Canadian Journal of Microbiology
Revue canadienne de de microbiologie

Genome-wide analysis for identification of adaptive diversification between hepatitis C virus subtypes 1a and 1b

Journal:	<i>Canadian Journal of Microbiology</i>
Manuscript ID	cjm-2016-0156.R1
Manuscript Type:	Article
Date Submitted by the Author:	14-Mar-2016
Complete List of Authors:	Li, Yan; Sichuan Agricultural University, College of Animal Science and Technology Wang, Ruirui; Yunnan University of Traditional Chinese Medicine Du, Xiaogang; Sichuan Agricultural University Zhang, Mingwang; Sichuan Agricultural University, College of Animal Science and Technology Xie, Meng; Sichuan Agricultural University
Keyword:	Hepatitis C virus, Subtypes 1a and 1b, Adaptive diversification, Genome-wide analysis

SCHOLARONE™
Manuscripts

1 **Genome-wide analysis for identification of adaptive diversification**
2 **between hepatitis C virus subtypes 1a and 1b**
3
4 Yan Li, Ruirui Wang, Xiaogang Du, Mingwang Zhang, and Meng Xie
5
6 **Y. Li and M. Zhang.** College of Animal Science and Technology, Sichuan Agricultural
7 University, Wenjiang, People’s Republic of China.
8 **R. Wang.** School of Pharmacy, Yunnan University of Traditional Chinese Medicine,
9 Kunming, Yunnan, People’s Republic of China.
10 **X. Du and M. Xie.** College of Life Science, Sichuan Agricultural University, Yaan,
11 People’s Republic of China.
12
13
14 **Corresponding authors:** Yan Li (e-mail: liyan@sicau.edu.cn) and Ruirui Wang (e-mail:
15 wangrr1980@163.com)
16

Abstract

Hepatitis C virus (HCV) is a major cause of liver disease and has been estimated to infect approximately 2–3% of the world's population. HCV genotype 1 is the subject of intense research and clinical investigations because of its worldwide prevalence and poor response to currently available therapies. The predominant subtypes 1a and 1b of HCV genotype 1 present considerable differences in epidemiological features. However, the genetic signature underlying such phenotypic functional divergence is still an open question. Here, we performed a genome-wide evolutionary study on HCV subtypes 1a and 1b. The results show that adaptive selection has driven the diversification between these subtypes. Furthermore, the major adaptive divergence-related changes have occurred on proteins E1, NS4B, NS5A, and NS5B. Structurally, a number of adaptively selected sites cluster in functional regions potentially relevant to (1) membrane attachment and (2) the interactions with viral and host cell factors and the genome template. These results might provide helpful hints about the molecular determinants of epidemiological divergence between HCV 1a and 1b.

38 **Keywords:**

39 Hepatitis C virus,

40 Subtypes 1a and 1b,

41 Adaptive diversification,

42 Genome-wide analysis.

43

Draft

Manuscript text

Introduction

Hepatitis C is a liver disease caused by the hepatitis C virus (HCV). The World Health Organization estimates that there are 185 million people infected with HCV worldwide (Mohd Hanafiah et al. 2013; Thomas 2013). The HCV can cause both acute and chronic hepatitis infection, resulting in 350, 000 to 500, 000 deaths each year from hepatitis C-related liver diseases (<http://www.who.int/mediacentre/factsheets/fs164/en/> 2014).

HCV possesses a single-stranded, positive-sense RNA genome, which was cloned and characterized in 1989 (Choo et al. 1989). The viral genome contains two untranslated regions at the 5' and 3' ends, and an open reading frame. The open reading frame encodes a polyprotein, which is further co- and post-translationally processed into ten mature polypeptides in the following order: core, envelope proteins (E1 and E2), p7, and, nonstructural (NS) proteins (NS2, NS3, NS4A, NS4B, NS5A, and NS5B) (reviewed in (Jackowiak et al. 2014; Moradpour et al. 2007)).

HCV is a member of the genus *Hepacivirus* within the virus family *Flaviviridae* (King et al. 2012). In addition to the *Hepacivirus*, *Flaviviridae* comprises the genera *Pestivirus* (e.g., classical swine fever virus (CSFV), bovine viral diarrhoea virus genotypes 1 and 2, and border disease virus) and *Flavivirus* (e.g., dengue and West Nile

65 viruses) (King et al. 2012). HCV is currently classified into 7 genotypes (1 to 7), which
66 are subdivided into 67 diverse subtypes (King et al. 2012; Kuiken and Simmonds 2009;
67 Smith et al. 2014). Besides phylogenetic separation, HCV genotypes and subtypes have
68 been associated with differences in geographical distribution (Esteban et al. 2008; Lu et
69 al. 2013; Sievert et al. 2011; Zein 2000) and responses to treatment under interferon
70 monotherapy or interferon/ribavirin combination therapy (Kamal and Nasser 2008;
71 Manns et al. 2001) and that under interferon-free regimens (Muir 2014; Webster et al.
72 2015).

73 Genotype 1 is distributed worldwide and contributes to a substantial percentage of
74 HCV infections (Mohd Hanafiah et al. 2013; Zein 2000). A number of sub-groupings
75 are apparent within genotype 1, of which subtypes 1a and 1b are most frequently
76 identified (Esteban et al. 2008; Lu et al. 2013; Sievert et al. 2011; Zein 2000). It has
77 been clearly shown that subtype 1a expanded explosively between the 1960s and around
78 1980, and subtype 1b grew exponentially from the late 1940s up to the 1980s
79 (Magiorkinis et al. 2009). It is noteworthy that subtypes 1a and 1b present considerable
80 differences in epidemiological features. Subtype 1a is prevalent in injection drug users,
81 which has been the major route of transmission in Australia, Europe and the United States
82 and now accounts for the prevalence of HCV infection in developing countries (Alter et
83 al. 1999; Dore et al. 2003a; Dore et al. 2003b; Esteban et al. 2008; Lavanchy 2011;
84 Nainan et al. 2006; Pawlotsky et al. 1995; Pybus et al. 2005; Sievert et al. 2011; Verachai
85 et al. 2002), whereas subtype 1b is acquired mainly from blood transfusions and unsafe

medical procedures (Hnatyszyn 2005; Pawlotsky et al. 1995; Simmonds 1997). Although HCV genotype 1 infection is now considered curable in >95% of treated patients with new drugs recently available, the high costs of the new treatment regimens make access to treatment a major problem (Muir 2014; Webster et al. 2015). Furthermore, there is still no effective vaccine for HCV because the obstacles to development of a vaccine for this virus are many, mainly from its evolutionary nature, including the genetic diversity and its effects on the host immune system (Abdel-Hakeem and Shoukry 2014; Abrignani et al. 1999; Hoofnagle 2002). Therefore, new infections and reinfections will remain a problem in developing countries and among marginalized populations like injection drug users (Abdel-Hakeem and Shoukry 2014; Drucker et al. 2001).

The epidemiological diversity between HCV 1a and 1b could be a reflection of viral adaptation in the host and provide important information not only for the basic virology of HCV, but also for the development of effective vaccines and future therapies. However, the molecular determinants underlying such distinctly epidemic phenotypes are poorly defined. An important step forward in detecting the decisive factors and understanding the diversification includes the description of the quality and quantity of genetic variations involved in the split of these subtypes. In this study, we analyzed the genetic divergence between HCV 1a and 1b in the complete polyprotein-coding region. Our results suggest that several proteins had experienced functional divergence in the process of dividing of the HCV genotype 1 progenitor into these subtypes. This study

provides helpful hints about the determinants of the epidemiological divergence between these subtypes and shed some light on the molecular mechanisms underlying such diversification.

Materials and methods

Sequence data collection and alignment

Complete genome sequences of HCV subtypes 1a and 1b with known dates of sampling were collected from the Los Alamos HCV sequence database (<http://www.hcv.lanl.gov>) (Kuiken et al. 2008). The genomes with nonsense mutations and ambiguous codons were manually excluded. The final analysis included 362 subtype 1a and 165 subtype 1b genomes. Table S1 and S2 show the GenBank accession numbers, dates of isolation, and origins of the included strains. For each subtype, the strains (1) up to year 1991 (31 subtype 1a and 10 1b strains, referred to as 1a and 1b 1991-datasets) and (2) up to year 2001 (90 subtype 1a and 39 1b strains, referred to as 1a and 1b 2001-datasets) recorded on the year of isolation were retrieved for the following evolutionary analyses. The genome sequences were aligned using ClustalW (Thompson et al. 1997) and the codon reading frames of alignment results were then checked manually.

Phylogenetic analysis of subtype 1a and 1b strains

To determine the phylogenetic distribution of the strains up to year 1991 within the global epidemic, we performed maximum likelihood (ML) analysis in MEGA5 using NS5B complete coding sequences of 362 subtype 1a and 165 subtype 1b strains (Tamura et al. 2011). The General Time Reversible model of evolution allowing gamma distribution plus a proportion of invariant sites was implemented in ML analysis on these datasets (Lanave et al. 1984). For each of the data sets, 1000 bootstrap replications were conducted to evaluate the reliabilities of the reconstructed trees.

McDonald-Kreitman (MK) test on the coding sequences of individual proteins

The MK test for protein coding data has been a commonly used method for detecting the presence of adaptive diversification (McDonald and Kreitman 1991). In this test, if the genetic variations are neutral, the ratio of replacement (nonsynonymous) to synonymous divergence between lineages should be roughly equal to the ratio of replacement to synonymous polymorphism within the lineages. By contrast, if some of the replacements have been influenced by positive adaptive selection, the ratio of replacement to synonymous fixed divergence should be greater than the ratio of replacement to synonymous polymorphism (Eyre-Walker 2006; McDonald and Kreitman 1991). Unfortunately, the estimate of adaptive diversification based on this method is biased downward by the segregation of low-frequency polymorphisms (Fay et al. 2001). It has been suggested that one way to cope with the impact of the bias is to remove the

low-frequency polymorphisms from the analysis (Charlesworth and Eyre-Walker 2008; Fay et al. 2001). However, it is technically difficult to determine the cutoff frequency below which we should exclude polymorphisms from the analysis (reviewed in (Charlesworth and Eyre-Walker 2008)). For HCV, because of the high mutation rate (Magiorkinis et al. 2009) and manifest existence of quasispecies rather than a defined genomic sequence in viruses (Farci et al. 2000; Noursbaum et al. 2000), it is not feasible to estimate the cutoff frequency for the MK test in the current study. Because the HCV subtypes 1a and 1b originated about 100 years ago and exploded during 1960-1980 and 1940-1980, respectively (Magiorkinis et al. 2009; Pybus et al. 2001), we can retrieve and analyze the early collected strains to detect the decisive factors relevant to their phenotypic functional divergence. This empirical strategy could reduce the downward bias from the accumulation of low-frequency polymorphisms in the MK test.

We first performed the MK test to identify adaptive evolution in all ten individual proteins separately with combined 1991-datasets of subtypes 1a and 1b (see above) (McDonald and Kreitman 1991). A Fisher's exact test of independence was performed (McDonald 2009). In order to validate potential bias in the use of a small amount of sequences, we repeated the MK test using the 2001-datasets.

Tests for the functional divergence on the coding sequences of all proteins

It is now widely accepted that the changes of amino acid physiochemical properties at critical sites could bring about the diversification as a response to different environmental

challenges (Li 1983; Nei 1987). We therefore tested for type II functional divergence at individual amino acid sites of each protein using the statistical method Gu-2006 in DIVERGE 3.0 (Gu 1999, 2001, 2006) with combined 1991-datasets of subtypes 1a and 1b. Type II functional divergence represents the evolutionary pattern where amino acids is conserved in both subtypes but their physiochemical properties are remarkably different. The method Gu-2006 calculates the coefficient of type II functional divergence (θ_{II}) between the generated subtypes (Gu 2006), and the Z-score test is performed for statistical evaluation based on the estimate of θ_{II} and its standard error (θ_{II} SE) (Gu 2001; Gu et al. 2013). The null hypothesis is $\theta_{II} = 0$, which means the absence of functional divergence between subtypes 1a and 1b. For the observed coefficient θ_{II} , a rejection of the null hypothesis indicates that a considerable shift of amino acid physiochemical properties is likely to have occurred between these subtypes. Furthermore, the posterior probability is used to predict the amino acid residues that could be responsible for the functional differences with a probability cutoff > 0.95 (Gu 2006).

Similar to the strategy for the MK test, we repeated the type II functional divergence analyses using the 2001-datasets to further validate the result obtained with the 1991-datasets.

Results

Signature of adaptive diversity between HCV subtypes 1a and 1b

We initially performed the MK test with the strains collected up to year 1991 (referred to as 1991-datasets) which closely follows the explosions of HCV subtypes 1a and 1b (Table S1 and S2). These datasets includes 31 subtype 1a and 10 1b strains. First, we carried out phylogenetic analysis separately for 362 subtype 1a and 165 subtype 1b strains with their NS5B coding sequences to estimate the potential representational bias of the 1991-datasets. For each subtype, the strains of the 1991-dataset were generally clustered into different branches and distributed over the span of the phylogenetic tree (Fig. S1). These results suggested that the impact of representational bias could be ruled out in the following analyses with the 1991-datasets. We then implemented the MK test to detect the occurrence of adaptive divergence with the split of the subtypes 1a and 1b from their common ancestor. Because the ten viral proteins potentially perform different functions, we assayed the evolution of individual proteins between these subtypes. Interestingly, for the coding sequences of E1, NS4B, NS5A, and NS5B, the Fisher's exact test of independence showed that the ratios of replacement versus fixed synonymous differences were significantly greater than those of replacement versus synonymous polymorphisms (Table 1). Additionally, the test on that of NS2 exhibited a deviation from the neutral mutation hypothesis with a marginal probability = 0.127. In contrast, the results from those of core, E2, p7, NS3 and NS4A did not show a similar pattern of evolution (Table 1).

Because a small amount of sequences counted tends to falsely estimate the genetic

variations (Li 1997), we then performed the MK test with the 2001-datasets, which include 90 subtype 1a and 39 1b strains. Consistent with those of 1991-datasets, for E1, NS4B, and NS5B, the Fisher's exact test of independence exhibited a deviation from neutrality in favor of adaptive divergence with a significant probability (Table S3). Meanwhile, the test on core, E2, p7, NS3 and NS4A suggested that no functional divergence took place (Table S3). These results indicated that recounting a few more sequences does not change the general pattern in the present case. On the whole, the accumulated results of the MK test under the empirical strategy implied that several proteins had experienced adaptive divergence with the split of HCV subtypes 1a and 1b.

Amino acid sites experienced type II functional divergence between HCV subtypes 1a and 1b

We implemented the method Gu-2006 in DIVERGE 3 with the 1991-datasets of HCV 1a and 1b to detect the implication of type II functional divergence at the individual sites (see Materials and methods) (Gu 2006; Gu et al. 2013). Similar to the results of the MK test (Table 1), differential patterns of divergence were found for the ten individual proteins' coding sequences (Table 2). The coefficient of type II functional divergence (θ_{II}) values on proteins E1, NS2, NS3, NS4B, NS5A, and NS5B were larger than 0 with significance ($P < 0.05$) (Table 2). In contrast, the results on the other proteins did not show a similar pattern of evolution. Furthermore, the site-specific score based on the posterior probability ratio showed that quite a number of amino acid sites, the majority of

which are within E1, NS5A, and NS5B, are related to the functional divergence at $P \geq 0.95$ level (Table 2). The amino acids at these sites present remarkable difference in physiochemical properties (e.g., charge, hydrophobic) between the subtypes 1a and 1b, while all of these sites are conserved within each of these subtypes (Table S4).

To further substantiate the result on the 1991-datasets, we repeated the test using the 2001-datasets. Generally consistent with those of 1991-datasets, the results with five proteins (E1, NS3, NS4B, NS5A, and NS5B) exhibited a significant deviation from neutrality in favor of the type II functional divergence (Table S5). The site-specific scores showed that the type II functional divergence might have contributed to adaptive changes at a number of sites with a probability ≥ 0.95 (Table S5). The major divergence-related sites are localized within the proteins E1, NS4B, NS5A, and NS5B. It is of note that a number of the identified sites were shared between the results with the 1991- and 2001-datasets (Fig. S2). Taken together, these observations suggested that the type II functional divergence had most likely occurred on the proteins E1, NS4B, NS5A, and NS5B with the split of HCV subtypes 1a and 1b and the substitutions at a number of sites could be relevant to the functional divergence.

Discussion

In this study, we characterized the genome-wide genetic variation between HCV

subtypes 1a and 1b using MK and DIVERGE functional divergence tests. The results of the MK test suggested that adaptive diversification had played an important role in shaping the variation on multiple individual proteins (Table 1 and S3). The MK test is robust to most demographic factors (see (Aguileta et al. 2009; Eyre-Walker 2006; Nielsen 2001) and references therein), with the exception that the slightly deleterious nonsynonymous mutations can be fixed during a rapid expansion in population size and thereby lead to artifactual evidence of adaptive evolution (Eyre-Walker 2006). In terms of HCV subtypes 1a and 1b, it has been shown that they expanded explosively recently (Magiorkinis et al. 2009). These findings raised further concerns about the impact of the demographic factor on the evolution of these subtypes. A demographic factor, such as expansion in population size, should have similar effects on the different genomic regions. In the present study, however, the results of the MK test on the coding sequences of core, E2, p7, NS3 and NS4A showed an opposite evolutionary pattern to those of E1, NS4B, NS5A, and NS5B (Table 1). Thus, the effect of the potential population expansion could be roughly ruled out - that is, the excess of fixed replacement in E1, NS4B, NS5A, and NS5B between subtypes is most likely to be a consequence of adaptive selection rather than a demographic factor. Moreover, the method Gu-2006 in DIVERGE suggested that the major functional divergence-related sites are distributed on the proteins E1, NS4B, NS5A, and NS5B (Table 2 and S5). In summary, the combined statistical results consistently indicated that the functional divergence took place with the split of the HCV genotype 1 progenitor into these subtypes.

275 Because the overall structures of HCV proteins E1, NS4B and NS5A are still lacking,
276 and their functions have been studied only to a limited extent, it is presently unknown
277 whether and how the variations at the divergence-related sites influence the function of
278 these proteins and clinical properties, including infectivity, virulence, and replication
279 potential, of HCV. For E1, it is noteworthy that the diversification at the adaptively
280 selected site 59 contributes to the difference of N-glycosylation at this site between
281 subtypes 1a and 1b (Table S5). Four sites for subtype 1a and five sites for 1b sites on
282 E1 are modified by N-linked glycosylation (Goffard and Dubuisson 2003). The
283 N-glycans of these sites have been shown to serve essential roles in virion assembly,
284 infectivity, core release, and evasion of the host immune response (Helle et al. 2011; Tarr
285 et al. 2015). Thus, the presence of adaptive evolution at site 59 might be a reflection of
286 alteration in these functions.

287 NS5A contains a N-terminal amphipathic alpha helix (amino acids 5–25) followed by
288 three structural domains (I, II, and III) (Penin et al. 2004; Tellinghuisen et al. 2004).
289 The structures of N-terminal alpha helix and domain I (AA 28-213) are relatively well
290 studied (Lambert et al. 2014; Love et al. 2009; Penin et al. 2004; Tellinghuisen et al.
291 2005). The adaptively selected sites 14, 24, and 25 are located on the alpha helix (Table
292 2 and S5). Of these, sites 14 and 24 are distributed on the hydrophobic-hydrophilic
293 borders. It has been suggested that the helix could not be buried deeply in the
294 hydrophobic core of the membrane and the side chains of the charged/polar amino acids
295 along the borders interact with the polar head of phospholipids or membrane proteins

(Penin et al. 2004). Consistent with the previous functional perspective, the adaptive changes at these sites might favor the interaction between the helix and different membrane-associated molecules in host cells (Dopico and Tigyi 2007; Dowhan 1997; Kjellen and Lindahl 1991).

Seven adaptively selected sites (30, 56, 58, 83, 101, 114, and 146) are located on the structural domain I of NS5A, and the majority of these sites fall into two functional regions (Table 2, S5, and Fig. S3). Sites 56, 58, and 83 are in the basic surface near the N-terminus (Fig. S3A). Considering the amphipathic alpha helix close to the N-terminus, the basic surface of domain I is likely in contact with the negatively charged head groups of the membrane (Tellinghuisen et al. 2005). Thus, consistent with that of sites 14 and 24 in the alpha helix, the amino acid changes at these sites might have constituted an adaptation to attach to different membrane-associated molecules. The adaptively selected sites 101, 114, and 146 are located on a contiguous patch which spans overall surface of domain I (Fig. S3B) (Tellinghuisen et al. 2005). Taking into account a 'pocket' and electrostatic potential plot generated on the surface, it has been suggested that the patch might represent an important molecular interaction surface (Love et al. 2009). In this scenario, considering the diverse interactions between NS5A and a large variety of partners (Macdonald and Harris 2004; Ross-Thriepland and Harris 2015), the existence of such adaptive changes within this region led us to propose that these identified sites could be responsible for the recognizing specificity of NS5A in interaction with different viral components and/or host cellular factors.

The RNA-dependent RNA polymerase (RdRp) NS5B is an essential component of the replicase complex (RC) in all single stranded RNA viruses, including HCV. Our results indicated that a number of NS5B sites may be related to the functional divergence between the subtypes 1a and 1b (Table 2 and S5). Structurally, the majority of identified sites cluster in four interesting regions: (1) on or nearby the β -hairpin (sites 436, 441 and 446) (Fig. S4 A and B) (Ago et al. 1999; Bressanelli et al. 1999; Lesburg et al. 1999), (2) at the entrance of the NTP channel (sites 307 and 329) (Fig. S4 A and B) (Mosley et al. 2012; O'Farrell et al. 2003), (3) on or close to a concave surface shaped by three helices (sites 77, 81, 181, 238, 246 and 248) (Fig. S4 A and C), and (4) at the top of the fingers domain (sites 111, 112 and 120) (Fig. S4 A).

Functionally, the β -hairpin might act as an initiation platform to ensure initiation of RNA synthesis from the 3' terminus of the template RNA (Ago et al. 1999; Cheney et al. 2002; Hong et al. 2001). It has been reported that substitutions at sites 436 and 446 may influence the RNA synthesis activity (Chinnaswamy et al. 2008). In addition, cyclophilin B, a host cellular factor that enhances HCV replication through modulating NS5B activity, interacts with the amino acids 398–469 (Fernandes et al. 2007). Thus, one potential explanation for this distribution feature is that the changes at these sites might contribute to the alteration of interacting behavior between polymerase NS5B and replication-related factors and/or template.

The identified sites 307 and 329 are located near the entrance to the NTP channel. Of note, they are also located at the base of an internal cavity, which has been suggested to

serve an important role in the structural step during the polymerase cycle (O'Farrell et al. 2003). With regard to the concave surface in which six identified sites lie (Fig. S4 A and C), the overall localization and secondary structure elements of this domain are well conserved in the homologues of viruses in the family *Flaviviridae* (Ago et al. 1999; Bressanelli et al. 1999; Choi et al. 2004; Lesburg et al. 1999; Malet et al. 2007; Yap et al. 2007). Interestingly, it has been reported that sites 202 and 367 of the classical swine fever virus NS5B, at which the amino acid changes are involved in functional divergence with the split of the classical swine fever virus major groups, fall into the homologous domain (Li and Yang 2015). The function of this surface remains, to some extent, elusive. In flaviviruses (dengue and West Nile viruses), it has been suggested that the surface could be involved in accommodating the N-terminal methyltransferase domain of RdRp (Yap et al. 2007). However, this domain is unique to flaviviruses and methylation events are required only for their genome replication (Choi and Rossmann 2009). Alternatively, the surface might be preserved in *Flaviviridae* RdRp to interact with other components of the RC. Viral NS1, NS2A, NS3, NS4A, and NS5A may function as part of the RC, which has been characterized extensively in the viruses in the family *Flaviviridae* (Brinton 2013; Grassmann et al. 2001; Ishido et al. 1998; Johansson et al. 2001; Khromykh et al. 2000; Koch and Bartenschlager 1999; Mackenzie et al. 1998; Neddermann et al. 1999; Sheng et al. 2010; Shirota et al. 2002; Westaway et al. 1997). Furthermore, several host cellular factors have been found to affect the replication of viruses, such as cyclophilin A and B for HCV (Fernandes et al. 2007; Nakagawa et al.

359 2005). Given this information, the residue changes at these sites, perhaps also including
360 other functional divergence-related ones at the surface of polymerase, would be linked to
361 the diversification of the interactions between NS5B and different adjacent molecules.

362 In conclusion, these results provided helpful hints about the decisive factors underlying
363 the phenotypic functional divergence of HCV 1a and 1b. It will be interesting to study
364 the functional significance of these identified changes. If the hypotheses are supported,
365 it might extend our knowledge regarding the function of these candidate proteins and
366 nature of epidemiological divergence between HCV 1a and 1b.

Draft

Appendix A. Supplementary data

Five supplementary tables and four supplementary figures are available:

Table S1. HCV subtype 1a sequences used for this study.

Table S2. HCV subtype 1b sequences used for this study.

Table S3. McDonald–Kreitman test on HCV subtypes 1a and 1b individual coding sequences with the strains collected up to year 2001.

Table S4. The amino acids at the Type-II divergence-related sites estimated with the strains collected up to year 1991.

Table S5. The estimates of the type-II functional divergences of individual coding sequences between HCV subtypes 1a and 1b with the strains collected up to year 2001.

Figure S1.

Maximum likelihood gene trees for the complete coding sequences of NS5B extracted from (A) 362 subtype 1a and (B) 165 subtype 1b genomes represented in Table S1 and S2.

The bootstrap values below 90% are not reported. Solid upward triangles highlighted in purple and the tips of strains show 1991-datasets recorded on the year of isolation, which

were used for evolutionary analysis (see Table S1 and S2). The tips are labeled with

GenBank accession numbers, country codes, and date of isolation. Country codes

officially assigned in ISO 3166-1: JP, Japan; US, United States of America

(http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2).

Figure S2

Similarity of the estimations of divergence-related sites on (A) E1, (B) NS4B, (C) NS5A, and (D) NS5B identified with 1991- and 2001-datasets. The overlapping regions represent divergence-related sites shared between two estimations.

Figure S3

Stick model showing the locations of divergence-related sites in the NS5A domain I dimer (A and B) (Tellinghuisen et al. 2005). In (A) and (B), the residues of identified sites are shown on the three-dimensional structures in purple. The coordinated zinc atoms are indicated in yellow.

Figure S4

Stick model showing the locations of divergence-related sites in the NS5B structure (Gentles et al. 2011). In (A), (B), and (C), the residues of identified sites are shown on the three-dimensional structures in purple.

Acknowledgements

We thank anonymous referees for their valuable comments and editor for his helpful English editing that led to improvement of the manuscript.

References

- Abdel-Hakeem, M.S., and Shoukry, N.H. 2014. Protective immunity against hepatitis C: many shades of gray. *Frontiers in immunology* **5**: 274.
- Abrignani, S., Houghton, M., and Hsu, H.H. 1999. Perspectives for a vaccine against hepatitis C virus. *Journal of hepatology* **31 Suppl 1**: 259-263.
- Ago, H., Adachi, T., Yoshida, A., Yamamoto, M., Habuka, N., Yatsunami, K., and Miyano, M. 1999. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Structure* **7**(11): 1417-1426.
- Aguileta, G., Refregier, G., Yockteng, R., Fournier, E., and Giraud, T. 2009. Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect Genet Evol* **9**(4): 656-670.
- Alter, M.J., Kruszon-Moran, D., Nainan, O.V., McQuillan, G.M., Gao, F., Moyer, L.A., Kaslow, R.A., and Margolis, H.S. 1999. The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. *The New England journal of medicine* **341**(8): 556-562.
- Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R.L., Mathieu, M., De Francesco, R., and Rey, F.A. 1999. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proceedings of the National Academy of Sciences of the United States of America* **96**(23): 13034-13039.
- Brinton, M.A. 2013. Replication cycle and molecular biology of the West Nile virus.

- 428 Viruses **6**(1): 13-53.
- 429 Charlesworth, J., and Eyre-Walker, A. 2008. The McDonald-Kreitman test and slightly
430 deleterious mutations. *Molecular biology and evolution* **25**(6): 1007-1015.
- 431 Cheney, I.W., Naim, S., Lai, V.C., Dempsey, S., Bellows, D., Walker, M.P., Shim, J.H.,
432 Horscroft, N., Hong, Z., and Zhong, W. 2002. Mutations in NS5B polymerase of hepatitis
433 C virus: impacts on in vitro enzymatic activity and viral RNA replication in the
434 subgenomic replicon cell culture. *Virology* **297**(2): 298-306.
- 435 Chinnaswamy, S., Yarbrough, I., Palaninathan, S., Kumar, C.T., Vijayaraghavan, V.,
436 Demeler, B., Lemon, S.M., Sacchettini, J.C., and Kao, C.C. 2008. A locking mechanism
437 regulates RNA synthesis and host protein interaction by the hepatitis C virus polymerase.
438 *The Journal of biological chemistry* **283**(29): 20535-20546.
- 439 Choi, K.H., Groarke, J.M., Young, D.C., Kuhn, R.J., Smith, J.L., Pevear, D.C., and
440 Rossmann, M.G. 2004. The structure of the RNA-dependent RNA polymerase from
441 bovine viral diarrhea virus establishes the role of GTP in de novo initiation. *Proceedings*
442 *of the National Academy of Sciences of the United States of America* **101**(13):
443 4425-4430.
- 444 Choi, K.H., and Rossmann, M.G. 2009. RNA-dependent RNA polymerases from
445 *Flaviviridae*. *Current opinion in structural biology* **19**(6): 746-751.
- 446 Choo, Q.L., Kuo, G., Weiner, A.J., Overby, L.R., Bradley, D.W., and Houghton, M. 1989.
447 Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis
448 genome. *Science (New York, N.Y)* **244**(4902): 359-362.

- 449 Dopico, A.M., and Tigyi, G.J. 2007. A Glance at the Structural and Functional Diversity
450 of Membrane Lipids. Humana Press.
- 451 Dore, G.J., Law, M., MacDonald, M., and Kaldor, J.M. 2003a. Epidemiology of hepatitis
452 C virus infection in Australia. *J Clin Virol* **26**(2): 171-184.
- 453 Dore, G.J., MacDonald, M., Law, M.G., and Kaldor, J.M. 2003b. Epidemiology of
454 hepatitis C virus infection in Australia. *Australian family physician* **32**(10): 796-798.
- 455 Dowhan, W. 1997. Molecular basis for membrane phospholipid diversity: why are there
456 so many lipids? *Annual review of biochemistry* **66**: 199-232.
- 457 Drucker, E., Alcabes, P.G., and Marx, P.A. 2001. The injection century: massive unsterile
458 injections and the emergence of human pathogens. *Lancet* **358**(9297): 1989-1992.
- 459 Esteban, J.I., Saucedo, S., and Quer, J. 2008. The changing epidemiology of hepatitis C
460 virus infection in Europe. *Journal of hepatology* **48**(1): 148-162.
- 461 Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends in ecology &*
462 *evolution* **21**(10): 569-575.
- 463 Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J.C., Strazzera, A.,
464 Chien, D.Y., Munoz, S.J., Balestrieri, A., Purcell, R.H., and Alter, H.J. 2000. The
465 outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science*
466 (New York, N.Y. **288**(5464): 339-344.
- 467 Fay, J.C., Wyckoff, G.J., and Wu, C.I. 2001. Positive and negative selection on the human
468 genome. *Genetics* **158**(3): 1227-1234.
- 469 Fernandes, F., Poole, D.S., Hoover, S., Middleton, R., Andrei, A.C., Gerstner, J., and

- 470 Striker, R. 2007. Sensitivity of hepatitis C virus to cyclosporine A depends on
471 nonstructural proteins NS5A and NS5B. *Hepatology* (Baltimore, Md **46**(4): 1026-1033.
- 472 Gentles, R.G., Sheriff, S., Beno, B.R., Wan, C., Kish, K., Ding, M., Zheng, X., Chupak,
473 L., Poss, M.A., Witmer, M.R., Morin, P., Wang, Y.K., Rigat, K., Lemm, J., Voss, S., Liu,
474 M., Pelosi, L., Roberts, S.B., Gao, M., and Kadow, J.F. 2011. Investigation of the mode
475 of binding of a novel series of
476 N-benzyl-4-heteroaryl-1-(phenylsulfonyl)piperazine-2-carboxamides to the hepatitis C
477 virus polymerase. *Bioorganic & medicinal chemistry letters* **21**(8): 2212-2215.
- 478 Goffard, A., and Dubuisson, J. 2003. Glycosylation of hepatitis C virus envelope proteins.
479 *Biochimie* **85**(3-4): 295-301.
- 480 Grassmann, C.W., Isken, O., Tautz, N., and Behrens, S.E. 2001. Genetic analysis of the
481 pestivirus nonstructural coding region: defects in the NS5A unit can be complemented in
482 trans. *Journal of virology* **75**(17): 7791-7802.
- 483 Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication.
484 *Molecular biology and evolution* **16**(12): 1664-1674.
- 485 Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional
486 divergence. *Molecular biology and evolution* **18**(4): 453-464.
- 487 Gu, X. 2006. A simple statistical method for estimating type-II (cluster-specific)
488 functional divergence of protein sequences. *Molecular biology and evolution* **23**(10):
489 1937-1945.
- 490 Gu, X., Zou, Y., Su, Z., Huang, W., Zhou, Z., Arendsee, Z., and Zeng, Y. 2013. An update

of DIVERGE software for functional divergence analysis of protein family. *Molecular biology and evolution* **30**(7): 1713-1719.

Helle, F., Duverlie, G., and Dubuisson, J. 2011. The hepatitis C virus glycan shield and evasion of the humoral immune response. *Viruses* **3**(10): 1909-1932.

Hnatyszyn, H.J. 2005. Chronic hepatitis C and genotyping: the clinical significance of determining HCV genotypes. *Antiviral therapy* **10**(1): 1-11.

Hong, Z., Cameron, C.E., Walker, M.P., Castro, C., Yao, N., Lau, J.Y., and Zhong, W. 2001. A novel mechanism to ensure terminal initiation by hepatitis C virus NS5B polymerase. *Virology* **285**(1): 6-11.

Hoofnagle, J.H. 2002. Course and outcome of hepatitis C. *Hepatology* (Baltimore, Md **36**(5 Suppl 1): S21-29.

<http://www.who.int/mediacentre/factsheets/fs164/en/>. 2014. Hepatitis C. *In* Fact sheet N°164. World Health Organization.

Ishido, S., Fujita, T., and Hotta, H. 1998. Complex formation of NS5B with NS3 and NS4A proteins of hepatitis C virus. *Biochemical and biophysical research communications* **244**(1): 35-40.

Jackowiak, P., Kuls, K., Budzko, L., Mania, A., Figlerowicz, M., and Figlerowicz, M. 2014. Phylogeny and molecular evolution of the hepatitis C virus. *Infect Genet Evol* **21**: 67-82.

Johansson, M., Brooks, A.J., Jans, D.A., and Vasudevan, S.G. 2001. A small region of the dengue virus-encoded RNA-dependent RNA polymerase, NS5, confers interaction with

512 both the nuclear transport receptor importin-beta and the viral helicase, NS3. The Journal
513 of general virology **82**(Pt 4): 735-745.

514 Kamal, S.M., and Nasser, I.A. 2008. Hepatitis C genotype 4: What we know and what we
515 don't yet know. Hepatology (Baltimore, Md **47**(4): 1371-1383.

516 Khromykh, A.A., Sedlak, P.L., and Westaway, E.G. 2000. cis- and trans-acting elements
517 in flavivirus RNA replication. Journal of virology **74**(7): 3253-3263.

518 King, A.M.Q., Adams, M.J., Carstens, E.B., and Lefkowitz, E.J. 2012. Family -
519 Flaviviridae. *In* Virus Taxonomy. Elsevier, San Diego. pp. 1003-1020.

520 Kjellen, L., and Lindahl, U. 1991. Proteoglycans: structures and interactions. Annual
521 review of biochemistry **60**: 443-475.

522 Koch, J.O., and Bartenschlager, R. 1999. Modulation of hepatitis C virus NS5A
523 hyperphosphorylation by nonstructural proteins NS3, NS4A, and NS4B. Journal of
524 virology **73**(9): 7138-7146.

525 Kuiken, C., Hraber, P., Thurmond, J., and Yusim, K. 2008. The hepatitis C sequence
526 database in Los Alamos. Nucleic acids research **36**(Database issue): D512-516.

527 Kuiken, C., and Simmonds, P. 2009. Nomenclature and numbering of the hepatitis C
528 virus. Methods in molecular biology (Clifton, N.J) **510**: 33-53.

529 Lambert, S.M., Langley, D.R., Garnett, J.A., Angell, R., Hedgethorne, K., Meanwell,
530 N.A., and Matthews, S.J. 2014. The crystal structure of NS5A domain 1 from genotype
531 1a reveals new clues to the mechanism of action for dimeric HCV inhibitors. Protein Sci
532 **23**(6): 723-734.

- 533 Lanave, C., Preparata, G., Saccone, C., and Serio, G. 1984. A new method for calculating
534 evolutionary substitution rates. *Journal of molecular evolution* **20**(1): 86-93.
- 535 Lavanchy, D. 2011. Evolving epidemiology of hepatitis C virus. *Clin Microbiol Infect*
536 **17**(2): 107-115.
- 537 Lesburg, C.A., Cable, M.B., Ferrari, E., Hong, Z., Mannarino, A.F., and Weber, P.C. 1999.
538 Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a
539 fully encircled active site. *Nature structural biology* **6**(10): 937-943.
- 540 Li, W.-H. 1983. Evolution of duplicate genes and pseudo genes. *In* *Evolution of Genes*
541 *and Proteins. Edited by M. Nei and R.K. Koehn. Sinauer Associates, Sunderland, Mass.*
542 *pp. 14-37.*
- 543 Li, W.-H. 1997. *Molecular evolution. . Sinauer Associates, Sunderland, MA.*
- 544 Li, Y., and Yang, Z. 2015. Episodic adaptive diversification of classical swine fever virus
545 RNA-dependent RNA polymerase NS5B. *Canadian journal of microbiology* **61**(12):
546 948-954.
- 547 Love, R.A., Brodsky, O., Hickey, M.J., Wells, P.A., and Cronin, C.N. 2009. Crystal
548 structure of a novel dimeric form of NS5A domain I protein from hepatitis C virus.
549 *Journal of virology* **83**(9): 4395-4403.
- 550 Lu, L., Tong, W., Gu, L., Li, C., Lu, T., Tee, K.K., and Chen, G. 2013. The current
551 hepatitis C virus prevalence in China may have resulted mainly from an officially
552 encouraged plasma campaign in the 1990s: a coalescence inference with genetic
553 sequences. *Journal of virology* **87**(22): 12041-12050.

- 554 Macdonald, A., and Harris, M. 2004. Hepatitis C virus NS5A: tales of a promiscuous
555 protein. *The Journal of general virology* **85**(Pt 9): 2485-2502.
- 556 Mackenzie, J.M., Khromykh, A.A., Jones, M.K., and Westaway, E.G. 1998. Subcellular
557 localization and some biochemical properties of the flavivirus Kunjin nonstructural
558 proteins NS2A and NS4A. *Virology* **245**(2): 203-215.
- 559 Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S.Y., Shapiro, B., Pybus, O.G.,
560 Allain, J.P., and Hatzakis, A. 2009. The global spread of hepatitis C virus 1a and 1b: a
561 phylodynamic and phylogeographic analysis. *PLoS medicine* **6**(12): e1000198.
- 562 Malet, H., Egloff, M.P., Selisko, B., Butcher, R.E., Wright, P.J., Roberts, M., Gruez, A.,
563 Sulzenbacher, G., Vornrhein, C., Bricogne, G., Mackenzie, J.M., Khromykh, A.A.,
564 Davidson, A.D., and Canard, B. 2007. Crystal structure of the RNA polymerase domain
565 of the West Nile virus non-structural protein 5. *The Journal of biological chemistry*
566 **282**(14): 10678-10689.
- 567 Manns, M.P., McHutchison, J.G., Gordon, S.C., Rustgi, V.K., Shiffman, M., Reindollar,
568 R., Goodman, Z.D., Koury, K., Ling, M., and Albrecht, J.K. 2001. Peginterferon alfa-2b
569 plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of
570 chronic hepatitis C: a randomised trial. *Lancet* **358**(9286): 958-965.
- 571 McDonald, J.H. 2009. *Handbook of biological statistics*. Sparky House Publishing
572 Baltimore, MD.
- 573 McDonald, J.H., and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in
574 *Drosophila*. *Nature* **351**(6328): 652-654.

- 575 Mohd Hanafiah, K., Groeger, J., Flaxman, A.D., and Wiersma, S.T. 2013. Global
576 epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to
577 HCV seroprevalence. *Hepatology (Baltimore, Md)* **57**(4): 1333-1342.
- 578 Moradpour, D., Penin, F., and Rice, C.M. 2007. Replication of hepatitis C virus. *Nature*
579 *reviews* **5**(6): 453-463.
- 580 Mosley, R.T., Edwards, T.E., Murakami, E., Lam, A.M., Grice, R.L., Du, J., Sofia, M.J.,
581 Furman, P.A., and Otto, M.J. 2012. Structure of hepatitis C virus polymerase in complex
582 with primer-template RNA. *Journal of virology* **86**(12): 6503-6511.
- 583 Muir, A.J. 2014. The rapid evolution of treatment strategies for hepatitis C. *The American*
584 *journal of gastroenterology* **109**(5): 628-635; quiz 636.
- 585 Nainan, O.V., Alter, M.J., Kruszon-Moran, D., Gao, F.X., Xia, G., McQuillan, G., and
586 Margolis, H.S. 2006. Hepatitis C virus genotypes and viral concentrations in participants
587 of a general population survey in the United States. *Gastroenterology* **131**(2): 478-484.
- 588 Nakagawa, M., Sakamoto, N., Tanabe, Y., Koyama, T., Itsui, Y., Takeda, Y., Chen, C.H.,
589 Kakinuma, S., Oooka, S., Maekawa, S., Enomoto, N., and Watanabe, M. 2005.
590 Suppression of hepatitis C virus replication by cyclosporin a is mediated by blockade of
591 cyclophilins. *Gastroenterology* **129**(3): 1031-1041.
- 592 Neddermann, P., Clementi, A., and De Francesco, R. 1999. Hyperphosphorylation of the
593 hepatitis C virus NS5A protein requires an active NS3 protease, NS4A, NS4B, and NS5A
594 encoded on the same polyprotein. *Journal of virology* **73**(12): 9984-9991.
- 595 Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

- 596 Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity*
597 **86**(Pt 6): 641-647.
- 598 Nousbaum, J., Polyak, S.J., Ray, S.C., Sullivan, D.G., Larson, A.M., Carithers, R.L., Jr.,
599 and Gretch, D.R. 2000. Prospective characterization of full-length hepatitis C virus NS5A
600 quasispecies during induction and combination antiviral therapy. *Journal of virology*
601 **74**(19): 9028-9038.
- 602 O'Farrell, D., Trowbridge, R., Rowlands, D., and Jager, J. 2003. Substrate complexes of
603 hepatitis C virus RNA polymerase (HC-J4): structural evidence for nucleotide import and
604 de-novo initiation. *Journal of molecular biology* **326**(4): 1025-1035.
- 605 Pawlotsky, J.M., Tsakiris, L., Roudot-Thoraval, F., Pellet, C., Stuyver, L., Duval, J., and
606 Dhumeaux, D. 1995. Relationship between hepatitis C virus genotypes and sources of
607 infection in patients with chronic hepatitis C. *The Journal of infectious diseases* **171**(6):
608 1607-1610.
- 609 Penin, F., Brass, V., Appel, N., Ramboarina, S., Montserret, R., Ficheux, D., Blum, H.E.,
610 Bartenschlager, R., and Moradpour, D. 2004. Structure and function of the membrane
611 anchor domain of hepatitis C virus nonstructural protein 5A. *The Journal of biological*
612 *chemistry* **279**(39): 40835-40843.
- 613 Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C., and Harvey, P.H.
614 2001. The epidemic behavior of the hepatitis C virus. *Science (New York, N.Y)* **292**(5525):
615 2323-2325.
- 616 Pybus, O.G., Cochrane, A., Holmes, E.C., and Simmonds, P. 2005. The hepatitis C virus

- 617 epidemic among injecting drug users. *Infect Genet Evol* **5**(2): 131-139.
- 618 Ross-Thriepland, D., and Harris, M. 2015. Hepatitis C virus NS5A: enigmatic but still
619 promiscuous 10 years on! *The Journal of general virology* **96**(Pt 4): 727-738.
- 620 Sheng, C., Zhu, Z., Yu, J., Wan, L., Wang, Y., Chen, J., Gu, F., and Xiao, M. 2010.
621 Characterization of NS3, NS5A and NS5B of classical swine fever virus through
622 mutation and complementation analysis. *Veterinary microbiology* **140**(1-2): 72-80.
- 623 Shiota, Y., Luo, H., Qin, W., Kaneko, S., Yamashita, T., Kobayashi, K., and Murakami, S.
624 2002. Hepatitis C virus (HCV) NS5A binds RNA-dependent RNA polymerase (RdRP)
625 NS5B and modulates RNA-dependent RNA polymerase activity. *The Journal of*
626 *biological chemistry* **277**(13): 11149-11155.
- 627 Sievert, W., Altraif, I., Razavi, H.A., Abdo, A., Ahmed, E.A., Alomair, A., Amarapurkar,
628 D., Chen, C.H., Dou, X., El Khayat, H., Elshazly, M., Esmat, G., Guan, R., Han, K.H.,
629 Koike, K., Largen, A., McCaughan, G., Mogawer, S., Monis, A., Nawaz, A., Piratvisuth,
630 T., Sanai, F.M., Sharara, A.I., Sibbel, S., Sood, A., Suh, D.J., Wallace, C., Young, K., and
631 Negro, F. 2011. A systematic review of hepatitis C virus epidemiology in Asia, Australia
632 and Egypt. *Liver Int* **31 Suppl 2**: 61-80.
- 633 Simmonds, P. 1997. Clinical relevance of hepatitis C virus genotypes. *Gut* **40**(3):
634 291-293.
- 635 Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., and
636 Simmonds, P. 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67
637 subtypes: updated criteria and genotype assignment web resource. *Hepatology* (Baltimore,

- 638 Md) **59**(1): 318-327.
- 639 Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. 2011.
- 640 MEGA5: molecular evolutionary genetics analysis using maximum likelihood,
- 641 evolutionary distance, and maximum parsimony methods. *Molecular biology and*
- 642 *evolution* **28**(10): 2731-2739.
- 643 Tarr, A.W., Khera, T., Hueging, K., Sheldon, J., Steinmann, E., Pietschmann, T., and
- 644 Brown, R.J. 2015. Genetic Diversity Underlying the Envelope Glycoproteins of Hepatitis
- 645 C Virus: Structural and Functional Consequences and the Implications for Vaccine
- 646 Design. *Viruses* **7**(7): 3995-4046.
- 647 Tellinghuisen, T.L., Marcotrigiano, J., Gorbalenya, A.E., and Rice, C.M. 2004. The NS5A
- 648 protein of hepatitis C virus is a zinc metalloprotein. *The Journal of biological chemistry*
- 649 **279**(47): 48576-48587.
- 650 Tellinghuisen, T.L., Marcotrigiano, J., and Rice, C.M. 2005. Structure of the zinc-binding
- 651 domain of an essential component of the hepatitis C virus replicase. *Nature* **435**(7040):
- 652 374-379.
- 653 Thomas, D.L. 2013. Global control of hepatitis C: where challenge meets opportunity.
- 654 *Nature medicine* **19**(7): 850-858.
- 655 Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The
- 656 CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment
- 657 aided by quality analysis tools. *Nucleic acids research* **25**(24): 4876-4882.
- 658 Verachai, V., Phutiprawan, T., Theamboonlers, A., Chinchai, T., Tanprasert, S., Haagmans,

- 659 B.L., Osterhaus, A.D., and Poovorawan, Y. 2002. Prevalence and genotypes of hepatitis C
660 virus infection among drug addicts and blood donors in Thailand. The Southeast Asian
661 journal of tropical medicine and public health **33**(4): 849-851.
- 662 Webster, D.P., Klenerman, P., and Dusheiko, G.M. 2015. Hepatitis C. Lancet **385**(9973):
663 1124-1135.
- 664 Westaway, E.G., Mackenzie, J.M., Kenney, M.T., Jones, M.K., and Khromykh, A.A. 1997.
665 Ultrastructure of Kunjin virus-infected cells: colocalization of NS1 and NS3 with
666 double-stranded RNA, and of NS2B with NS3, in virus-induced membrane structures.
667 Journal of virology **71**(9): 6650-6661.
- 668 Yap, T.L., Xu, T., Chen, Y.L., Malet, H., Egloff, M.P., Canard, B., Vasudevan, S.G., and
669 Lescar, J. 2007. Crystal structure of the dengue virus RNA-dependent RNA polymerase
670 catalytic domain at 1.85-angstrom resolution. Journal of virology **81**(9): 4753-4765.
- 671 Zein, N.N. 2000. Clinical significance of hepatitis C virus genotypes. Clinical
672 microbiology reviews **13**(2): 223-235.

673

674

675

Tables

Table 1. McDonald–Kreitman test on HCV subtypes 1a and 1b individual coding sequences with the strains collected up to year 1991.

Subtype 1a vs. subtype 1b		Fixed	Polymorphic
core	Replacement	2	8
	Synonymous	12	100
	<i>Probability^a</i>	0.603	
E1	Replacement	29	49
	Synonymous	39	157
	<i>Probability</i>	0.005**	
E2	Replacement	17	79
	Synonymous	57	305
	<i>Probability</i>	0.641	
p7	Replacement	6	20
	Synonymous	9	57
	<i>Probability</i>	0.348	
NS2	Replacement	24	73
	Synonymous	42	207
	<i>Probability</i>	0.127	
NS3	Replacement	20	99
	Synonymous	114	659
	<i>Probability</i>	0.581	

Table 1. McDonald–Kreitman test on HCV subtypes 1a and 1b individual coding sequences with the strains collected up to year 1991 (Table continued).

Subtype 1a vs. subtype 1b		Fixed	Polymorphic
NS4A			
	Replacement	1	8
	Synonymous	6	50
	<i>Probability</i>		1.000
NS4B			
	Replacement	29	43
	Synonymous	48	273
	<i>Probability</i>		< 0.001**
NS5A			
	Replacement	47	108
	Synonymous	83	398
	<i>Probability</i>		< 0.001**
NS5B			
	Replacement	49	119
	Synonymous	77	432
	<i>Probability</i>		< 0.001**

^a The significance according to the Fisher's exact test (two tailed). Double asterisks correspond to *Probability* < 0.01.

Table 2. The estimates of the type-II functional divergences of individual coding sequences between HCV subtypes 1a and 1b with the strains collected up to year 1991.

Subtype 1a vs. subtype 1b	Estimate (Z-score)	Divergence-related amino acid sites
core	$\theta_{II}^a = 0.007 \pm 0.018$ (0.389)	None (at $P_{II} \geq 0.95$)
E1	$\theta_{II} = 0.174 \pm 0.047$ (3.702**)	2, 17, 30, 55, 57, 59, 71, 79, 84, 97, 108, 144, 153, 157, 171, 184, 192 (at $0.99 > P_{II} \geq 0.95$)
E2	$\theta_{II} = 0.064 \pm 0.04$ (1.600)	57, 288, 325, 350 (at $0.99 > P_{II} \geq 0.95$)
p7	$\theta_{II} = 0.119 \pm 0.093$ (1.280)	28, 36, 59 (at $0.99 > P_{II} \geq 0.95$)
NS2	$\theta_{II} = 0.114 \pm 0.053$ (2.151*)	None (at $P_{II} \geq 0.95$)
NS3	$\theta_{II} = 0.047 \pm 0.018$ (2.611**)	None (at $P_{II} \geq 0.95$)
NS4A	$\theta_{II} = 0.036 \pm 0.063$ (0.571)	None (at $P_{II} \geq 0.95$)
NS4B	$\theta_{II} = 0.091 \pm 0.029$ (3.138**)	1, 85, 111, 255 (at $0.99 > P_{II} \geq 0.95$)
NS5A	$\theta_{II} = 0.130 \pm 0.030$ (4.333**)	14, 24, 25, 30, 56, 58, 83, 101, 114, 146, 288, 335, 387, 391, 407, 438 (at $0.99 > P_{II} \geq 0.95$)
NS5B	$\theta_{II} = 0.074 \pm 0.022$ (3.364**)	77, 81, 111, 112, 120, 181, 238, 246, 248, 272, 307, 329, 376, 436, 441, 446 (at $0.99 > P_{II} \geq 0.95$)

^a Coefficient of type-II functional divergence.

Single and double asterisks correspond to $P < 0.05$ and < 0.01 (Z-score test), respectively.