ELSEVIER

FEMS Microbiology Letters 214 (2002) 143–152

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

MiniReview

Genetic diversity and models of viral evolution for the hepatitis C virus

M.P.H. Stumpf [a,1,*], O.G. Pybus [b,1]

[a] *Department of Biology, University College London, Gower Street, London WC1E 6BT, UK*
[b] *Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

Received 28 June 2002; received in revised form 11 July 2002; accepted 11 July 2002

First published online 1 August 2002

Abstract

In this review we discuss the application of theoretical frameworks to the interpretation of viral gene sequence data, with particular reference to the hepatitis C virus (HCV). The increasing availability of such data means that it is now possible (and necessary) to proceed from simple qualitative models of viral evolution, to more quantitative frameworks based on statistical inference, notably population genetics and molecular phylogenetics. We argue that these approaches are invaluable tools to the virologist and are essential for understanding the dynamics of viral infection and the outcome of therapeutic strategies. We use several recent HCV data-sets to illustrate the methods. © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.

Keywords: RNA virus; Virus dynamics; Sequence evolution; Statistical genetics

1. Introduction

Traditionally, the taxonomy of viruses has not been straightforward. Very similar viruses can present radically different aetiologies and, vice versa, similar clinical outcomes may be caused by viruses differing hugely in their molecular structures, cellular dynamics and replication – the human hepatitis viruses are a prime example. We have come a long way from the first reports, more than 100 years ago, of unusual 'filterable' infectious agents, to the establishment of modern viral gene sequence databases. Significant progress has been made during the past 20 years through the application of sequencing technology and population genetic theory, and this path is likely to continue to provide the most fruitful route towards understanding the evolution of viruses. Viral taxonomy that is not based on evolutionary (i.e. DNA or RNA sequence) criteria may have to be revised and could hinder our understanding of infectious disease.

In this review we shall go beyond the level of phylogeny and discuss how population genetic and epidemiological processes affect the genetic diversity of the hepatitis C virus (HCV). HCV is currently the most widely studied and best understood positive-sense RNA virus, largely because of its enormous clinical importance. An estimated 170 million people worldwide are at risk of liver disease due to chronic HCV infection [1] and the virus is responsible for approximately 10 000 deaths per year in the United States [2]. The wealth of data being generated for HCV makes it a potential model system for studying genetic diversity in other positive-sense RNA viruses, despite the absence of a satisfactory animal model. Furthermore, HCV rarely recombines ([3] is the only example to date), which simplifies the application and interpretation of evolutionary analyses.

We show that an assessment of viral genetic diversity can have important implications for the clinical treatment of HCV, and that certain aspects of the aetiology of HCV infection can only be understood when genetic diversity

* Corresponding author. Tel.: +44 (207) 679 2263; Fax: +44 (207) 679 2887.
E-mail addresses: m.stumpf@ucl.ac.uk (M.P.H. Stumpf), oliver.pybus@zoo.ox.ac.uk (O.G. Pybus).

[1] The authors contributed equally to this work.

0378-1097/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.
PII: S0378-1097(02)00854-6

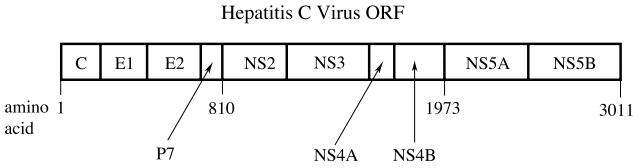

Fig. 1. The HCV genome contains a single open reading frame (ORF). The genes for structural proteins (C, E1, E2, P7) are situated towards the N-terminus of the ORF. Genes coding for proteins necessary for viral replication are found towards the C-terminus of the ORF.

and selective pressures are taken into account. Our discussion proceeds as follows. First we give a brief review of the theoretical frameworks that can be used to interpret viral genetic diversity. Second, we discuss the genetic diversity of HCV in the context of these frameworks, and outline its epidemiological and clinical relevance. Lastly, we describe how a simple population genetic model can be used to examine the relationship between HCV diversity and disease progression.

HCV is a single-stranded positive-sense RNA virus that belongs to the taxonomic family Flaviviridae and has a genome approximately 10 kb in length. All its proteins are encoded in a single open reading frame (see Fig. 1), with genes coding for structural proteins situated towards the N-terminus of the genome and non-structural genes located near the C-terminus. The structural genes code for the capsid protein (C, or Core) and the envelope glycoproteins (E1, E2). The first 27 amino acids of the E2 gene constitute the hyper variable region 1 (HVR1), which is the most variable region of the genome and appears to be involved in immune escape and disease progression. The non-structural genes code for a protease (NS2, NS3) and its cofactor (NS4A), a helicase (NS3), a protein of unknown function (NS4B), a phosphoprotein (NS5A), and an RNA-dependent RNA polymerase (NS5B). In addition, the HCV genome has 5′ and 3′ untranslated regions (UTRs) that are involved in the control of viral translation.

Upon cell entry, viral RNA is released and replicated in the cytoplasm. The polycistronic mRNA of the virus is translated into a single polyprotein precursor that is subsequently cleaved by a combination of viral and host proteases. The glycoproteins E1 and E2 are found in the endoplasmic reticulum, where new viral RNA appears to be enveloped. The resulting virions then bud from the endoplasmic reticulum and are released from the cell by exocytosis.

HCV replication via RNA-dependent RNA-polymerase is very error-prone and generates mutations at an estimated rate of 10^{-5} mutations per nucleotide per replication. This high mutation rate is the ultimate source of the virus' genetic diversity. The correct interpretation of genetic diversity should provide valuable information about the processes determining HCV transmission and disease progression.

2. Frameworks for understanding viral genetic diversity

Three mainstream frameworks exist for the analysis and interpretation of viral genetic diversity, (i) the quasi-species model, (ii) standard population genetics, and (iii) molecular phylogenetics. Before discussing recent experimental findings relating to HCV diversity, we first describe and compare these approaches.

2.1. The quasi-species model

Although originally conceived by Eigen [4] to model early pre-biotic evolutionary processes, the quasi-species model is frequently referred to in the viral literature, where it is usually used to denote populations containing significant genetic variation. However, the original mathematical formulation of the quasi-species goes much further than this. In this model, the quasi-species represents a very large set of replicating nucleotide sequences. Each sequence has an associated fitness, a constant value that corresponds to the reproductive potential of that variant. Replication is assumed to be error-prone and thus creates sequence diversity. The sequence with the highest fitness is called the master sequence, but it can only persist if its fitness (denoted a_0) exceeds a_1/Q, where a_1 is the fitness of the next most fit sequence, and Q is the error rate during replication. In other words, if $a_0 > a_1/Q$ is true then the master sequence will persist, if not, then the master sequence will go extinct and be replaced by its 'mutant tail' (the set of sequences that it produces during replication by mutation).

The dynamics of the quasi-species model have been exhaustively examined [5,6]. One of the central theoretical results is that the master sequence does not dominate the population if its 'mutant tail' has a low average fitness. If an alternative sequence exists with a mutant tail that has a higher average fitness, then this alternative sequence (and its mutant tail) will come to dominate the quasi-species population. Experimental evidence of such behaviour has been reported for the vesicular stomatitis virus [7], although this finding can also be explained using non-quasi-species concepts [8]. In short, the quasi-species evolves in a deterministic fashion towards a state where a balance between selection and mutation is achieved, although stochastic models which lead to similar results have been developed.

Because the quasi-species was originally conceived to describe the biochemical evolution of RNA sequences [4] it seemed natural to apply it to RNA viruses, and it has since served as a useful conceptual model. However, the quasi-species has a number of shortcomings that limit its applicability to natural viral populations: (i) it ignores effects of population size and neglects random genetic drift; (ii) it makes very strong assumptions about natural selection; in particular, it assumes that selection acts on the viral population as a whole. In contrast, conventional

population genetics assumes that selection acts separately on each individual genome. (iii) The quasi-species is assumed to be in mutation-selection equilibrium, whereas selection pressures in natural populations will fluctuate as a result of immune responses. (iv) Most importantly, it is very difficult to adapt the quasi-species theory to make quantitative inferences from sampled viral gene sequence data.

Thus the applicability of the quasi-species framework to natural populations must be carefully assessed [8], and the quasi-species concept should not be applied to biological systems whose essential dynamics are not solely controlled by mutation and preferential replication of some variants [6].

2.2. Population genetics

The central quantities in population genetics are allele frequencies. The frequency of an allele is influenced by a combination of processes, namely mutation, recombination, natural selection and demography. In general, population genetics treats these processes as stochastic in nature, and unlike the quasi-species model, statistical methods for inferring population genetic processes from genetic data are well developed [9,10]. In recent years, faster computers have greatly increased the power of inference methods based on coalescent theory [9]. Coalescent theory enables researchers to infer the parameters of genetic models from a small sample of gene sequences. Population genetic theory and inference are well developed disciplines, further details of which can be found in the literature [10,11].

The null hypothesis for population genetic inference is selective neutrality. The neutral theory of molecular evolution, originally developed by Kimura [12] and extended by Ohta [13], assumes that most mutations are either selectively neutral or slightly deleterious. Because genetic processes are stochastic and population sizes are finite, the spread of such mutations is affected by random factors. Under the neutral model, if generations are discrete and non-overlapping then allele frequencies vary from generation to generation according to a binomial sampling process. This process is called random genetic drift and will lead to fixation of a new neutral mutation in $2N$ generations on average (N is the population size). If the population size is sufficiently small then genetic drift can even bring disadvantageous mutations to fixation. Of course, selectively advantageous mutations usually become fixed, and this occurs much more quickly than the fixation of a neutral variant. However, even advantageous mutations are affected by genetic drift when they are rare and are occasionally lost from the population [6,7,9,10].

Considerable effort has gone into debating whether the neutral model is generally applicable, and we will not add to that effort here. The most important point for our current discussion is that the assumption of selective neutrality for most mutations serves as a useful, even necessary, null hypothesis for evolutionary hypothesis testing.

2.3. Molecular phylogenetics

Molecular phylogenetics is the most commonly used of the three frameworks described here and is used to reconstruct the shared history of sampled viral strains. The phylogenetic framework has two components, (i) a tree topology, representing the evolutionary relationships among the sequences, and (ii) a nucleotide substitution model, describing the processes by which the sequences have evolved. Both components can be estimated from sequence data using maximum likelihood inference methods (see [14] for a comprehensive review).

The branch lengths of a viral phylogeny typically represent genetic distances, i.e. estimates of the amount of evolutionary change among the sampled sequences. When used in this way phylogenies are simply intuitive graphical representations of the sequence data and are unconnected to population genetic models. However, population genetic concepts can be brought into the phylogenetic framework by analysing genetic distances in one of two ways.

First, normal genetic distances can be decomposed into synonymous distances (dS) and non-synonymous distances (dN) [15]. The former represents 'silent' nucleotide changes that do not alter the amino acid encoded, whereas the latter represents 'replacement' nucleotide changes that do result in a different amino acid. If the neutral theory is correct and all mutations are unselected, then synonymous and non-synonymous mutations will spread at the same rate and the ratio dN/dS will equal one. Positively selected mutations spread faster – and negatively selected mutations spread slower – than neutral mutations, so a dN/dS ratio greater than one is interpreted as evidence for the past action of positive selection, whereas a dN/dS ratio smaller than one is thought to result from negative selection. Several statistical methods based on these principles have been developed (e.g. [16]), but they are not applicable to every problem and can be difficult to interpret [17].

Second, genetic distances can be transformed into measures of time if the evolutionary rate of the sequences is known [6]. Adding a time scale to a phylogeny has many practical advantages, but also adds the implicit assumption that all mutations are fixed at the same rate (a molecular clock). This assumption is met under the neutral theory [6].

Population genetics and molecular phylogenetics are closely related [10,11] but differ in their emphasis on the evolutionary tree that relates sampled individuals. The central aim of molecular phylogenetics is the reconstruction of the tree itself, whereas in population genetic models the tree is implicitly represented but is not the focus of attention, and is thus considered as a 'nuisance variable' [9].

3. The genetic diversity of HCV

With over 100 complete HCV genomes currently available, our knowledge of the genetic diversity of HCV is considerable – perhaps unsurpassed by any other organism except HIV. Like many RNA viruses, HCV shows considerable genetic diversity at the nucleotide level, generated by the high mutation rate of the virus. We now discuss how this genetic diversity, when interpreted in the context of the theoretical frameworks described above, can provide insights into the processes that determine the epidemiological (among-host) and infection (within-host) dynamics of HCV.

3.1. Genetic diversity among infected individuals

In order to impose some structure on the global genetic diversity of HCV, a classification scheme based on molecular phylogenetic analysis of viral sequences has been developed [18]. The phylogeny shown in Fig. 2 illustrates the estimated relationships among all current HCV complete genome sequences. The tree has six main branches, labelled 1–6, corresponding to the six types (or 'genotypes') of HCV [18]. Each type is phylogenetically subdivided into a number of subtypes, labelled alphabetically in their order of discovery (thus subtype 2a was identified before subtype 2c and both belong to genotype 2). More than 50 subtypes have been reported to date and are normally identified on the basis of partial gene sequences from E1 and NS5B. Only a small proportion of identified subtypes have had their entire genomes sequenced, so Fig. 2 underestimates the genetic diversity of HCV within each genotype.

Complete genomes from different types differ at approximately 30–35% of nucleotide sites, whereas those from different subtypes (of the same genotype) typically differ at about 20–25% of nucleotides (although the subtypes of type 6 are more diverse than this) [19]. The scale bar in Fig. 2 shows that the average genetic distance among HCV types is approximately one substitution per nucleotide site. As only 30–35% of nucleotides actually differ, there is obviously considerable heterogeneity in evolutionary rates among nucleotide sites in the genome. This heterogeneity is the result of variable evolutionary constraints. The 5′ UTR contains extensive secondary RNA structure and is correspondingly the slowest evolving genomic region [20]. The next slowest region is the C (Core) gene, which evolves three times faster than the 5′ UTR. The envelope genes E1 and E2 constitute the most diverse genome region and evolve about nine times faster than the 5′ UTR [20], probably as a result of their presumed role in evading the host immune response. Evolutionary constraints are also evident at the codon level: third codon positions (where approximately 70% of possible mutations are synonymous) evolve nine times faster than second codon positions (where all mutations are non-synonymous).

The six HCV types are roughly genetically equidistant from each other and there is little statistical support for phylogenetic relationships among types, thus the root of the HCV phylogeny resembles a 'star' with six arms [19,20]. There is no obvious explanation for this pattern and none is likely to arise unless a non-human source population of HCV is found. Such populations have been found for many other members of HCV's taxonomic group, the *Flaviviridae*. Identification of the source of HCV should enable us to determine whether the six types originated from none, one, or several cross-species transmissions.

HCV types and subtypes exhibit complex patterns of geographic distribution, relative prevalence and modes of transmission that can be best understood by categorising them into three groups. The 'epidemic' group contains subtypes 1a, 1b, 2a, 2b and 3a, which are distributed globally and account for the majority of HCV infections worldwide [21,22]. The rapid spread and global dissemination of these subtypes arises from their efficient transmission via certain transmission routes, namely, infected blood products and injecting drug use. Subtypes 1b and 2a are more strongly associated with the former route and the relative prevalence of these subtypes has decreased in recent years due to improved blood screening [21,24]. Subtypes 1a and 3a most often infect injecting drug users and appear to be increasing in prevalence [21,23].

The 'endemic' group of HCV strains are less prevalent than the epidemic subtypes and tend to have restricted geographic distributions. For example, the subtypes of type 6 are found only in South East Asia. The high genetic diversity of endemic strains points to a long period of infection in these areas, where transmission is thought to be maintained by a variety of relatively inefficient social and domestic routes, including sexual transmission [22]. As HCV was only identified in 1989, differences in the long-term transmission dynamics of the endemic and epidemic strains should be impossible to discover. However, using methods based on coalescent theory [9], the epidemic history of different HCV strains can be reconstructed from observed viral genetic diversity [24].

The third group are the 'local epidemic' strains of HCV that are found at high prevalence but only in specific locations and risk groups. The best example is subtype 4a which infects more than 10% of the Egyptian population but is rare outside the Middle East. Epidemiological studies suggest that this strain was widely transmitted in Egypt during the twentieth century by mass injectable drug treatment campaigns against schistosomiasis [25].

Taken together, these patterns suggests that the varied epidemiological behaviour of HCV types and subtypes is determined by transmission route (or the social and medical practices that generate such routes) rather than by genetic differences among strains [22,24]. If this is true, then any of the 50 or more known HCV subtypes could 'emerge' and generate a future epidemic, so long as effi-

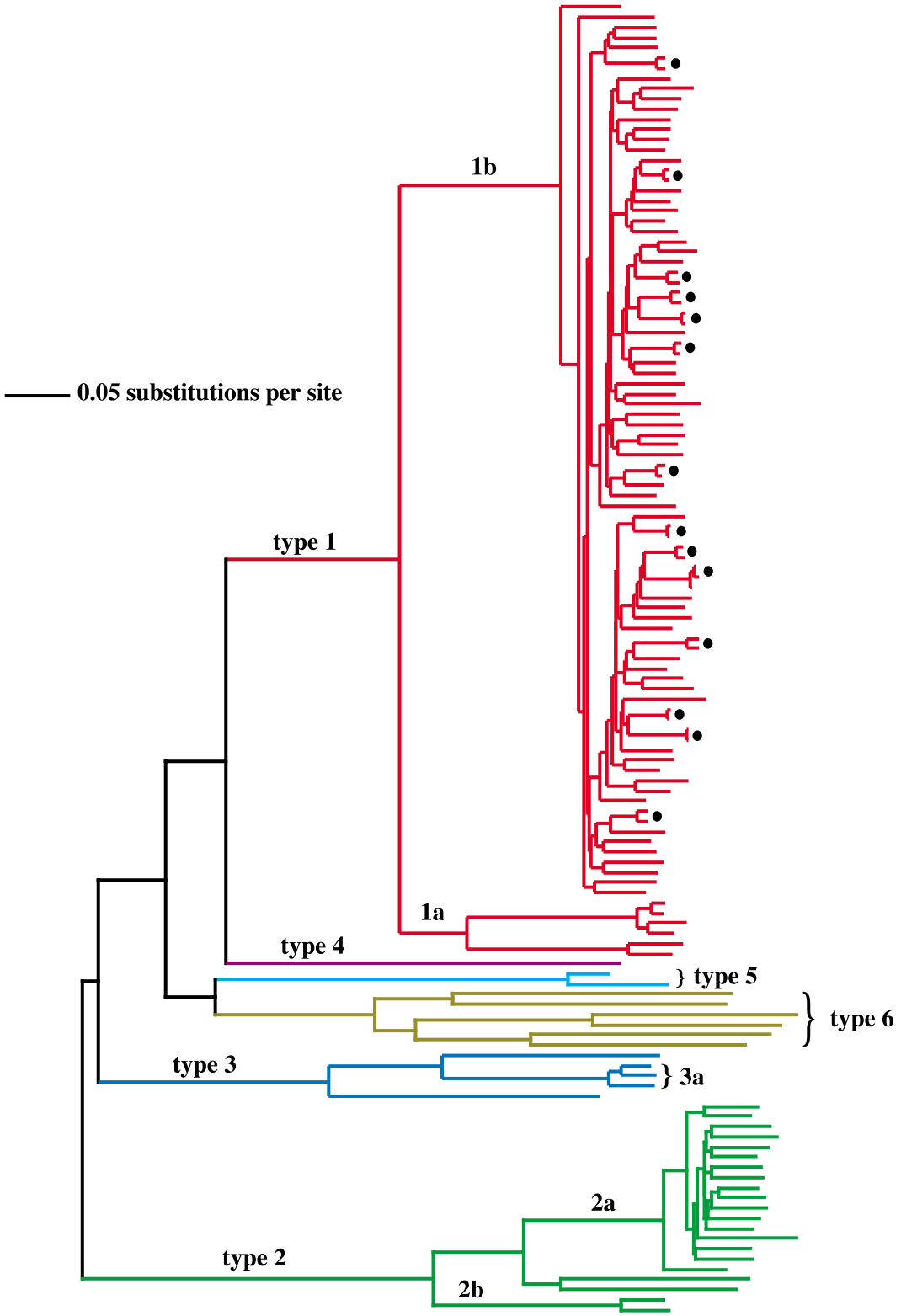

Fig. 2. A maximum likelihood phylogeny of all complete HCV genome sequences. The six main HCV types are shown in different colours and subtypes 1a, 1b, 2a, 2b and 3a are labelled. The black dots represent pairs of sequences which have been sampled from the same infected individual, illustrating the difference between within-patient and among-patient HCV diversity.

cient routes of transmission are available to the virus. Thus HCV genetic diversity poses a significant challenge to the prospective development of protective vaccines against HCV.

Although transmission route appears to be the dominant factor determining the global epidemiology of HCV, genetic variation among subtypes does generate significant differences in clinical outcome [21]. Repeated stud-

ies have shown that the response to anti-viral therapy is lower in patients with subtypes 1a, 1b and 4a than in those infected with type 2 and 3 strains [26]. Viral subtyping is therefore a useful tool in the management and treatment of chronic HCV infection. The causes of variation in treatment response are not well understood. Studies of Japanese patients infected with subtype 1b indicated that the outcome of interferon therapy was correlated with genetic variability in a portion of the NS5A gene (the interferon sensitivity determining region, ISDR) [27], although subsequent studies of European patients could not confirm this result [28].

In addition, several studies have indicated that HCV genetic diversity partially determines the variation among patients in disease progression and severity of liver disease, with the conclusion that subtype 1b infections are more likely to cause liver disease than other subtypes [21,29]. However an equal number of studies have found no such association [21,30]. Subtype 1b has also been linked with high rates of hepatocellular carcinoma than other subtypes (e.g. [31]). The statistical analysis of such studies is difficult, because subtype 1b is more often found in older patients, who might be expected to have more severe disease as a result of a longer period of infection.

3.2. Genetic diversity within infected individuals

The previous section touched upon differences in HCV disease progression among different hosts. Here we review recent experimental results concerning the within-host evolution of HCV, with subtype 1b being the most commonly investigated strain. The studies discussed are affected by a multitude of factors and often yield contradictory results. Both the number of patients investigated and number of virions extracted per patient are generally small. Moreover, the studies focus on many different viral genome regions, making it difficult to draw general conclusions. Observed levels of viral diversity depend on the tissue sampled, disease stage, drug treatment, and the strength of humoral and cellular immune responses. Each of these factors are addressed in turn below.

HCV predominantly replicates in hepatocytes but low levels of replication have also been reported in peripheral blood mononuclear cells [32,33]. Studies report differences in the genetic diversity of samples taken from different tissues [34,35], in particular the blood and the liver [33,34,36–39]. Genetic diversity in the blood can be lower, equal, or higher than in the liver [39]. Higher diversity in the liver could result from some aspect of the viral replication process in hepatocytes, or from the accumulation of impaired sequences that are not able to leave the cell in functional virions [40]. Higher diversity in the blood is difficult to explain, as levels of HCV replication outside the liver are low [39]; it may be due to the accumulation of deficient virions which cannot infect liver cells. Findings are highly dependent on the genomic region investigated, and genetic differences in the E1/E2 genes have been interpreted as regulating cell tropism [32]. Natural selection, whether due to immune pressure, differences in viral replication rates, or cell-entry ability, may be highly variable between different tissue compartments. This puts into question the notion that a single quasi-species population exists inside each infected host [40].

The correlation of viral diversity with disease outcome is more clear-cut. After an initial acute phase following HCV inoculation, serum levels of the virus are generally very low and the infection may be cleared spontaneously. However, in approximately 85% of cases infection becomes persistent and may cause a spectrum of diseases, ranging from minor liver damage to liver cirrhosis and hepatocellular carcinoma [2]. In a study by Farci et al. [41], patients were grouped into four classes according to their disease outcome and genetic diversity was measured as the number of amino acid substitutions among sequences. Diversity was measured at three or more time-points; at the time of the first PCR-positive sample, just before the first PCR-positive sample, and just after sero-conversion. Despite having the highest viremia, patients with fulminant hepatitis showed the least amino acid diversity. Viral genetic diversity in resolving patients was slightly higher than in the fulminant cases, but considerably less than in those who progressed to chronic disease. In all cases, changes in genetic diversity were significantly higher in HVR1 than in other regions of the E1 and E2 genes. Crucially, it was found that non-synonymous substitutions in HVR1 were more common in rapid and slow progressors, compared to the fulminant and resolving cases. Although there are instances where complexity in the HVR1 does not seem to correlate with disease severity [42], reports generally agree that diversity is higher in asymptomatic carriers compared to severe disease cases [41,43,44]. In the final section we outline a simple model that describes these observed relationships between genetic diversity and disease progression.

Because of the lack of a vaccine, great effort has been invested in developing drug treatments against HCV [45]. The high mutability of the virus has unfortunately been a major obstacle for the success of most drug treatments to date. HCV anti-viral therapy involves either α-interferon alone, or interferon in combination with ribavirin [46], but about 60% of treated patients fail to clear the virus [47]. Generally, drug response appears to be correlated with HCV diversity during the early stages of the treatment [47–51], although not necessarily with diversity before the onset of treatment [49,52]. Diversity in non-responders is higher than in responders, suggesting that escape mutants arise [53]. Again we see significant variation between different regions in the viral genome; at the onset of therapy diversity of the HVR1 in non-responders appears to be higher than in sustained responders [47]. However, low genetic diversity during the early stages of drug therapy is usually found to be predictive of therapy success. Viral

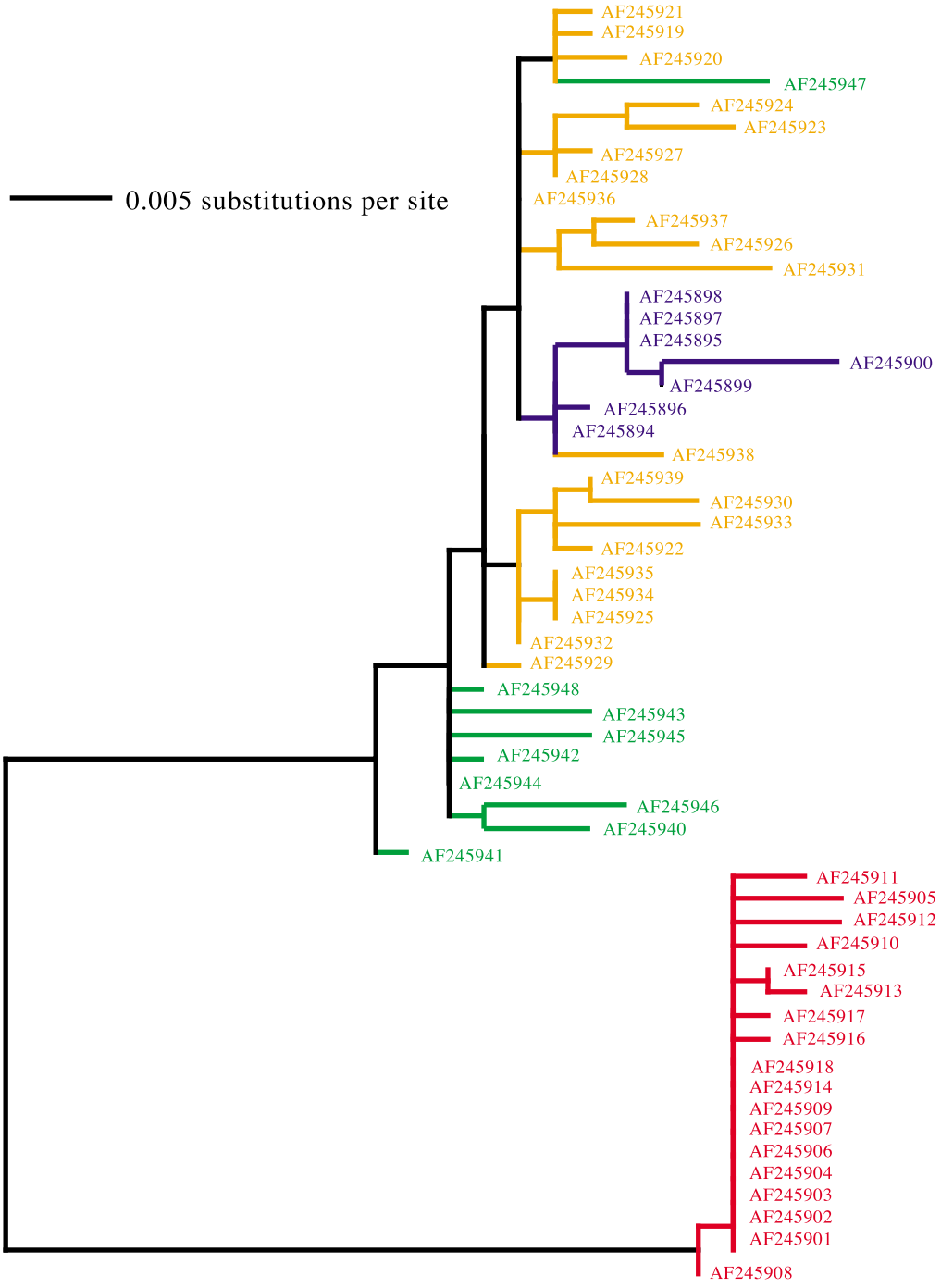

Fig. 3. A phylogeny of viral sequences obtained from an individual with rapidly progressing chronic infection. Viral isolates were extracted at four different times during the course of the infection. In chronological order, the samples are labelled as follows: time 1 (purple), time 2 (green), time 3 (orange), time 4 (red). Note how the viral population changes over time and how a selective sweep drives a new viral type to fixation between time points 3 and 4.

populations in non-responders and patients who relapse tend to be characterised by one resistant strain, which may already be present before therapy in some cases [47].

Genetic diversity is of course also affected by host factors, most importantly the strength of the humoral and cellular immune responses. Over the course of an infection, immune pressure could increase diversity by repeatedly selecting for escape mutants, and individuals that are immuno-suppressed following liver transplantation have almost homogenous viral populations [44,54,55]. Thus the success or failure of drug treatment will almost certainly be influenced by the patient's immune background.

Many of the above studies are limited by their reliance on partial genome sequences and simplistic measures of viral diversity. Advances in sequencing technology mean that whole genomes should become available in the near

future, providing us with a much more comprehensive picture of viral diversity. However, crude genetic diversity statistics throw away most of the information contained in sequence data and, at best, only inform us about viral mutation rate and population size. We should aim to collect viral isolates from multiple points in time, sequence large genomic regions, and analyse this data using powerful phylogenetic and population genetic methods. As an illustration, consider Fig. 3, which shows a phylogeny of sequences taken from a patient with rapidly progressing chronic HCV infection (data from [24]). Between the third and forth time points a new viral strain becomes fixed in the population, showing that a 'selective sweep' [10] has rapidly driven an advantageous variant to fixation. High-quality data, combined with a suitable quantitative framework for interpretation, will allow us to understand viral dynamics in much more detail than was possible previously.

4. A unified model of viral diversity in HCV

We have discussed three conceptual approaches that can be used to understand the genetic diversity of HCV. Recently, a simple qualitative model of genetic diversity in HIV has gained some prominence, and this model can also be straightforwardly applied to HCV [56,57]. The model contends that viral adaptation during an infection is controlled by two principal factors: selective pressure exerted by the immune system, and viral effective population size, which determines both the efficacy of selection and the rate at which advantageous viral variants are produced.

The interaction of these two factors is illustrated schematically in Fig. 4. The number of virus copies in a host will decrease as the immune response increases (green curve) thus reducing the ability of the virus to adapt. Conversely, stronger immune pressure leads to greater selection for escape mutants, which generate viral diversity (blue curve). Together these two factors give rise to a scenario that can be summarised as follows: for low levels of immune response there is little selective pressure, so very few adaptive mutations will be observed, despite a large viral population size. For high levels of immune response the viral effective population size will become small and therefore selection for escape mutants will be weak. Only at intermediate levels of population size and immune selection will there be a large number of escape mutants (red curve), resulting from the strong co-evolutionary 'arms race' between the virus and the immune system. It is this ability of the virus to adapt (i.e. to generate non-neutral amino acid changes) that therapeutic approaches need to decrease.

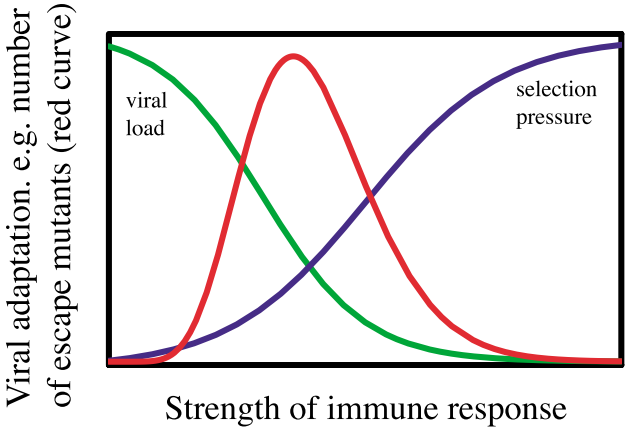

Fig. 4. A qualitative model for within-host viral evolution under the influence of immune (selection) pressure. The number of adaptive mutants (red curve) increases as immune pressure (blue curve) increases to moderate levels, despite a decrease in viral load (green curve). Only when immune pressure increases further will the viral load decrease sufficiently to impede adaptive change in the virus population.

This simple model seems to capture the essential dynamics underlying the genetic diversity of HCV and it makes sense to re-evaluate the findings of Farci et al. [41] in this context. In cases with fulminant hepatitis the immune system is too weak to control the virus, resulting in a large, homogeneous viral population. For the same reason, the model predicts that low diversity should be observed in immuno-compromised patients, as is the case [44,54,55]. In resolving cases, the immune system is sufficiently strong to reduce the selection of escape mutants by decreasing the viral population size, or is capable of dealing with those mutants that do arise, leading to intermediate levels of viral diversity. For intermediate strengths of the immune response, the virus population is only partially controlled and is therefore large enough to generate multiple escape mutants in response, resulting in chronic infection and a high observed dN/dS ratio. Thus Farci et al.'s [41] results can be described by a simple model that only incorporates selection and population size. In contrast, the quasi-species model incorporates theoretical assumptions that are not easily reconciled with realistic virus dynamics. Specifically, the quasi-species assumes that viral adaptation is not affected by population size or by mutations whose fitnesses are frequency-dependent. The above model has the same intuitive appeal as the quasi-species, but has the advantage of being firmly based on population genetic processes.

5. Conclusion

Viral diversity is influenced by a variety of factors that are not always separable. Qualitative descriptions, such as the quasi-species concept and the abovementioned population genetic model, can provide useful guides to the interpretation of genetic diversity in natural virus populations. Increasingly, however, we need to be able to assess genetic diversity more quantitatively. Population genetics and phylogenetics potentially provide us with powerful statistical frameworks to interpret observed sequence

data and to make testable predictions about the outcome of proposed drug and vaccine treatments. Progress over the next few years will show if we can correctly utilise these frameworks to investigate and quantify the dynamic processes involved in virus transmission and disease progression.

Acknowledgements

M.P.H.S. and O.G.P. are funded by Wellcome Trust research fellowships.

References

[1] Organization WH (1995) Wkly. Epidemiol. Rec. In: Book Wkly. Epidemiol. Rec.
[2] Prevention CfDCa (1998) Morb. Mortal. Wkly. Rep., 47, RR-19.
[3] Kalinina, O., Norder, H., Mukomolov, S. and Magnius, L.O. (2002) A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. J. Virol. 76, 4034–4043.
[4] Eigen, M. (1971) Self-organization of matter and the evolution of biological molecules. Naturwissenschaften 58, 465–523.
[5] Nowak, M.A. and May, R.M. (2000) Virus Dynamics. Oxford University Press, Oxford.
[6] Domingo, E., Escarmis, C., Menendez-Arias, L. and Holland, J.J. (1999) Viral quasispecies and fitness variations. In: Origin and Evolution of Viruses (Domingo, E., Webster, R. and Holland, J.J., Eds.), Academic Press, San Diego, CA, pp. 141–162.
[7] de la Torre, J.C. and Holland, J.J. (1990) RNA virus quasispecies populations can suppress vastly superior mutant progeny. J. Virol. 64, 6278–6281.
[8] Holmes, E.C. and Moya, A. (2002) Is the quasispecies concept relevant to RNA viruses? J. Virol. 76, 460–465.
[9] Rosenberg, N.A. and Nordborg, M. (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3, 380–390.
[10] Li, W.H. (1997) Molecular Evolution. Sinauer, Sunderland.
[11] Hartl, D.L. and Clark, A.G. (1998) Principles of Population Genetics, 3rd edn. Sinauer, Sunderland.
[12] Kimura, M. (1968) Evolutionary Rate at the molecular level. Nature 217, 624–626.
[13] Ohta, T. (1973) Slightly deleterious substitutions in evolution. Nature 246, 96–98.
[14] Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1999) Phylogenetic inference. In: Molecular Systematics (Hillis, D.M., Mable, B.K. and Moritz, C., Eds.), pp. 407–543. Sinauer, Sunderland.
[15] Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymus and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.
[16] Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.M. (2000) Codon-substitutions models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.
[17] Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C. and Salzman, N.P. (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. 16, 372–382.
[18] Robertson, B., Myers, G., Howard, C., Brettin, T., Bukh, J., Gaschen, B., Gojobori, T., Maertens, G., Mizokami, M., Nainan, O., Robertson, B., Netesov, S., Nishioka, K., Shin-i, T., Simmonds, P., Smith, D., Stuyver, L. and Weiner, A. (1998) Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy. Arch. Virol. 143, 2493–2503.
[19] Simmonds, P. (2000) Hepatitis C virus genotypes. In: Hepatitis C (Liang, T.J. and Hoofnagle, J.H., Eds.), pp. 53–83. Academic Press, London.
[20] Salemi, M. and Vandamme, A.M. (2002) Hepatitis C virus evolutionary patterns studied through analysis of full-genome sequences. J. Mol. Evol. 54, 62–70.
[21] Mondelli, M.U. and Silini, E. (1999) Clinical significance of hepatitis C virus genotypes. J. Hepatol. 31 (Suppl. 1), 65–70.
[22] Smith, D.B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P.L. and Simmonds, P. (1997) The origin of hepatitis C virus genotypes. J. Gen. Virol. 78, 321–328.
[23] Kalinina, O., Norder, H., Vetrov, T., Zhdanov, K., Barzunova, M., Plotnikova, V., Mukomolov, S. and Magnius, L.O. (2001) Shift in predominating subtype of HCV from 1b to 3a in St. Petersburg mediated by increase in injecting drug use. J. Med. Virol. 65, 517–524.
[24] Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. and Harvey, P.H. (2001) The epidemic behavior of the hepatitis C virus. Science 292, 2323–2325.
[25] Frank, C., Mohamed, M.K., Strickland, G.T., Lavanchy, D., Arthur, R.R., Magder, L.S., El Khoby, T., Abdel Wahab, Y., Aly Ohn, E.S., Anwar, W. and Sallam, I. (2000) The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. Lancet 355, 887–891.
[26] Bell, H., Hellum, K., Harthug, S., Maeland, A., Ritland, S., Myrvang, B., von der Lippe, B., Raknerud, N., Skaug, K., Gutigard, B.G., Skjaerven, R., Prescott, L.E. and Simmonds, P. (1997) Genotype, viral load and age as independent predictors of treatment outcome of interferon-alpha 2a treatment in patients with chronic hepatitis C. Construct group. Scand. J. Infect. Dis. 29, 17–22.
[27] Enomoto, N., Sakuma, I., Asahina, Y., Kurosaki, M., Murakami, T., Yamamoto, C., Izumi, N., Marumo, F. and Sato, C. (1995) Comparison of full-length sequences of interferon-sensitive and resistant hepatitis C virus 1b. Sensitivity to interferon is conferred by amino acid substitutions in the NS5A region. J. Clin. Invest. 96, 224–230.
[28] Zeuzem, S., Lee, J.H. and Roth, W.K. (1997) Mutations in the nonstructural 5A gene of European hepatitis C virus isolates and response to interferon alfa. Hepatology (Baltimore, MD) 25, 740–744.
[29] Boothe, J.C., Foster, G.R., Levine, T., Thomas, H.C. and Goldin, R.D. (1997) The relationship of histology in chronic HCV infection. Liver 17, 144–151.
[30] Kleter, B., Brouwer, J.T., Nevens, F., van Doorn, L.J., Elewaut, A., Versieck, J., Michielsen, P.P., Hautekeete, M.L., Chamuleau, R.A., Brenard, R., Bourgeois, N., Adler, M., Quint, W.G., Bronkhorst, C.M., Heijtink, R.A., Hop, W.J., Fevery, W.J. and Schalm, S.W. (1998) Hepatitis C virus genotypes: epidemiological and clinical associations. Benelux Study Group on Treatment of Chronic Hepatitis C. Liver 18, 32–38.
[31] Takada, A., Tsutsumi, M., Zhang, S.C., Okanoue, T., Matsushima, T., Fujiyama, S. and Komatsu, M. (1996) Relationship between hepatocellular carcinoma and subtypes of hepatitis C virus: a nationwide analysis. J. Gastroenterol. Hepatol. 11, 166–169.
[32] Laskus, T., Radkowski, M., Wang, L.F., Nowicki, M. and Rakela, J. (2000) Uneven distribution of hepatitis C virus quasispecies in tissues from subjects with end-stage liver disease: confounding effect of viral adsorption and mounting evidence for the presence of low-level extrahepatic replication. J. Virol. 74, 1014–1017.
[33] Jang, S.J., Wang, L.F., Radkowski, M., Rakela, J. and Laskus, T. (1999) Differences between hepatitis C virus 5′ untranslated region quasispecies in serum and liver. J. Gen. Virol. 80, 711–716.
[34] Navas, S., Martin, J., Quiroga, J.A., Castillo, I. and Carreno, V. (1998) Genetic diversity and tissue compartmentalization of the hepatitis C virus genome in blood mononuclear cells, liver, and serum from chronic hepatitis C patients. J. Virol. 72, 1640–1646.

[35] Radkowski, M., Wilkinson, J., Nowicki, M., Adair, D., Vargas, H., Ingui, C., Rakela, J. and Laskus, T. (2002) Search for hepatitis C virus negative-strand RNA sequences and analysis of viral sequences in the central nervous system: evidence of replication. J. Virol. 76, 600–608.

[36] Cabot, B., Gomez, J., Martel, M., Esteban, J.I., Otero, T., Esteban, R. and Guardia, J. (2000) Evolution of replicating and circulating hepatitis C virus quasispecies in non-progressive chronic hepatitis C. Hepatology 32, 414.

[37] Afonso, A.M.R., Jiang, J.J., Penin, F., Tareau, C., Samuel, D., Petit, M.A., Bismuth, H., Dussaix, E. and Feray, C. (1999) Nonrandom distribution of hepatitis C virus quasispecies in plasma and peripheral blood mononuclear cell subsets. J. Virol. 73, 9213–9221.

[38] Honda, M., Kaneko, S., Sakai, A., Unoura, M., Murakami, S. and Kobayashi, K. (1994) Degree of diversity of hepatitis C virus quasispecies and progression of liver disease. Hepatology (Baltimore, MD) 20, 1144–1151.

[39] Cabot, B., Martell, M., Esteban, J.I., Sauleda, S., Otero, T., Esteban, R., Guardia, J. and Gomez, J. (2000) Nucleotide and amino acid complexity of hepatitis C virus quasispecies in serum and liver. J. Virol. 74, 805–811.

[40] Stumpf, M.P.H. and Zitzmann, N. (2001) RNA replication kinetics, genetic polymorphism and selection in the case of the hepatitis C virus. Proc. R. Soc. Lond. Ser. B Biol. Sci. 268, 1993–1999.

[41] Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J.C., Strazzera, A., Chien, D.Y., Munoz, S.J., Balestrieri, A., Purcell, R.H. and Alter, H.J. (2000) The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. Science 288, 339–344.

[42] Lopez-Labrador, F.X., Ampurdanes, S., Gimenez-Barcons, M., Guilera, F.X., Costa, J., de Anta, M.T.J., Sanchez-Tapias, J.M., Rodes, J. and Saiz, J.C. (1999) Relationship of the genomic complexity of hepatitis C virus with liver disease severity and response to interferon in patients with chronic HCV genotype 1b interferon. Hepatology 29, 897–903.

[43] Curran, R., Jameson, C.L., Craggs, J.K., Grabowska, A.M., Thomson, B.J., Robins, A., Irving, W.L. and Ball, J.K. (2002) Evolutionary trends of the first hypervariable region of the hepatitis C virus E2 protein in individuals with differing liver severity. J. Gen. Virol. 83, 11–23.

[44] Doughty, A.L., Painter, D.M. and McCaughan, G.W. (2000) Post-transplant quasispecies pattern remains stable over time in patients with recurrent cholestatic hepatitis due to hepatitis C virus. J. Hepatol. 32, 126–134.

[45] Herrmann, E., Neumann, A.U., Schmidt, J.M. and Zeuzem, S. (2000) Hepatitis C virus kinetics. Antiviral Ther. 5, 85–90.

[46] Poynard, T., Marcellin, P., Lee, S.S., Niederau, C., Minuk, G.S., Ideo, G., Bain, V., Heathcote, J., Zeuzem, S., Trepo, C. and Albrecht, J. (1998) Randomised trial of interferon α2b plus ribavirin for 48 weeks or for 24 weeks versus interferon α2b plus placebo for 48 weeks for treatment of chronic infection with hepatitis C virus. International Hepatitis Interventional Therapy Group (IHIT). Lancet 352, 1426–1432.

[47] Farci, P., Strazzera, R., Alter, H.J., Farci, S., Degioannis, D., Coiana, A., Peddis, G., Usai, F., Serra, G., Chezza, L., Diaz, G., Balestrieri, A. and Purcell, R.H. (2002) Early changes in hepatitis C viral quasispecies during interferon therapy predict the therapeutic outcome. Proc. Natl. Acad. Sci. USA 99, 3081–3086.

[48] Farci, P., Strazzera, R., DeGjonnis, D., Peddis, G., Chessa, L., Setzu, R., Ghiani, A., Coiana, A., Wong, D., Balestrieri, A. and Purcell, R.H. (1998) Evolution of the viral quasispecies tracked by sequence analysis during interferon treatment of chronic hepatitis C. Hepatology 28, 938.

[49] Sandres, K., Dubois, M., Pasquier, C., Payen, J.L., Alric, L., Duffaut, M., Vinel, J.P., Pascal, J.P., Puel, J. and Izopet, J. (2000) Genetic heterogeneity of hypervariable region 1 of the hepatitis C virus (HCV) genome and sensitivity of HCV to alpha interferon therapy. J. Virol. 74, 661–668.

[50] Hino, K., Yamaguchi, Y., Fujiwara, D., Katoh, Y., Korenaga, M., Okazaki, M., Okuda, M. and Okita, K. (2000) Hepatitis C virus quasispecies and response to interferon therapy in patients with chronic hepatitis C: a prospective study. J. Viral Hepatitis 7, 36–42.

[51] Thelu, M.A., Brengel-Pesce, K., Leroy, V., Attuil, V., Drouet, E., Seigneurin, J.M. and Zarski, J.P. (2001) Influence of three successive antiviral treatments on viral heterogeneity in nonresponder chronic hepatitis C patients. J. Med. Virol. 65, 698–705.

[52] McKechnie, V.M., Mills, P.R. and McCruden, E.A.B. (2000) The NS5a gene of hepatitis C virus in patients treated with interferon-α. J. Med. Virol. 60, 367–378.

[53] Sanchez-Fueyo, A., Gimenez-Barcons, M., Puig-Basagoiti, F., Rimola, A., Sanchez-Tapias, J.M., Saiz, J.C. and Rodes, J. (2001) Influence of the dynamics of the hypervariable region 1 of hepatitis C virus (HCV) on the histological severity of HCV recurrence after liver transplantation. J. Med. Virol. 65, 266–275.

[54] Lawal, Z., Petrik, J., Wong, V.S., Alexander, G.J. and Allain, J.P. (1997) Hepatitis C virus genomic variability in untreated and immunosuppressed patients. Virology 228, 107–111.

[55] Ni, Y.H., Chang, M.H., Chen, P.J., Hsu, H.Y., Lu, T.W., Lin, K.H. and Lin, D.T. (1999) Decreased diversity of hepatitis C virus quasispecies during bone marrow transplantation. J. Med. Virol. 58, 132–138.

[56] McMichael, A.J. and Rowland-Jones, S.L. (2001) Cellular immune response to HIV. Nature 410, 980–987.

[57] Holmes, E.C. (2001) On the origin and evolution of the human immunodefiency virus (HIV). Biol. Rev. 76, 239–254.