Journal of General Virology (2003), **84**, 000–000 **DOI** 10.1099/vir.0.19273-0

Sampling and repeatability in the evaluation of hepatitis C virus genetic variability

Manuela Torres-Puente, M. Alma Bracho, Nuria Jiménez, Inmaculada García-Robles, Andrés Moya and Fernando González-Candelas

Institut Cavanilles de Biodiversitat i Biologia Evolutiva and Departament de Genètica, Universitat de València, Apartado Oficial 2085, 46071 València, Spain

Correspondence
Fernando González-Candelas
fernando.gonzalez@uv.es

Received 4 April 2003
Accepted 9 May 2003

Among the experimental techniques available to study the genetic variability of RNA virus populations, the most informative involve reverse transcription (RT), amplification, cloning and sequencing. The effects of several aspects of these techniques on the estimation of genetic variability in a virus population were analysed. Hepatitis C virus populations from four patients were examined. For each patient, ten series of data derived from independent PCR amplifications of a single RT reaction were obtained. The sample size of each data set was 10 sequences (in nine series) and 100 sequences (in one series). An additional data set derived from an independent RT reaction (about 10 sequences) performed on RNA extracted from the same serum sample was also analysed. The availability of data sets of different sample sizes allowed the effect of sample size on the amount and nature of the genetic variability recovered to be examined. The repeatability of the data obtained in different amplification experiments as well as from different RT reactions was also determined, together with the best strategy to obtain a given number of sequences by comparing the set of 100 sequences obtained from a single amplification with those obtained by pooling the nine sets of 10 sequences. In all cases, these results confirm the high repeatability of the conclusions and parameters derived from the sets of 10 sequences. These results validate the use of relatively small sample sets for the evaluation of genetic variability and for the estimation of phylogenetic relationships of RNA viruses in population and epidemiological studies.

INTRODUCTION

RNA virus populations are extremely variable due to their large population sizes, short generation times and high replication and mutation rates (Domingo & Holland, 1997; Drake & Holland, 1999). These factors account for their fast evolutionary rates and adaptability to new selection pressures, which help them escape the immune system response of their hosts. Hence, RNA viruses constitute excellent model organisms for population and evolutionary genetics, both in the laboratory and in nature (Moya *et al.*, 2000).

To analyse and characterize such extremely variable populations, there are a number of techniques that allow the estimation of genetic variability, such as heteroduplex mobility assays (Woodward *et al.*, 1994), single-strand conformation polymorphism assays (Spinardi *et al.*, 1991), multiple-site-specific tracking assays (Resch *et al.*, 2001), mutant analysis by PCR and restriction enzyme cleavage (Chumakov *et al.*, 1991) or denaturing gradient gel electrophoresis (Fodde & Losekoot, 1994; Woodward *et al.*, 1994). However, to analyse many of the properties of such populations, it is necessary to know the nucleotide sequence of the constituting genomes. There are three main methodologies for this. The first method proceeds through reverse transcription (RT), amplification and direct sequencing of the resulting cDNA (Leitner *et al.*, 1993), hence rendering a single, consensus sequence on which variability is usually estimated by the analysis of variable positions in the electrophoregrams. The second method also amplifies cDNA by PCR but the resulting products are cloned into an appropriate vector. Clones derived from a single DNA molecule are sequenced, thus providing individual sequences representative of the initial, variable population. The third method, denoted PCR-based limited dilution assay, is also aimed at providing individual sequences but avoiding the cloning steps. This is achieved through limiting dilutions prior to PCR amplification (Rodrigo *et al.*, 1997; Taswell, 1981), thus assuring that only a single molecule acts as template for the reaction. Later, these PCR products are sequenced directly.

All of these methods have their advantages and limitations

The nucleotide sequence data from the variants reported in this paper will appear in the EMBL and GenBank databases under accession nos AJ560333–AJ560620.

and none is universally best for all applications and in all circumstances. Since hepatitis C virus (HCV)-infected individuals usually harbour 10^{10}–10^{12} virus particles (Neumann *et al.*, 1998), it could seem evident that the low numbers of sequences or clones obtained in these studies, usually in the tens at most, would hardly be a truly representative sample of the whole population and that increasing the number of sequences would provide a much better evaluation of the underlying diversity. The question can then be restated as: do the conclusions obtained with a relatively small number of sequences (about 10) still hold when compared with those obtained using a larger number of sequences (say 100) from the same serum sample?

A second interesting question is the repeatability of results obtained with an experimental protocol that involves one RT and several independent PCR amplifications. One process that might introduce a bias in the PCR products is PCR drift (Wagner *et al.*, 1994). This kind of bias could be due to stochastic variation in the early cycles of amplification and could result in poor repeatability in replicate PCR amplifications. Consequently, for many applications, such as molecular epidemiology or forensic studies, it is important to ascertain what levels of repeatability can be obtained with these sample sizes and techniques. Once again, our interest is not simply in the reproduction of the same raw sequences, since using such small sample sets makes it very unlikely to obtain exactly the same ones, but in the conclusions that can be derived from their analysis.

Lastly, and also as a consequence of previous considerations, we are interested in which of two alternative strategies is best for obtaining a large number of individual sequences, either cloning and sequencing a large number, namely 100, of DNA amplified products from a single PCR reaction or dividing the total number of sequences into several PCR reactions and cloning and sequencing a smaller number of products from each.

We have used a factorial design to analyse these questions. Since the initial level of genetic variability in each sample is a likely factor affecting diversity analyses, we decided to use four HCV-infected patients whose viruses covered a wide range of genetic variability. Our results indicate that essentially the same conclusions can be obtained from a moderately small sample set than from a large sample set, although, as expected, the larger the sample set the more detailed the description of the virus population will be.

METHODS

Cloning and sequencing of virus populations. Serum samples from four patients were chosen for this study. Previous estimates of HCV genetic diversity in these samples showed markedly different levels (Table 1). The individuals selected encompass the full range of HCV genetic variability found in a previous study (unpublished data). One RT reaction was performed for each patient. Aliquots of each resulting cDNA were amplified independently in 10 different PCR reactions to obtain a 742 nt fragment in the E1–E2 region of the viral genome. Products from each amplification were cloned and a number of recombinant plasmids were sequenced. In 9 of 10 cases, about 10 clones were sequenced. Around 100 clones were sequenced from the remaining case. Hence, we generated 10 different data sets from 10 different PCR amplifications, derived from a common RT reaction.

Viral RNA was extracted from 140 μl serum using the QIAamp Viral RNA kit (Qiagen). RT was performed on a 40 μl volume containing

Table 1. Summary of genetic variability in the E1–E2 region of HCV in the four patients analysed

For each patient, the tenth set corresponds to the transformation from which 100 clones were sequenced, the one denoted 9 × 10 corresponds to the analysis of the pooled results for the nine data sets of 10 sequences. Prev. corresponds to the set obtained in a previous experiment and the last row corresponds to the analysis of all the sequences pooled. *n*, Number of sequences; Nhap, number of different haplotypes; ▲, nucleotide diversity after Jukes–Cantor correction.

Patient 21				Patient 16				Patient 45				Patient 13			
Set	*n*	Nhap	▲	Set	*n*	Nhap	▲	Set	*n*	Nhap	▲	Set	*n*	Nhap	▲
2101	11	10	0·05039	1601	10	7	0·01778	4501	11	4	0·00180	1301	11	6	0·00224
2102	9	9	0·05308	1602	11	8	0·01499	4502	11	7	0·00269	1302	11	5	0·00216
2103	10	8	0·03650	1603	10	6	0·01693	4503	11	5	0·00396	1303	12	2	0·00075
2104	12	10	0·04721	1604	12	8	0·01408	4504	12	4	0·00206	1304	11	2	0·00045
2105	11	7	0·03148	1605	11	8	0·01864	4505	10	3	0·00099	1305	11	6	0·02690
2106	10	9	0·03854	1606	9	6	0·01444	4506	13	5	0·00152	1306	9	2	0·00165
2107	9	6	0·04807	1607	14	9	0·01827	4507	12	6	0·00247	1307	12	4	0·00123
2108	11	11	0·04251	1608	13	11	0·01738	4508	12	6	0·00247	1308	12	4	0·00123
2109	10	10	0·04872	1609	8	6	0·01867	4509	9	5	0·00316	1309	10	4	0·00148
2110	96	56	0·04222	1610	100	30	0·01241	4510	99	31	0·00227	1310	99	27	0·00178
9 × 10	93	57	0·04359	9 × 10	98	42	0·01665	9 × 10	101	36	0·00234	9 × 10	99	22	0·00153
Prev.	20	15	0·04481	Prev.	10	6	0·01040	Prev.	10	5	0·00197	Prev.	10	1	0
21	189	105	0·04292	16	198	62	0·01477	45	200	61	0·00231	13	198	45	0·00166

10 µl eluted RNA, 8 µl 5× RT buffer, 500 µM of each dNTP, 1 µM antisense primer (see below), 100 U MMLV reverse transcriptase (USB) and 20 U RNaseOUT (Gibco-BRL). The reaction was incubated at 42 °C for 45 min, followed by 3 min at 95 °C.

Amplifications were performed in a 100 µl volume containing 4 µl of the RT product, 10 µl 10× PCR buffer, 200 µM of each dNTP, 400 nM of each primer (sense, 5′-CGCCAYTGGACRACGCAA-3′, positions 1230–1247 in the reference sequence accession no. M62321; antisense, 5′-RCAMCCRAACCAATTGCC-3′, positions 1997–1980) and 2·5 U *Pfu* DNA polymerase (Stratagene). PCR was performed in a Perkin Elmer 2400 thermal cycler with the following thermal profile: 94 °C for 3 min, then 5 cycles at 94 °C for 30 s, 55 °C for 30 s and 72 °C for 3 min, followed by 35 cycles at 94 °C for 30 s, 52 °C for 30 s and 72 °C for 3 min. A final extension at 72 °C for 10 min was also carried out.

Amplification products were cloned directly into the *Eco*RV-digested pBluescript II SK (+) phagemid (Stratagene). Recombinant clones with our insert were selected by PCR-colony isolation and were purified by manual precipitation. Clones were sequenced using primers 5′-RGCCATCTTGGAYATGATYGC-3′ (sense, positions 1367–1387) and 5′-YTTGGRGGGTAGTGCCARCARTA-3′ (antisense, positions 1816–1794) and the ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems) in an ABI 3700 automated sequencer (Applied Biosystems). Sequences were verified and both strands assembled using the Staden package (Staden *et al.*, 1999). Sequences obtained in the previous RT reaction were obtained in a similar manner, except that the same primers were used for amplification and sequencing, thus rendering 406 nt long sequences from the same genome region.

Statistical analysis. For all analyses, 13 data sets for each patient were used. Of these, 10 corresponded to the 10 independent amplifications, nine with a sample size of around 10 sequences and one with a sample size of about 100 sequences. A further set was obtained by combining the nine sets of 10 sequences into a single set. Another set was composed of all of the sequences from the 10 independent amplifications. The last set corresponded to a previous study (unpublished data), in which we obtained a similar number of sequences ($n=10$) by the same procedure and from the same region for each sample, although from a different RT reaction. Hence, a test on the effect of the RT reaction was possible by comparing results from two different RT reactions. Also, we obtained information on the repeatability of the results, by comparing the nine samples of 10 sequences among themselves, on the effect of sample size, by comparing each of these with the samples of 100 sequences, and on the effect of sampling a similar number of sequences from a single amplification (of 100 sequences) or from different amplifications (nine amplifications of 10 sequences).

Genetic variation for each data set was evaluated using DNAsp, version 3.51 (Rozas & Rozas, 1999). Pairwise comparisons between data sets from the same patient were obtained with Arlequin, version 2000 (Schneider *et al.*, 2000), as estimates of the population subdivision statistic Fst. The statistical significance for this statistic was evaluated by 1000 random permutations in each case. Phylogenetic trees were constructed using the neighbour-joining algorithm (Saitou & Nei, 1987) based on the general time reversible evolutionary model for nucleotide substitution (Posada & Crandall, 2001). These analyses were done with PAUP*, version 4.0b10 (Swofford, 1998). Estimates of synonymous and non-synonymous substitutions among sequences from each data set were obtained using the Nei–Gojobori method (Nei & Gojobori, 1986), as implemented in the program MEGA (Kumar *et al.*, 2000). Exact, unbiased estimates of *P* values in contingency tables were obtained using the Metropolis algorithm implemented in the program RxC (Miller, 1997).

RESULTS

Previous estimates of genetic variability by means of nucleotide diversity were largely consistent with those obtained in this experiment for the four patients (Table 1). The only noticeable difference corresponded to patient 13, the one with the lowest initial nucleotide diversity. In the previous analysis, the 10 sequences obtained from this patient were identical, whereas several variants were obtained in this experiment. Nevertheless, this patient still presented the lowest overall genetic variability in the study.

For the four patients, nucleotide diversity was very similar for the nine data sets of 10 sequences, as well as for the one with 100 sequences and the one obtained by pooling the previous sets (9×10). Only for patient 16 was there a certain difference both in the number of different haplotypes and in nucleotide diversity, which were larger for the pooled set than for the large set (1×100). In all cases, estimates of nucleotide diversity for the large and pooled sets were intermediate among those obtained for the nine small data sets. A similar result was obtained when the estimates from the previous experiment were compared with this one, with the exception of patient 13, as already noted. It is also noticeable that although different haplotypes were sequenced, largely similar values of genetic variability were obtained. For instance, of the 113 different haplotypes sequenced from patient 21 when the large and the pooled sets were considered (56+57), only eight were coincident between both groups, the remaining 105 were different. The same pattern was obtained in the other patients.

A summary of the results from genetic differentiation analyses is shown in Table 2. Pairwise genetic differentiation analyses of the large sample set (1×100) with respect to each small size data set and the pooled set (9×10) from each patient were obtained. After correction for multiple, non-independent comparisons using Bonferroni's method (Miller, 1966), there were only two statistically significant Fst values and both corresponded to patient 16. One of them was from one of the small samples (series 1603) and the other corresponded to the pooled sample (9×10). This result was largely due to differences arising in three different data sets (series 1603, 1605 and 1607). In these cases, the intergroup component of variation was close to or even larger than 10 %, whereas in none of the other sets for this patient was it larger than 5 %.

When the small data sets from each patient were compared to each other in a pairwise manner using the same procedure (data not shown), there was only one statistically significant case, appearing in patient 45 for series 4504 and 4507. A few other cases (one for patient 21, two for patient 45 and one for patient 13) presented marginal significance, but

Table 2. Genetic differentiation statistics (Fst) between the large sample set (1 × 100) and each small sample set and their pooled data (9 × 10) for the four patients analysed

Prev. corresponds to the set obtained in a previous experiment. Significant differences are indicated by one ($P<0{\cdot}05$), two ($P<0{\cdot}01$) or three ($P<0{\cdot}001$) asterisks. Bold type indicates Fst value with $P<0{\cdot}05$ after Bonferroni's correction.

Patient 21	Fst	Patient 16	Fst	Patient 45	Fst	Patient 13	Fst
2101 versus 2110	0·04289	1601 versus 1610	0·04846	4501 versus 4510	−0·01574	1301 versus 1310	0·00414
2102 versus 2110	−0·06950	1602 versus 1610	−0·00942	4502 versus 4510	0·01039	1302 versus 1310	0·02625
2103 versus 2110	−0·03008	1603 versus 1610	**0·14625****	4503 versus 4510	0·02857	1303 versus 1310	−0·01194
2104 versus 2110	0·01687	1604 versus 1610	−0·02759	4504 versus 4510	−0·00171	1304 versus 1310	−0·04160
2105 versus 2110	−0·00212	1605 versus 1610	0·10176*	4505 versus 4510	−0·02213	1305 versus 1310	0·00397
2106 versus 2110	−0·00594	1606 versus 1610	0·00544	4506 versus 4510	−0·00809	1306 versus 1310	−0·00117
2107 versus 2110	−0·02131	1607 versus 1610	0·09379**	4507 versus 4510	−0·00905	1307 versus 1310	−0·00786
2108 versus 2110	−0·03584	1608 versus 1610	−0·00424	4508 versus 4510	0·00576	1308 versus 1310	−0·00803
2109 versus 2110	0·02039	1609 versus 1610	0·03659	4509 versus 4510	0·04680	1309 versus 1310	−0·01454
9 × 10 versus 2110	0·00960	9 × 10 versus 1610	**0·03540*****	9 × 10 versus 4510	−0·00030	9 × 10 versus 1310	−0·00114
Prev. versus 2110	−0·01901	Prev. versus 1610	0·00666	Prev. versus 4510	−0·00312	Prev. versus 1310	−0·04812

none of them was significant after application of Bonferroni's correction.

Our final test for the homogeneity of sequences obtained from different transformation experiments came from their phylogenetic analysis. If there were substantial differences among the data sets obtained from different amplification experiments, then we would expect to obtain highly structured phylogenetic trees, with most sequences derived from each set grouped into separate clusters. The phylogenetic trees obtained for the different haplotypes from the almost 200 sequences from each patient are shown in Fig. 1 and the frequency distribution of sequences from each set into haplotypes is shown in Table 3. As expected, each phylogenetic tree reflected the genetic variability levels described previously, with a higher degree of branching in the tree for patient 21, the one with the largest variability. Nevertheless, in all cases, it was evident that sequences derived from any data set did not group into separate clusters and, instead, they mixed in quite a random manner. This was also true for the sequences obtained in the previous experiment, derived from an independent RT reaction followed by PCR amplification, as in the four phylogenetic trees they grouped similarly to the other small sample size data sets from the corresponding patient.

A further test of the random assignment of sequences from each data set to different clusters in each phylogenetic tree also allowed us to check whether resampling of single, different variants had been produced in each amplification as a result of the preferential amplification of a sequence in the early stages of the PCR that would show in its overrepresentation in the subsequent cloning and sequencing. This would lead to incorrect estimates of genetic diversity, both at the within and among series levels, and to lack of repeatability of the whole procedure. Table 3 summarizes the frequency distribution of the different sequence variants obtained from each patient. Contingency tests, with unique sequences grouped into a single class, resulted in non-significant deviations from the null hypothesis of homogeneity among all data sets from each patient. Only for patient 16 was there an apparent deviation between observed and expected counts (resulting in a *P* value of 0·067 for the test statistic) for the most frequent variant – in apparent excess in the 1 × 100 data set – and the unique variants – in apparent excess in the 9 × 10 set.

Rates of synonymous and non-synonymous substitutions were not significantly different among data sets from each patient (data not shown, available upon request), although there were significant differences among patients. These differences correlate with the levels of variability described previously, especially for non-synonymous substitutions, ranging from an average of 0·0005 substitutions per site ($s\ s^{-1}$) for patient 13 to 0·0487 $s\ s^{-1}$ for patient 21 (0·0015 for patient 45 and 0·0185 for patient 16). Interestingly, synonymous substitutions were less variable among patients, with average values of 0·0043 $s\ s^{-1}$ for patient 13, 0·0051 $s\ s^{-1}$ for patient 45, 0·0053 for patient 16 and 0·0295 for patient 21.

DISCUSSION

The extent and nature of genetic variation in a virus population are among the most important factors that determine the short- and long-term evolution of a virus and its interaction with the host. Despite controversies about the units of selection and evolution in RNA virus populations and the main factors driving their evolution (Domingo, 2002; Holmes & Moya, 2002), the presence of escape mutants, highly virulent and/or faster replicating variants will undoubtedly have an effect on their fate. One of the most difficult tasks faced by virologists is the documentation and evaluation of genetic variability in these populations. Despite technological advances allowing a faster, global evaluation of diversity and the introduction of new methods that allow the search for specific variants, sequencing recombinant plasmids remains the best method

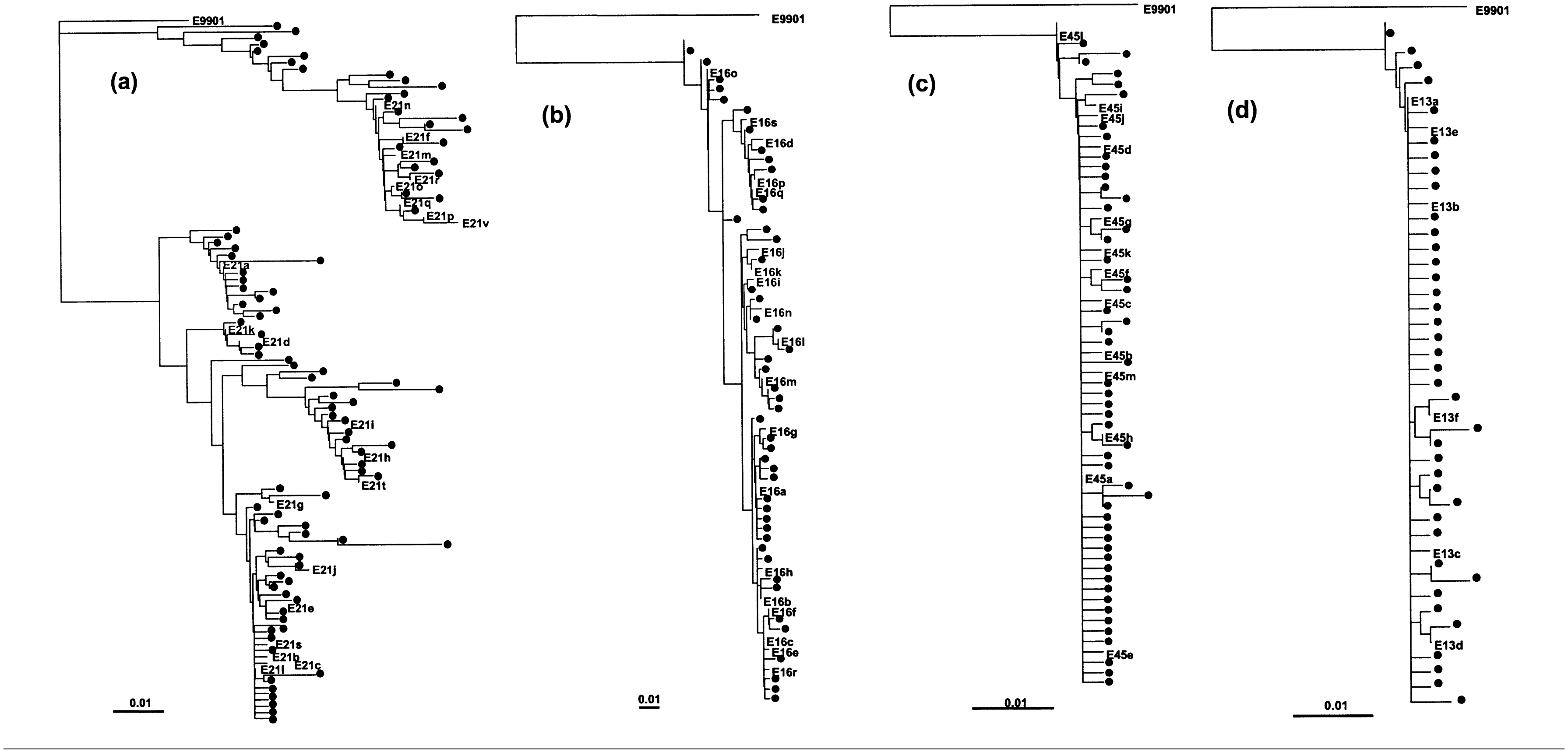

Fig. 1. Phylogenetic tree obtained by the neighbour-joining method for the different haplotypes sequenced from patients 21 (a), 16 (b), 45 (c) and 13 (d). Frequencies of each haplotype are reported in Table 3, except for those represented by only one sequence (Table 3, column u), shown here as black circles. E9901 corresponds to a common outgroup sequence. Bar, 0·01 substitutions per nucleotide position.

Table 3. Frequency distribution of sequences in the different haplotypes (a–u) obtained from each patient in the different data sets

Prev. corresponds to the set obtained in a previous experiment. Column u corresponds to haplotypes represented by only one sequence.

Series	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	v	u
Patient 21																						
2101	1											1	2			1						6
2102					1			1						1								6
2103	2		1				2		1			1										3
2104	2				2									1	1		1					5
2105	4				2									1								4
2106	2				1			1														6
2107	3							1				2										3
2108	1				1						1	1			1	1						5
2109	1								1					1	1							6
2110	10	2	1	2	7	2		4		2	1	11		4			1	5	2	1		41
Prev.	3		1					1				3			1		1			1	2	7
Patient 16																						
1601	4									1		1			1	1			1			1
1602	2		2							1					2							4
1603	2		1												3		2					2
1604	4		1			2			1							1	1					2
1605	2										2				1	1			2			3
1606	4		1														1					3
1607	3							1		2					1	3			1			3
1608	3									1	1											8
1609	2	1											2			1						2
1610	36	4	13	2	2	4	4	1	2			3		3	1	4	5	2				15
Prev.	5														1							4
Patient 45																						
4501	8											1										2
4502	5									1			1									4
4503	7																					4
4504	9																					3
4505	8																					2
4506	9	1																				3
4507	7		1					1														3
4508	7							1														4
4509	4			1			2															2
4510	60	1	2	1	2	2			2	1	2	5										21
Prev.	6												1									3
Patient 13																						
1301	6					1																4
1302	7				1																	3
1303	10					2																
1304	10				1																	
1305	6			1	1																	3
1306	8																					1
1307	9		1																			2
1308	9		1																			2
1309	7					1																2
1310	67	2		2	4	2																22
Prev.	10																					

to ascertain the linkage relationships among variants in different genome positions. But the potential uses of ascertaining genetic variability in virus populations extend into many realms, including evolutionary and epidemiological reconstructions. In these contexts, it is necessary to document not only the nature of the variants but also their frequencies and sequencing is burdened with economical and experimental constraints. Consequently, most studies using this methodology usually analyse only a few variants among those initially present in the population. In this paper, we have explored the validity of the inferences drawn from such necessarily reduced sample sizes.

Our genetic differentiation and phylogenetic analyses reflect the existence of a substantial homogeneity among the data derived from the different series obtained in each of the four patients included in our study. No significant differences were observed when the genetic variability parameters obtained from small sample data sets were compared to those obtained from the corresponding large ones. In all cases, we found that large data set values were intermediate among those obtained from the small sets, thus indicating a smaller accuracy for the estimates derived from the latter. Although, as expected, the larger the size of a data set the more precise the derived estimates are, according to our results, the variation found among small sample sets and between these and the large ones was not significant. Hence, our study shows that it is adequate to use relatively small sample sizes to evaluate genetic variability in virus populations by means of RT, PCR amplification, cloning and sequencing of recombinant plasmids.

Comparisons among the previous conclusions from the four patients with different levels of variability included in our study show that this variability has no influence on what we have just considered. There is neither a patient effect for the different sample sizes nor an interaction with respect to the level of variability of the samples analysed, at least for the range of variability we have worked with.

Since we have found a considerable similarity between the data derived from the large data set obtained from a single amplification and those obtained with the pooled series from different amplifications, our results also indicate that both strategies employed to obtain large samples are equally valid. Therefore, the choice between methods can be based upon other considerations.

The comparison to an additional data set of the same patients from a different, previous experiment allowed us to test the role that the RT reaction could play as a biasing factor with regards to the reproducibility of the data. We found consistency between the conclusions extracted from data sets from two different RT reactions, as the previous set was undistinguishable from the small size sets derived from the same RT reaction in this experiment. The only difference was observed for patient 13, in which the previous experiment sample showed no variation, whereas all samples derived in the new experiment harbour at least two variants. However, there is no statistical significance in the differences, again due to the small sample sizes used. This result is relevant for those cases in which an independent validation of the results obtained in a laboratory has to be performed in a different one. In these cases, it is not the absolute identity of the sequences obtained from both settings what should be expected. Rather, it is the concordance in the genetic variability parameters and phylogenetic relationships that should be compared. We must emphasize that these conclusions hold only for general evaluations of variability. In any case, the search for specific variants in the virus population in different experiments should provide identical results.

Furthermore, our experimental design allowed us to address another important issue in the estimation of genetic variability in virus populations, i.e. the error introduced by random preferential amplification of some variants by DNA polymerases used in PCR. Even a relatively low preferential amplification during the first rounds would lead to increased frequency estimates for some variants as a result of the exponential growth in subsequent replication rounds. Our data do not provide support for this, since there is homogeneity in the distribution of variants among different data sets for the same patient, even including a set from a separate RT reaction. Consequently, for the estimation of genetic variability of HCV in these patients, it would be legitimate to pool the data from all the sets, thus obtaining a more accurate estimate of the true value in each case.

In summary, our main conclusion is that although the raw data in the different sets were all distinct, we have found a great consistency between the conclusions derived from them, not only in genetic variability but also in phylogenetic relationship estimates. This consistency is maintained regardless of sample size or the amplification and cloning strategy used.

ACKNOWLEDGEMENTS

This work was supported by the Conselleria de Sanidad, Generalitat Valenciana. Further support was provided by the Spanish Ministerio de Educación y Ciencia, Plan Nacional I+D (project 1FD97-2328) and the Ministerio de Ciencia y Tecnología (BMC2001-3096). We also thank the comments of two anonymous reviewers.

REFERENCES

Chumakov, K. M., Powers, L. B., Noonan, K. E., Roninson, I. B. & Levenbook, I. S. (1991). Correlation between amount of virus with altered nucleotide sequence and the monkey test for acceptability of oral poliovirus vaccine. *Proc Natl Acad Sci U S A* **88**, 199–203.

Domingo, E. (2002). Quasispecies theory in virology. *J Virol* **76**, 463–465.

Domingo, E. & Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annu Rev Microbiol* **51**, 151–178.

Drake, J. W. & Holland, J. J. (1999). Mutation rates among RNA viruses. *Proc Natl Acad Sci U S A* **96**, 13910–13913.

Fodde, R. & Losekoot, M. (1994). Mutation detection by denaturing gradient gel electrophoresis (DGGE). *Hum Mutat* **3**, 83–94.

Holmes, E. C. & Moya, A. (2002). Is the quasispecies concept relevant to RNA viruses? *J Virol* **76**, 460–465.

Leitner, T., Halapi, E., Scarlatti, G., Rossi, P., Albert, J., Fenyo, E. M., Uhlén. &, M. (1993). Analysis of heterogeneous viral populations by direct DNA sequencing. *Biotechniques* **15**, 120–127.

Miller, R. G. (1966). *Simultaneous Statistical Inference*. New York: McGraw-Hill.

Miller, M. P. (1997). RxC, a program for the analysis of contingency tables. Northern Arizona University, Flagstaff, USA.

Moya, A., Elena, S. F., Bracho, A., Miralles, R. & Barrio, E. (2000). The evolution of RNA viruses: a population genetics view. *Proc Natl Acad Sci U S A* **97**, 6967–6973.

Neumann, A. U., Lam, N. P., Dahari, H., Gretch, D. R., Wiley, T. E., Layden, T. J. & Perelson, A. S. (1998). Hepatitis C viral dynamics *in vivo* and the antiviral efficacy of interferon-α therapy. *Science* **282**, 103–107.

Posada, D. & Crandall, K. A. (2001). Selecting the best-fit model of nucleotide substitution. *Syst Biol* **50**, 580–601.

Resch, W., Parkin, N., Stuelke, E. L., Watkins, T. & Swanstrom, R. (2001). A multiple-site-specific heteroduplex tracking assay as a tool for the study of viral population dynamics. *Proc Natl Acad Sci U S A* **98**, 176–181.

Rodrigo, A. G., Goracke, P. C., Rowhanian, K. & Mullins, J. I. (1997). Quantitation of target molecules from polymerase chain reaction-based limiting dilution assays. *AIDS Res Hum Retroviruses* **13**, 737–742.

Rozas, J. & Rozas, R. (1999). DNAsp, version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

Schneider, S., Roessli, D. & Excoffier, L. (2000). Arlequin, version 2000: a software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Spinardi, L., Mazars, R. & Theillet, C. (1991). Protocols for an improved detection of point mutations by SSCP. *Nucleic Acids Res* **19**, 4009.

Staden, R., Beal, K. & Bonfield, J. (1999). The Staden package, 1998. In *Computer Methods in Molecular Biology*, pp. 115–130. Edited by S. Misener & S. Krawetz. Totowa: Humana Press.

Swofford, D. L. (1998). PAUP*. Phylogenetic Inference Using Parsimony (* and other methods), version 4.0b10. Sunderland, MA: Sinauer Associates.

Taswell, C. (1981). Limiting dilution assays for the determination of inmunocompetent cell frequencies. *J Immunol* **126**, 1614–1619.

Wagner, A., Blackstone, N., Cartwright, P. & 7 other authors (1994). Surveys of gene families using polymerase chain reactions: PCR sequences and PCR drift. *Syst Biol* **43**, 250–261.

Woodward, T. M., Carlson, J., McClelland, C. & DeMartini, J. C. (1994). Analysis of lentiviral genomic variation by denaturing gradient gel electrophoresis. *Biotechniques* **17**, 366–371.

JGV Journal of General Virology

Offprint Order Form

PAPER vir19273st **Please quote this number in any correspondence**

Authors M. Torres-Puente and others **Date** ____________________

I would like 25 free offprints, plus ☐ additional offprints, giving a **total of** ☐ **offprints**

Dispatch address for offprints (BLOCK CAPITALS please)

____

____

____

____

Please complete this form **even if you do not want extra offprints.** Do not delay returning your proofs by waiting for a purchase order for your offprints: the offprint order form can be sent separately.

Please pay by credit card or cheque with your order if possible. Alternatively, we can invoice you. All remittances should be made payable to **'Society for General Microbiology'** and crossed **'A/C Payee only'**.

Tick one

☐ Charge my credit card account (give card details below)
☐ I enclose a cheque/draft payable to Society for General Microbiology
☐ Purchase order enclosed

Return this form to: JGV Editorial Office, Marlborough House, Basingstoke Road, Spencers Wood, Reading RG7 1AG, UK.

CHARGES FOR ADDITIONAL OFFPRINTS

Copies	25	50	75	100	125	150	175	200	Per 25 extra
No. of pages									
1-2	£21	£37	£53	£69	£85	£101	£117	£133	£21
3-4	£32	£53	£74	£95	£117	£138	£159	£175	£27
5-8	£42	£69	£95	£122	£148	£175	£201	£228	£32
9-16	£53	£85	£117	£148	£180	£212	£244	£276	£37
17-24	£64	£101	£138	£175	£212	£249	£286	£323	£42
each 8pp extra	£16	£21	£27	£32	£37	£42	£48	£53	

OFFICE USE ONLY
Issue:
Vol/part:
Page nos:
Extent:
Price:
Invoice: VR/

PAYMENT BY CREDIT CARD *(Note: we cannot accept American Express)*

Please charge the sum of £___________ to my credit card account.

My Access/Eurocard/Mastercard/Visa number is (*circle appropriate card; no others acceptable*):

Expiry date

Signature: ________________________ Date: ______________

Cardholder's name and address*: ____

____

Address to which your credit card statement is sent. Your offprints will be sent to the address shown at the top of the form.

April 2003