Current Topics in Virology
Vol. 9, 2011

Mini-Review

Analysis of hepatitis C variability by cloning and sequencing: Technical recommendations

Rémy Moenne-Loccoz[1,2], Aurélie Velay[1,2], François Habersetzer[3], Michel Doffoël[3], Jean-Pierre Gut[1,2,4], Thomas F. Baumert[1,2,3,4], Françoise Stoll-Keller[1,2,4,*], and Evelyne Schvoerer[1,2,4]

[1]Unité Inserm U748, [2]Université de Strasbourg, [3]Pôle Hépato-digestif, Hôpitaux Universitaires de Strasbourg, F-67000 Strasbourg, [4]Laboratoire de Virologie, Hôpitaux Universitaires de Strasbourg, F-67091 Strasbourg, France

ABSTRACT

Hepatitis C variability is a field of numerous investigations. Information drawn from these studies is useful for a better comprehension of hepatitis C virus-related disease and for development of new preventive or therapeutic tools. With this aim, a critical analysis of viral quasispecies obtained by cloning and sequencing needs to follow some recommendations. From extraction procedure of high quality with careful conservation of RNA extracts, results should be obtained by reliable RT-PCR systems with random and/or degenerated primers in order to avoid any possible selection bias. The recommendations are exposed and discussed in detail to obtain reliable experimental data for valid biological findings.

KEYWORDS: reverse transcription, polymerase chain reaction, cloning, bias, technical recommendations, quasispecies, hepatitis C virus

I. INTRODUCTION

Hepatitis C virus (HCV) contains a single-strand positive RNA genome and displays a high genetic diversity. The mean frequency of nucleotidic mutations is 1.4-1.9 $\times$ 10^{-3} substitutions/nucleotide/year, due to defect in repair activity of the RNA-dependent RNA polymerase. Virus isolates were classified into at least 6 different genetic groups differing from each other by about 30% at the nucleotide level. The regions of HCV genome with crucial functions (translation, replication) and with structural constraints (non-coding genome extremities) are rather conserved. The most variable regions correspond to envelope glycoproteins E1 and E2, especially hypervariable regions (HVR)-1, -2 and -3. The N-terminal 27 amino acids fragment of the envelope glycoprotein E2, HVR-1, is the most divergent among HCV isolates and contains epitopes which are recognized by antibodies produced by patients. HVR-1 variability is essentially due to strong host-related humoral pressure [1]. At the same time, chemico-physical properties and conformation of HVR-1 are quite conserved [2], consistently with its important role in HCV entry. This example of opposite forces between conservation of structures/functions in HCV virions and variations of certain regions under immunological and/or treatment-related pressure lead to a chronic phase of HCV infection in 80% of the cases, and to frequent anti-HCV treatment failures, respectively. The analysis of these opposite constraints and, for our topic, of viral variability, is indeed an interesting approach to understand

*Corresponding author: Pr Françoise Stoll-Keller, Laboratoire de Virologie, Hôpitaux Universitaires de Strasbourg, F-67091 Strasbourg, France.
francoise.stoll@unistra.fr

better the physiopathology of HCV infection, frequent failure of antiviral treatment and how to develop vaccination programs.

This short review will focus on technical recommendations to take into account for critical analysis of viral quasispecies by cloning and sequencing in this context. From careful preparation of RNA extracts, strong data should be obtained by reliable Reverse Transcription-Polymerase Chain Reaction (RT-PCR) systems with random and/or degenerated primers in order to avoid any possible selection bias and/or artefacts. Bio-informatical analyses should be performed with critical precautions. The recommendations to be followed will be exposed in detail to obtain strong experimental data able to draw strong biological hypotheses.

II. Technical approach of viral variability

HCV variability was first explored by direct sequencing, allowing investigators to collect and analyze numerous viral strains, to define types and subtypes and to mount efficient pre-treatment orientation of successive therapy by pegylated interferon and ribavirin. This method produces a consensus nucleotide sequence, without detailed information on the composition of viral variants evolving in infected patients.

Elsewhere, study of HCV variability can be approached by the analysis of viral quasispecies, defined by HCV variants slightly different but genetically linked and present at a certain time point. These investigations started a few years ago and were facilitated by automated sequencing. They give a good reflect of different variants distribution in a quasispecies for one patient, in one biological compartment, at a precise date of the evolution of viral infection and can be followed according to time [3, 4]. Two main technical approaches have been used in this context.

The preliminary studies in this field often were based on analysis of single-strand conformation polymorphisms (SSCPs) after migration of PCR products, giving a general overview of viral quasispecies content without detailed description of nucleotide composition [5, 6].

Afterwards, to describe the composition of quasispecies precisely, the majority of teams performed their investigations by sequencing a sufficient number of clones, i.e. 10 or even more variants up to 40 clones ideally, giving a reliable picture of viral variants molecular features [7, 8]. However, this elegant exploration of HCV according to various clinical prognostic, to different biological compartments and to chronological evolution of viral infection is quite complex and should take into account some technical precautions which will be developed in the following part of this manuscript.

III. Technical recommendations for viral RT-PCR-cloning and sequencing with the aim to explore quasispecies distribution

In order to obtain a reliable viral quasispecies by RT-PCR-cloning and sequencing, whatever the scientific question is, one has to examine the different factors able to lead to amplification of a quasispecies which could be a bad reflect of the mixture of interest genes, such as HCV variants infecting a patient.

A few variations in PCR reactions could favour either a subgroup of HCV variants or another subgroup. The selection can occur in case of preferential denaturation, differential primers hybridization or variable polymerase elongation. Another potential hypothesis for bias is linked to a possible selection either due to the HCV genomic matrix itself which can be partially altered or to the micro-environment in RNA extract, potentially disturbing primers annealing. Complex molecular interactions giving artefacts such as molecular chimeras can also disturb reliability of the analyses. As an evidence, enzyme fidelity has to be verified before its extensive use in PCR process. At last, a bias due to cloning selection has also to be eliminated.

The key points of technical bias/artefacts risks will be discussed.

III.a. Quality of primers used in (RT)-PCR

A PCR selection or non random PCR bias can occur if certain viral variants in a quasispecies are advantaged and, as a consequence, are artificially overestimated.

The selection of viruses can occur in case of preferential denaturation due to GC content (either

in genomic matrix or in primers), differential efficacy of primers hybridization (which can be reduced by the use of degenerated primers) or, elsewhere, variable polymerase elongation linked to secondary structures in genomic target [9, 10]. More generally, factors determining the impact of mismatches in primers annealing to genomic targets include primers length, nature and location of mismatches, hybridization temperature, presence of solvents, primers concentration and quantity of cationic ions [10].

An approach tending to improve the reliability of (RT)-PCR amplification consists of the best choice of specific primers. It is generally admitted that partially degenerated primers are more confident in large spectrum amplification. They are supposed to represent a good alternative in the balance between too high matching of the primers with certain targets subsequently advantaged and too high degree of variability in degenerated primers mix decreasing affinity of the molecular tools for different variants of viral quasispecies. However, Polz and Cavanaugh observed a PCR selection with degenerated primers. Indeed, by choosing a degenerated primers pair frequently used in characterization of bacterial 16S rDNA (27F/ 1492R), they obtained a preferential amplification of genes containing G or C compared to A or T within hybridization sites targeted by the primers at the degenerated positions. One explanation of this bias could be the stronger binding energy between G and C (due to triple hydrogen bound) than between A and T [9, 11]. This point has to be taken into account when designing degenerated primers.

Bracho *et al.* studied genetic variability first in E1 and E2 (HVR-1 et -2) and second in NS5A (ISDR - interferon sensitivity determining region, and variable region V3) in eight patients infected by HCV genotype 1. Each genetic fragment was amplified twice with two primers sets slightly different but both degenerated. After cloning and sequencing, genetic variability was explored and gave significantly different genetic distributions for all patients except one. The authors attributed this differential amplification to differences in degeneracy level existing between both primers sets. Finally, the two main conclusions drawn by the authors were as following: the viral variability observed by the way of PCR procedures has to be considered as a minimum level, and the viral sequences obtained after (RT)-PCR depend on the quality of the primers design [10].

In the same context, Fan *et al.* investigated HVR-1 variability of HCV by using four RT-PCR procedures and concluded that mismatches between primers used in RT steps and targeted matrix could lead to sequences selection [12].

As a summary of this part, when analyzing variability of viruses such as HCV showing a potential high level of genetic plasticity, random hexameric primers could represent the best choice for RT step. In any case, a cautious design of gene specific primers has to be performed, by comparison with numerous corresponding targeted sequences in international database, within a conserved segment of the genome of interest. However, in case of polymorphism(s) in the conserved site, partially degenerated primers could represent a good compensatory solution. Additionally, a second run of PCR amplification by using a second primers set can be informative and reinforce data reliability if reproducible variants distribution is observed.

III.b. Possible artefacts by molecular interactions during PCR amplification of multiple genes

Bias and artefacts in the amplification of a mixture of genes which are close one to the others have to be avoided. Indeed, non specific molecular interactions tend to occur during the last PCR amplification cycles.

Suzuki and Giovannoni compared mixtures of two different bacterial 16S rDNA according to various initial ratios (i.e., 1/4, 2/3, 3/2 and 4/1) after amplification with primers pair 27F-338R. After 35 PCR cycles, ratios of final products were always next to 1/1. One hypothesis consists of re-hybridization of PCR products during annealing step where temperature is under DNA melting point and which can occur when concentration of PCR products is high enough. During amplification of a genes mixture, re-hybridization happens faster for the most abundant PCR product and prevents primers annealing. This leads to an amplification rate which declines faster for the most abundant than for the rare PCR product in the last PCR cycles [11, 13].

Thus, artificial overestimation of certain genes could be diminished by decreasing the number of PCR cycles. In the same context, Kurata *et al.* also hypothesized that re-hybridization could occur, but, rather by formation of homoduplexes during transition from denaturation to hybridization [14]. This could be avoided by thermocyclers with a fast ramp rate able to perform quick transition steps.

Moreover, Kanagawa listed two distinct mechanisms of PCR artefacts both occurring late during PCR run: either formation of hetero-duplexes giving after cloning possible artificially corrected sequences, or production of molecular chimeras favoured by partially elongated primers or template-switching [11]. Finally, these artefacts give a wrong reflect of the initial genes mixture and could be also minimized by limiting the number of PCR cycles.

III.c. Crucial role of enzyme fidelity used in RT-PCR process

Typically, mean error rate by a *Taq* polymerase during PCR is estimated to vary between 2×10^{-4} and $< 1.2 \times 10^{-5}$ mutations/nucleotide/cycle according to PCR conditions and targeted templates [15]. Oppositely, DNA polymerases with proofreading activity exhibits a mean error rate globally estimated to 10^{-6}. This part insists on the interests of DNA polymerases with proofreading activity which are unanimously admitted by authors exploring gene variability by the mean of PCR.

In an investigation analyzing quasispecies distribution, Mullan *et al.* observed different results after PCR amplification of identical cDNA derived from three distinct HCV-positive sera, either by a *Taq* polymerase or by a *Pwo* polymerase with proofreading activity. After cloning and sequencing, their data collected in HVR-1 of E2 and NS5A ISDR showed a strong advantage for the proofreading enzyme. Considering only non-synonymous mutations, 10-19 and 0-11 polymorphisms were observed for *Taq* and for *Pwo*, respectively [16, 17]. In the same context, Bracho *et al.* obtained similar results with 8-12 non-synonymous mutations in two 500 bp-length regions of Vesicular Stomatitis Virus amplified by a *Taq* polymerase compared to the low rate observed with proofreading thermostable *Pfu* enzyme, i.e. 0-3 non-synonymous mutations [17, 18]. Importantly, in these two studies, *Taq*-derived clones globally showed an increased proportion of unique polymorphisms which argued in favour of punctual *Taq*-introduced mutations rather than real quasispecies variability.

Malet *et al.* tested four different proofreading DNA polymerases (*eLONGase* (Invitrogen), *Expand High Fidelity* (Roche), *Pfu* (Promega), *Pwo* (Roche)) and one reference *Taq* polymerase (Applied Biosystems) which were evaluated by performing cloning and sequencing on PCR products from 5'-UTR (untranslated region) belonging to a known HCV plasmid. The authors recovered errors rates in the range expected by the manufacturers with proofreading DNA polymerases producing between 3 and 10 times less errors compared to the rate obtained with the *Taq* polymerase [17, 19].

One way to decrease errors rates due to *Taq* polymerases consists of the use of equimolar concentrations of dNTPs and $MgCl_2$ co-factor. However, this procedure requires a very long PCR extension time and fidelity is under that obtained with proofreading enzymes, i.e., 10^{-5} versus 10^{-6} errors/site/cycle, respectively [15, 17, 20].

Concerning reverse transcriptases, they contain an intrinsic error rate and their ability to polymerize cDNA can be affected by primer-matrix mismatches. Nevertheless, the fidelity of reverse transcriptases is significantly higher than that of *Taq* polymerases and primer-matrix mismatches affect significantly less cDNA polymerization than polymerization during *Taq*-mediated amplification [17].

By using a DNA polymerase with proofreading activity (i.e. *Pfu*), Domingo-Calap *et al.* recently investigated reliability of RT-PCR-cloning procedure in the characterization of HCV E1 and E2 glycoproteins variability. In order to assess artefacts linked to RT-PCR process, RNA was obtained by *in vitro* transcription of one unique HCV molecular clone and then was reverse transcripted with random hexamers. The authors diluted cDNA at 1/10 and 1/5000 and amplified both dilutions with *Pfu* and degenerated primers. If mutations were introduced during PCR amplification, sequences derived from cDNA 1/5000 should have shown an increased error rate

compared to cDNA 1/10 because approximately 9 additional PCR cycles were applied to the 1/5000 dilution. No statistically significant difference was detected between both groups of sequences, indicating that for their experiments, most of mutations could occur during *in vitro* transcription and/or RT step. Importantly, the same RT-PCR-cloning process (without cDNA dilution) was performed on RNA from 18 patients infected by HCV genotype 3a. Since average intra-patient mutation frequency was approximately 15-fold higher than average frequency of RT-PCR-induced errors, the authors concluded that RT-PCR and subsequent cloning is reliable for the investigation of viral variability in their experimental conditions [21]. However, they insisted in possible RT-PCR artefacts – potentially linked to RNA secondary structures –, which preferentially occurred in HVR-2 of HCV E2 in their study for example. These particular RNA structures could have a negative impact on reverse transcriptase fidelity.

As a consequence, some recommendations can be suggested: an initial RNA denaturing step should be done before any RT-PCR; moreover, a stable reverse transcriptase working at high temperatures between 42°C and 55°C should be preferred in order to allow a better RNA linearization during cDNA processing. Obviously, proofreading enzymes are a crucial need in good procedures devoted to exploration of genetic variability.

III.d. Bias risk during molecular cloning of RT-PCR products

After RT-PCR, the analysis of molecular clones obtained by individual insertion of PCR products in a plasmidic vector followed by transformation in a bacterial host (generally *E. Coli*), is a

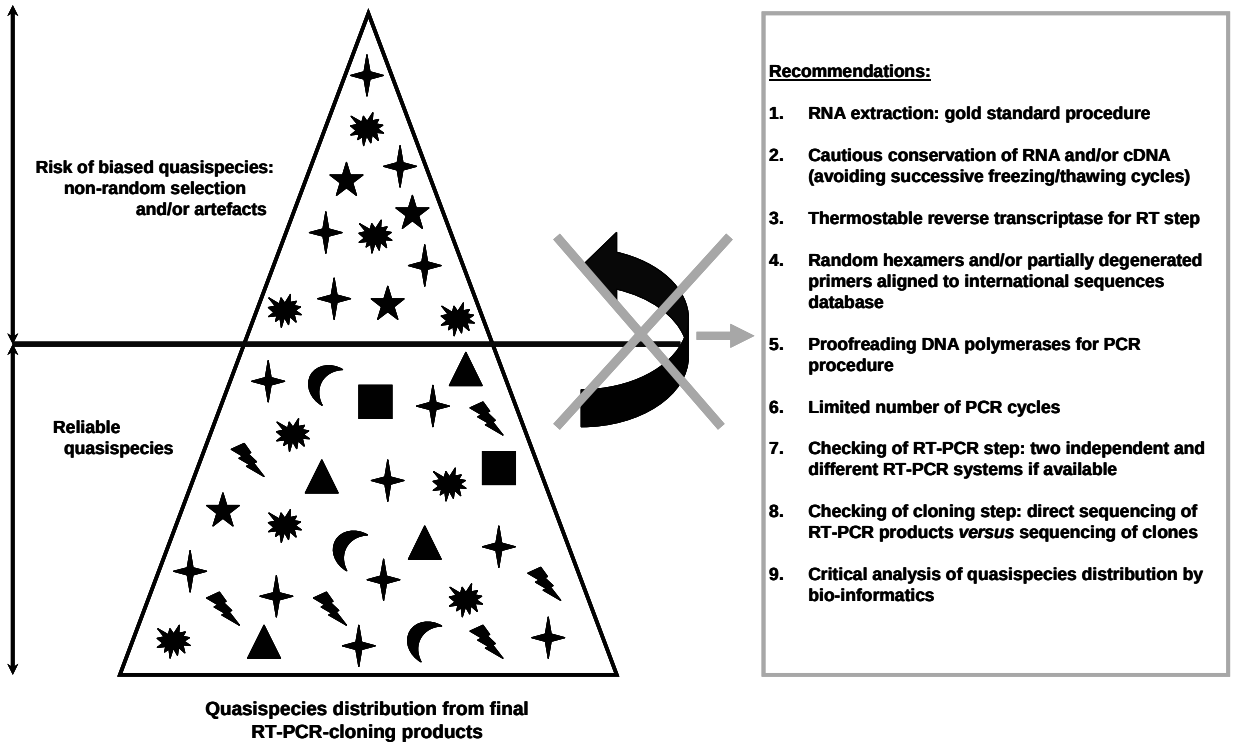

Figure 1. Schematic representation of recommendations to avoid bias during investigation of viral variability by RT-PCR-cloning procedure. Symbols within the big triangle correspond to quasispecies variants obtained by cloning after RT-PCR process. If selection bias occurs, certain variants which were advantaged during RT-PCR-cloning can be recovered, although they are not highly represented in the real quasispecies (black five-pointed star). Conversely, other variants can be artificially absent (e.g. black square). Recommendations aiming to avoid bias are listed in the right side of the figure.

standard procedure used in order to investigate genetic variability. Drake *et al.* reported that mutation rate during DNA replication in *E. Coli* was very low (5.4×10^{-10} mutations/base pair/ replication) [22].

However, Forns *et al.* observed a cloning bias when transforming *E. Coli* with plasmids containing structural HCV genes (nucleotides 1-2646 of the ORF) amplified by RT-PCR from plasma viral RNA of patient H (strain H77, genotype 1a). Indeed, all the clones contained stop codons or frame-shift mutations giving defective constructs for the translation, maybe due to a potential toxicity of HCV proteins for *E. Coli* [23]. One way to diminish or eliminate this kind of selection is to change or modify either vector or insert, such as suggested Forns *et al.* in their publication.

Moreover, systematic direct sequencing of RT-PCR products should be performed and compared with sequences subsequently obtained after cloning, in order to verify the similarity between the two steps and to avoid any selection bias. In this way, the sequences have to be concordant with an identity approaching 100%.

IV. CONCLUSIONS

As conclusive recommendations, given the risk of possible selection bias and/or artefacts with RT-PCRs, technical features are important in this context and a few precautions have to be taken into account (see conclusive Figure 1).

All the experiments have to be performed after RNA extraction following gold standard procedures and conservation of RNA extracts and/or cDNA has to be cautious by avoiding successive freezing/thawing cycles. A systematic initial denaturing step of RNA template should be done before RT-PCR and a stable reverse transcriptase working at high temperatures should be preferred for cDNA processing. Random hexameric primers could represent the best choice for RT step and, more generally, careful design of gene specific primers has to be performed, by comparison with numerous corresponding targeted sequences within a conserved genomic region. In case of polymorphism(s), partially degenerated primers should be chosen. In this kind of molecular approach, the use of polymerases with proofreading activity has to be respected for PCR in order to avoid the production of inaccurate and misleading data. Moreover, only a limited number of PCR cycles should be performed in order to prevent non-specific molecular interactions occurring late during amplification run.

Ideally, concordant results obtained by two RT-PCR assays with two different RT-PCR systems using different primers sets on two distinct nucleic acid extracts should be checked before further, in-depth characterization of HCV quasispecies. This will eliminate risks of artificial selection due to either microenvironment of nucleic acid extract, to genomic matrix features, or to primers-mediated bias.

Additionally, a comparison should be done between sequences obtained after direct sequencing of PCR products and variants produced by cloning, in order to exclude a non-random selection of viral clones.

At last, when performing final analyses, viral variability observed by way of PCR procedures has to be considered as a minimum level, cautiously considering highly homogeneous variants clusters which are distinct from the global quasispecies distribution. Critical phylogenetic analyses have to be performed before further functional investigations.

ACKNOWLEDGEMENTS

We are indebted to Fondation Transplantation (Saint-Apollinaire, France), Région Alsace (France), Abbott Molecular and Roche Pharma (French departments) for their support of Rémy Moenne-Loccoz (PhD student). Aurélie Velay received financial help from A.R.S. Lorraine (France) (one year fellowship for her Master 2 level).

REFERENCES

1. von Hahn, T., Yoon, J. C., Alter, H., Rice, C. M., Rehermann, B., Balfe, P., and McKeating, J. A. 2007, Gastroenterology, 132, 667-678.
2. Penin, F., Dubuisson, J., Rey, F. A., Moradpour, D., and Pawlotsky, J. M. 2004, Hepatology, 39, 5-19.

3. Pellerin, M., Lopez-Aguirre, Y., Penin, F., Dhumeaux, D., and Pawlotsky, J. M. 2004, J. Virol., 78, 4617-4627.
4. Roque-Afonso, A. M., Ducoulombier, D., Di Liberto, G., Kara, R., Gigou, M., Dussaix, E., Samuel, D., and Féray, C. 2005, J. Virol., 79, 6349-6357.
5. Asselah, T., Martinot, M., Cazals-Hatem, D., Boyer, N., Auperin, A., Le Breton, V., Erlinger, S., Degott, C., Valla, D., and Marcellin, P. 2002, J. Viral Hepat., 9, 29-35.
6. Neau, D., Jouvencel, A. C., Legrand, E., Trimoulet, P., Galperine, T., Chitty, I., Ventura, M., Le Bail, B., Morlat, P., Lacut, J. Y., Ragnaud, J. M., Dupon, M., Fleury, H., and Lafon, M. E. 2003, J. Med. Virol., 71, 41-48.
7. Torres-Puente, M., Bracho, M. A., Jiménez, N., García-Robles, I., Moya, A., and González-Candelas, F. 2003, J. Gen. Virol., 84, 2343-2350.
8. Gao, G., Stuver, S. O., Okayama, A., Tsubouchi, H., Mueller, N. E., and Tabor, E. 2005, J. Viral Hepat., 12, 46-50.
9. Polz, M. F. and Cavanaugh, C. M. 1998, Appl. Environ. Microbiol., 64, 3724-3730.
10. Bracho, M. A., García-Robles, I., Jiménez, N., Torres-Puente, M., Moya, A., and González-Candelas, F. 2004, Virol. J., 1, 13.
11. Kanagawa, T. 2003, J. Biosci. Bioeng., 96, 317-323.
12. Fan, X., Lyra, A. C., Tan, D., Xu, Y., and Di Bisceglie, A. M. 2001, Biochem. Biophys. Res. Commun., 284, 694-697.
13. Suzuki, M. T. and Giovannoni, S. J. 1996, Appl. Environ. Microbiol., 62, 625-630.
14. Kurata, S., Kanagawa, T., Magariyama, Y., Takatsu, K., Yamada, K., Yokomaku, T., and Kamagata, Y. 2004, Appl. Environ. Microbiol., 70, 7545-7549.
15. Eckert, K. A. and Kunkel, T. A. 1990, Nucleic Acids Res., 18, 3739-3744.
16. Mullan, B., Kenny-Walsh, E., Collins, J. K., Shanahan, F., and Fanning, L. J. 2001, Anal. Biochem., 289, 137-146.
17. Mullan, B., Sheehy, P., Shanahan, F., and Fanning, L. 2004, J. Viral Hepat., 11, 108-114.
18. Bracho, M. A., Moya, A., and Barrio, E. 1998, J. Gen. Virol., 79 (Pt 12), 2921-2928.
19. Malet, I., Belnard, M., Agut, H., and Cahour, A. 2003, J. Virol. Methods, 109, 161-170.
20. Kwiatowski, J., Skarecky, D., Hernandez, S., Pham, D., Quijas, F., and Ayala, F. J. 1991, Mol. Biol. Evol., 8, 884-887.
21. Domingo-Calap, P., Sentandreu, V., Bracho, M. A., González-Candelas, F., Moya, A., and Sanjuán, R. 2009, J. Virol. Methods, 161, 136-140.
22. Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. 1998, Genetics, 148, 1667-1686.
23. Forns, X., Bukh, J., Purcell, R. H., and Emerson, S. U. 1997, Proc. Natl. Acad. Sci. USA, 94, 13909-13914.