. The population genetic analysis showed that, for the three genotypes, new lineages are generated continuously, but in some case these lineages can either increase and become dominant or decrease and finally become extinct. In the case of HCV-1a, there was a dominant lineage that prevailed for nearly 3 years, and even when "new" different sequences were cloned at each time-point, none of these mutants were fixed in the population (negative selection). Contrastingly, for HCV-2a, almost every time-point was characterized (at population level) by co-dominance of two lineages. Moreover, each lineage showed its own evolutionary process and some of these even became extinct. HCV-3a showed a more complex behavior with dominance of one lineage that allowed the fixation of some variants in the first 3 years, but then switched to the co-dominance of several variants that prevailed (in different proportion) for the rest of the follow-up. It is worth noting that these evolutionary patterns have been previously described in the follow-up of chronic HCV infections in different hosts (Ramachandran et al., 2011). Interestingly, in this work, HCV evolution occurred in the same host, which limits the impact of the particular host background driving the virus evolution. This highlights the complexity of the host-virus interaction, since not only HCV-1a and HCV-2a showed different manifestations of the evolutionary process in the same host at the same time, but also HCV-3a showed a change between two patterns in the period analyzed.

It is possible that the three HCV genotypes in this patient represent different starting points in an immune escape process leading to different outcomes. While HCV-1a probably reaches an equilibrium point where immune evasion was not necessary (most of the cloned sequences present identical HVR-1 amino acid sequences), both HCV-2a and HCV-3a were in a continuous process of selection of new evasion mutants (different sequences were observed for HVR-1 in the clones).

The recognition of splitting and extinction of lineages can only be achieved at intra-host viral population level with the aid of cloned sequences. Novel technologies like pyrosequencing may also allow modeling the evolutionary dynamics at this level (Wang et al., 2010) by providing a number of readings (virtual clones) that may detect even very scarce viral variants, but this information may not be tractable in the Bayesian-coalescent framework.

Unfortunately, the low HIV viral load and the steady levels of CD4+ cells observed in the patient throughout the period analyzed do not provide enough information to make further speculations about the impact of HIV co-infection (and therapy) over the evolution of HCV. Likewise, the temporal distance of the direct (and cloned) HCV sequences to the INF therapy makes it difficult to link any feature of the HCV evolution to this event.

Finally, these results support a mechanism of evolution for HCV where new lineages are generated continuously and then, in some cases, these lineages can either increase and become dominant or decrease and finally become extinct. Moreover, the magnitude of this process may be different for different genotypes and variable over the time for a same genotype which suggests that HCV may explore a number of different ways to persist in the host.

Each genotype evolving in the same patient and at the same time presents a different evolutionary pattern apparently modulated by the immune pressure of the host.

This study provides useful information for the management of co-infected patients and provides insights into the mechanisms behind the intra-host evolution of HCV.

Materials and methods

Ethical statement

Written informed consent to participate in this study was obtained from the patient. The study protocol was approved by the ethics committees of the "Academia Nacional de Medicina" and "Facultad de Farmacia y Bioquímica de la Universidad de Buenos Aires" (record number 732575/2010) in accordance with the 1975 Helsinki Declaration.

Samples

A retrospective study was carried out using plasma samples from a patient with severe hemophilia A, currently older than 60 years. In his youth, the patient received first-generation non-inactivated clotting factor VIII concentrates which led to the infection with HIV and HCV. Clinical records (Supplementary Fig. 1) include CD4 cell counts, HIV and HCV viral loads from 1993 to 2008. Since 1990, he has received uninterrupted antiretroviral therapy first with zidovudine only and then with different schemes of highly active antiretroviral therapies. In 2003/2004, he also received anti-HCV treatment (PEG-Interferon plus Ribavirin) and was considered a partial responder. The samples analyzed in this study were coded based on their date, counted as days after an arbitrary date in the past. The numbers coding the dates were selected so that the resulting dates (including the possible ancestor's dates of the different analyses carried out) were higher than zero in order to avoid possible unhandled exceptions in the programs due to a "negative time".

RT-PCR-sequencing

The RNA was extracted from plasma samples using the QIAmp Viral RNA Mini Kit (QIAGEN) and then transcribed with MMLV retrotranscriptase (Promega) with random hexamer primers following the manufacturer's recommendations. Then the cDNA was used as template for the nested polymerase chain reaction (PCR) protocols with the Go Taq reagents (Promega) and with specific primers designed to amplify the 5′UTR (Davidson et al., 1995) and NS5B (Chen and Weck, 2002) regions used for genotyping. 5′UTR genotyping was carried out by Restriction Fragment Length Polymorphism (RFLP) analysis, as already described (Davidson et al., 1995). The samples were subtypified by phylogenetic analysis of NS5B sequences.

After genotyping, new PCRs were carried out for each sample using genotype specific primers for the NS5A and E2 regions (Supplementary Table 1). These regions represent both structural and non-structural genomic regions which are thought to be subject to different kinds of selective pressures.

For the NS5A and E2 regions, a simple PCR protocol was carried out: 5 μl of the cDNA was used as template for the first round PCR in a final volume of 50 μl with 1.5 mM of Mg^{++}, 200 μM of dNTP (each), 0.4 μM of direct and reverse primers (see Supplementary Table 1) and 1.25 units of *Taq* polymerase. The first round comprised initial denaturation at 95 °C for 5 min followed by 30 cycles of denaturation at 95 °C for 1 min, annealing at 55 °C for 1 min and elongation at 72 °C for 1 min. Final elongation at 72 °C was carried out for 10 min. The same conditions were used for a second round with 2 μl of first round product used as template. The thermocycler program for the second round was similar to the first one, with the exception of the number of cycles (40 instead of 30) and the annealing temperature (57 °C instead of 55 °C).

All PCR products were purified using the General Electric Health Care Illustra GFX PCR and Gel Band DNA extraction kit and sequenced in an automatic sequencer.

Molecular cloning

The E2 amplicons were cloned into the p-GEM-T Easy Cloning Vector (Promega) following the manufacturer's protocol, which provides a blue/white system for clone selection. Briefly, the E2 amplicons were purified to remove PCR byproducts and primers, and used as inserts in a ligation reaction with the linearized vector. After the ligation, the vector was transfected into *Escherichia coli* by heat shock, and plated in selective media (LB agar with Ampicilin, IPTG and X-Gal). White clones were picked and grown in LB broth overnight. The plasmidic DNA was purified by in-house alkaline lysis coupled with Phenol-Chloroform extraction. Then, the DNA obtained was sequenced using the universal T7 promoter primer.

Sequence relationship assessment

The phylogenetic relatedness of the E2 sequences cloned was assessed by phylogenetic reconstruction of the patient's viral sequences plus the set (non-redundant) of the 10 best hits (lowest "*E* value") retrieved with BLAST for each. To do this automatically, a BASH script called Blast-Fishing was written (available upon request). This script performs the BLASTN query retrieving a desired number of hits (10 in this work). Then, it collects all BLASTN hit accession numbers, removes the duplicated ones and downloads the full sequences from GenBank.

The sequences downloaded were then aligned with those obtained using ClustalX v 2.12 (Larkin et al., 2007), and trimmed so that they matched the region analyzed. After selecting the best fit substitution model with ModelTest 3.07 (Posada and Crandall, 1998), the phylogeny was reconstructed by maximum likelihood using bootstrapping as branch support (1000 pseudoreplica) with the program PhyML 3.0 (Guindon and Gascuel, 2003) for Linux.

The sequences were considered to be related if they formed a monophyletic group with a bootstrap value higher than 95.

Phylogenetic analyses

The sequences obtained were visualized and manipulated with BioEdit v5.7 (Hall, 1999) software and aligned using ClustalX. The phylogenetic reconstructions were carried out with a maximum likelihood criterion. The best nucleotide substitution model was selected for each dataset analyzed according to the Akaike information criterion implemented in the program ModelTest. In the case of small datasets (genotype-specific NS5A and E2 direct sequences), the tree search strategy was an exhaustive topology search performed with the program PAUP* v4.10b (Swofford, 2003)). In the datasets with more than 10 sequences, the tree search strategy was a heuristic search using the hill climbing algorithm implemented in the program PhyML v3.0. Regardless of the tree search strategy, all parameters were adjusted as suggested by ModelTest. In the case where branch support values were required, the non-parametric bootstrap analysis was carried out with PhyML (1000 pseudoreplica).

Molecular clock analyses

Regression of branch lengths versus time

As an initial and intuitive method to measure whether or not the sequences were correlated with the time of sampling, a

root-to-tip regression analysis was carried out. Although this analysis allows the numerical (and chart) representation of the genetic change as a function of time and is thus useful to explore the temporal structure of the datasets, the assumptions made for the regression are not fulfilled by a phylogeny, making the estimation of the substitution rate by this means unreliable (Drummond et al., 2003). This method is based on the fact that maximum likelihood tree reconstruction optimizes the length of the branches so that they represent the amount of genetic change ($\mu \times T \times s$; where μ represents the substitution rate, T the time and s an arbitrary scaling factor) that relates sequences, and thus the linear regression of the "amount of genetic change" (Branch Length) with the time span between the pairs formed with the older and each sequence would result in a positive non-zero slope if the tree topology is comb-like and the sequence has a temporal structure.

For each dataset, the best tree was transformed into a distance matrix using the "Pairwise distance matrix" option of the program HyPhy v2.0 (Kosakovsky Pond et al., 2005) for Windows. Then, the linear regression was carried out with Prism GraphPad v4.10 (GraphPad Software, San Diego, California, USA, www.graphpad.com) using the branch length and time distances between each sample and the oldest one. Thus, all sequences except the oldest one were considered only one time for the regression.

The software also computes the probability that randomly selected points result in the same slope as the estimated one (full description of the test available at http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?istheslopesignificantlydifferentthanzero_.htm).

Bayesian coalescent analysis

The datasets with the cloned sequences were analyzed using the Bayesian Coalescent framework implemented by BEAST v1.6.1 (Drummond and Rambaut, 2007). After selecting the best fit substitution model, the datasets were run in BEAST using the available population (constant, exponential growth, logistic growth, expansional growth and Bayesian skyline) and clock models (strict, and relaxed uncorrelated lognormal) for 5×10^7 generations. The results were analyzed using Tracer v1.5 (available from: http://tree.bio.ed.ac.uk/software/tracer/). The runs were inspected for convergence, i.e. whether or not the estimated parameter (*clock.rate* or *meanRate*, *treemodel.rootHeight*), likelihood, prior and posterior yielded an Effective Sample Size (ESS) greater than 200, and visually to check if there were any trends in the trace of each parameter. The convergent runs were further compared with Bayes Factor (Kass and Raftery, 1995) to choose the population and clock models that best fitted the data. The substitution rate estimations were taken from these runs and since the time data were coded in days, the estimators obtained (in s/s/d) were converted into s/s/y by multiplying by 365.25. The trees of the selected runs were summarized (annotated) using the TreeAnnotator utility (BEAST v1.6.1 program) to obtain a representation of the coalescent events. The annotated trees were drawn with the Figtree v1.3.1 program (available from http://tree.bio.ed.ac.uk/software/Figtree).

Acknowledgments

This study was carried out thanks to the financial support of the grants: PIP 112-200801-01169 (Consejo Nacional de Investigaciones Científicas y Técnicas), PICT 2006-00285 (Agencia Nacional de Promoción Científica y Tecnológica) and UBACyT 2008-B601 (Universidad de Buenos Aires). The founders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2013.11.034.

References

Bull, R.A., Luciani, F., McElroy, K., et al., 2011. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog. 7 (9), e1002243, http://dx.doi.org/10.1371/journal.ppat.1002243.

Chen, Z., Weck, K.E., 2002. Hepatitis C virus genotyping: interrogation of the 5' untranslated region cannot accurately distinguish genotypes 1a and 1b. J. Clin. Microbiol. 40 (9), 3127–3134.

Davidson, F., Simmonds, P., Ferguson, J.C., Jarvis, L.M., Dow, B.C., Follett, E.A.C., Seed, C.R.G., Krusius, T., Lint, C., Medgyesi, G.A., Kiyokawa, H., Olim, G., et al., 1995. Survey of major genotypes and subtypes of hepatitis C virus using RFLP of sequences amplified from the 5' non-coding region. J. Gen. Virol. 76, 1197–1204.

Drummond, A., Pybus, O.G., Rambaut, A., 2003. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitol. 54, 331–358.

Drummond, A.J., Rambaut, A., 2007. BEAST Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214–222.

Farci, P., Quinti, I., Farci, S., et al., 2006. Evolution of hepatitis C viral quasispecies and hepatic injury in perinatally infected children followed prospectively. Proc. Natl. Acad. Sci. USA 103 (22), 8475–8480.

Gray, R.R., Salemi, M., Klenerman, P., Pybus, O.G., 2012. A new evolutionary model for hepatitis C virus chronic infection. PLoS Pathog. 8 (5), e1002656, http://dx.doi.org/10.1371/journal.ppat.1002656.

Guan, M., Wang, W., Liu, X., Tong, Y., Liu, Y., Ren, H., Zhu, S., Dubuisson, J., Baumert, T.F., Zhu, Y., Peng, H., Aurelian, L., Zhao, P., Qi, Z., 2012. Three different functional microdomains in the hepatitis C virus hypervariable region 1 (HVR1) mediate entry and immune evasion. J. Biol. Chem. 287 (42), 35631–35645, http://dx.doi.org/10.1074/jbc.M112.382341.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52 (5), 696–704.

Hall, T.A., 1999. BioEdit a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids Symp. Ser. 41, 95–98.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22 (2), 160–174.

He, Y., Staschke, K.A., Tan, S.L., 2006. HCV NS5A: a multifunctional regulator of cellular pathways and virus replication. In: Tan, S.L. (Ed.), Hepatitis C Viruses: Genomes and Molecular Biology. Horizon Bioscience, Norfolk (UK), pp. 267–292. (Chapter 9).

Huber, T., Faulkner, G., Hugenholtz, P., 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20 (14), 2317–2319.

Kass, R.E., Raftery, A.E., 1995. Bayes Factors. J. Am. Stat. Assoc. 90 (430), 773–795.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21 (5), 676–679.

Kühnert, D., Wu, C.H., Drummond, A.J., 2011. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. Infect. Genet. Evol. 11 (8), 1825–1841, http://dx.doi.org/10.1016/j.meegid.2011.08.005.

Kuntzen, T., Timm, J., Berical, A., Lewis-Ximenez, L.L., Jones, A., Nolan, B., Schulze zur Wiesch, J., Li, B., Schneidewind, A., Kim, A.Y., Chung, R.T., Lauer, G.M., Allen, T.M., 2007. Viral sequence evolution in acute hepatitis C virus infection. J. Virol. 81 (21), 11658–11668.

Lavanchi, D., 2009. The global burden of hepatitis C. Liver Int. 29, 74–81.

Larkin, M.A., Blackshields, G., Brown, N.P., et al., 2007. Clustal W and clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lauer, G.M., Walker, B.D., 2001. Hepatitis C virus infection. N. Engl. J. Med. 345, 41–52.

Lee, C.A., 2009. The best of times, the worst of times: a story of haemophilia. Clin. Med. 9 (5), 453–458.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14 (9), 817–818.

Qin, H., Shire, N.J., Keenan, E.D., Rouster, S.D., Eyster, M.E., Goedert, J.J., Koziel, M.J., Sherman, K.E., 2005. Multicenter Hemophilia Cohort Study Group, 2005. HCV quasispecies evolution: association with progression to end-stage liver disease in hemophiliacs infected with HCV or HCV/HIV. Blood 105 (2), 533–541.

Ramachandran, S., Campo, D.S., Dimitrova, Z.E., Xia, G.L., Purdy, M.A., Khudyakov, Y.E., 2011. Temporal variations in the hepatitis C virus intrahost population during chronic infection. J. Virol. 85 (13), 6369–6380, http://dx.doi.org/10.1128/JVI.02204-10.

Saludes, V., González-Candelas, F., Planas, R., Solà, R., Ausina, V., Martró, E., 2013. Evolutionary dynamics of the E1–E2 viral populations during combination therapy in non-responder patients chronically infected with hepatitis C virus subtype 1b. Infect. Genet. Evol. 13, 1–10, http://dx.doi.org/10.1016/j.meegid.2012.09.012.

Smith, J.A., Aberle, J.H., Fleming, V.M., Ferenci, P., Thomson, E.C., Karayiannis, P., McLean, A.R., Holzmann, H., Klenerman, P., 2010. Dynamic coinfection with multiple viral subtypes in acute hepatitis C. J. Infect. Dis. 202 (12), 1770–1779, http://dx.doi.org/10.1086/657317.

Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10 (3), 512–526.
von Hahn, T., Yoon, J.C., Alter, H., Rice, C.M., Rehermann, B., Balfe, P., McKeating, J.A., 2007. Hepatitis C virus continuously escapes from neutralizing antibody and T-cell responses during chronic infection in vivo. Gastroenterology 132 (2), 667–678.
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 (12), 6218, http://dx.doi.org/10.1128/JVI.02271-09.