Journal of General Virology (1998), **79**, 2921–2928. Printed in Great Britain

Contribution of *Taq* polymerase-induced errors to the estimation of RNA virus diversity

M. A. Bracho, A. Moya and E. Barrio

Departament de Genètica, Universitat de València, Dr. Moliner 50, 46100 Burjassot, València, Spain

The genetic diversity of a vesicular stomatitis virus population was analysed by RT–PCR, cloning and sequencing of two ~ 500 nucleotide regions of the virus genome. PCR amplifications were performed in parallel experiments with both *Taq* and *Pfu* DNA polymerases, and important differences were observed. Between 10 and 22 mutations were detected when virus populations were analysed by *Taq* amplification (20 clones from each region), whereas amplification of the same samples with *Pfu* revealed between 0 and 5 mutations. PCR fidelity assays, performed under the same PCR conditions as those used in the population analysis, showed that the *Taq* error-rate estimate of $0{\cdot}27 \times 10^{-4}$ misincorporations per bp per cycle was within the range estimated elsewhere from PCR amplification of recombinant plasmids ($0{\cdot}27$–$0{\cdot}85 \times 10^{-4}$ errors per bp per cycle) or from functional assays ($0{\cdot}2$–2×10^{-4} errors per bp per cycle). The error rate of *Taq* was found to be 9·3 times higher than the error rate of *Pfu* with DNA as a template, and about 10 times higher with cDNAs obtained by reverse transcription of viral RNA templates from natural populations. In the present study, we discuss (i) the implications of *Taq* errors on the analysis of genetic variability, based on both the frequency and nature (replacement vs synonymous) of the observed substitutions and (ii) the sample size required to assess the genetic variability in a virus population generated by a single infection.

Introduction

Since *Taq* DNA polymerase substitution errors were surveyed by Saiki *et al.* (1988), new error-rate estimates have been reported in several other studies of *Taq* fidelity and PCR-based determination of virus variability (Kwiatowski *et al.*, 1991; Smith *et al.*, 1997 and references therein). A wide range of substitution error rates, measured by sequencing recombinant plasmids, has been reported, which suggests that the accuracy of published rates is low, probably due to differences in the amplification conditions during PCR. Eckert & Kunkel (1990) have determined the conditions required to minimize *Taq* errors.

Error rates for *Taq* calculated from *in vitro* reversion frequency assays range from $0{\cdot}2$ to 2×10^{-4} errors per bp per cycle (Smith *et al.*, 1997). However, error rates obtained from forward mutation assays for *Pfu* DNA polymerase, which unlike *Taq* exhibits $3' \rightarrow 5'$ exonuclease proof-reading activity, have been reported to be three to four times (Cline *et al.*, 1996) or ten times (Lundberg *et al.*, 1991) lower than those estimated for *Taq*.

Of course, such RT–PCR amplification errors have a direct impact on the estimation of genetic variability of RNA viruses. Firstly, the method of analysis uses retrovirus reverse transcriptase to polymerize cDNA *in vitro* from viral RNA templates, so it is not possible to distinguish between misincorporations that occur during *in vitro* retrotranscription and mutations already present in the viral RNA population under study. In fact, the error rate of Moloney murine leukaemia virus reverse transcriptase on RNA templates *in vitro* is less than one error per 28000 nucleotides (Ji & Loeb, 1992); i.e., the same order of magnitude, or one order of magnitude smaller, than the error rates estimated for viral RNA polymerases. A second, and more important, source of error is introduced during cyclic DNA polymerization by PCR amplification after reverse transcription. There are three possible ways to sample closely related viral genomes: (i) isolation of virus plaques derived from single virus particles (plaque purification or biological cloning) followed by RT–PCR and direct sequencing of the amplified products, (ii) direct sequencing of PCR products obtained from single target sequences

Author for correspondence: M. Alma Bracho.
Fax +34 96 398 30 29. e-mail alma.bracho@uv.es

isolated by limiting dilution and (iii) molecular cloning of a PCR product and sequencing of several recombinant plasmids. In the first two strategies, inaccuracies associated with PCR-induced errors can be reduced or eliminated. However, in the third approach, errors introduced by thermostable polymerases during PCR amplification are added to mutations present in the natural virus populations. Thus, if the nucleotide substitution rate of a given virus is similar to or lower than the misincorporation rate of the DNA polymerase used for PCR, the genetic analysis of the population variability could be affected by those errors.

Virus genetic variability can be studied at various levels, from low degrees of divergence between viral genomes from a single infected host to higher degrees of heterogeneity between different virus serotypes or species. The determination of variability at the within-population level will be more greatly affected by errors caused by the polymerases used in RT–PCR than the analysis of variation at higher levels.

In this report we have analysed natural viral RNA populations by RT–PCR with both *Taq* and *Pfu*, to determine how amplification errors contribute to the observed virus variability. In the light of our results, we discuss how errors may affect estimates of virus variability and suggest ways of avoiding such problems.

Methods

■ **Samples and RNA extraction.** BHK_{21} cells were infected with vesicular stomatitis virus (VSV; Indiana serotype, Mudd–Summers strain) as described elsewhere (Holland *et al.*, 1991). Three of the four samples analysed were obtained from a sequential experiment in which cultured cells were infected as follows with a single plaque of VSV. A single plaque was isolated and diluted in 0·5 ml Dulbecco's modified Eagle's minimum essential medium (DMEM). A fraction of this diluted plaque (3×10^4 p.f.u./ml) was kept for analysis and designated 0 h (representing 0 h of infection). The remainder was used to infect a flask of cultured cells. After this initial infection, samples were taken from the supernatant at 28 h (28h; titre not determined) and 48 h (48 h; $1{\cdot}2 \times 10^5$ p.f.u./ml). Sample 40 d came from another experiment, where virus was passaged on cultured cells, and was taken at passage 40 (2×10^{10} p.f.u./ml), 40 days after initial infection from a single plaque (Elena *et al.*, 1998). Samples were kept at −80 °C until needed. RNA was extracted by a guanidinium thiocyanate–phenol–chloroform method described by Logemann *et al.* (1987), with slight modifications.

■ **RT–PCR, cloning and sequencing.** RNA was reverse-transcribed into cDNA and two regions of the genome were amplified by PCR with VSV-specific primers. One region corresponded to 514 nucleotides of the *G* gene (encoding the glycoprotein) and the other to a 510 nucleotide fragment comprising the 3′ end of the *P* gene (encoding one of the subunits of the viral polymerase), a short non-coding intergenic sequence and a small portion of the *M* gene (encoding the matrix protein). These two regions will be referred to as the G and P regions, respectively. Primers (see below) had about ten bases added to the 5′ end, six of them corresponding to *Xba*I or *Pst*I sites (underlined), allowing their eventual use in restriction site cloning. Position of primers are given according to GenBank accession no. J02428. Primers for the G region were XG1817 (sense, 5′ GGAGTCTAGACTCCCATCAGGTGTCTGGTT 3′, 3813–3835) and XG2370 (antisense, 5′ GAAGTCTAGAAG(C/T)AGCGTCT-(T/G)GAATGTG 3′, 4369–4350). Primers for the P region were PP1751 (sense, 5′ GAGCCTGCAGAAAGACCTTACGGTTGAC 3′, 1751–1770) and PP2280 (antisense, 5′ ATTTCTGCAGATTTCTTACCTTT-CCCCTT 3′, 2300–2280).

Reverse transcription prior to *Taq* amplification was performed in a 20 µl reaction volume containing half the extracted RNA (to ensure sampling of low-frequency viral RNA molecules), 50 mM Tris–HCl pH 8·3, 40 mM KCl, 10 mM DTT, 6 mM $MgCl_2$, 1 mM total dNTPs, 44 U M-MuLV reverse transcriptase (Boehringer Mannheim), 20 U RNase inhibitor (Promega) and 1 µM sense primer. The mixture was incubated at 42 °C for 45 min and 94 °C for 5 min. PCR was then carried out by adding the whole amount of synthesized cDNA (20 µl) to 80 µl PCR pre-mix containing 10 mM Tris–HCl pH 9·0, 50 mM KCl, 1·5 mM $MgCl_2$, 0·25 µM antisense primer and 2·5 U *Taq* (Pharmacia). PCR was performed according to the following profile: initial denaturation at 94 °C for 30 s; 5 cycles of 94 °C for 30 s, 50 °C for 30 s and 72 °C for 30 s; 25 cycles of 94 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s; and a final extension at 72 °C for 7 min.

Reverse transcription prior to *Pfu* amplification was performed in a 20 µl reaction volume containing the same amount of RNA as described above, 25 mM Tris–HCl pH 8·3, 75 mM KCl, 3 mM $MgCl_2$, 4 mM total dNTPs, 50 U M-MLV reverse transcriptase (Amersham), 20 U RNase inhibitor and 1 µM sense primer. The mixture was incubated at 42 °C for 45 min and 94 °C for 5 min. PCR was carried out by adding 10 µl of this cDNA to 90 µl PCR pre-mix containing 22·2 mM Tris–HCl pH 8·0, 11·1 mM $(NH_4)_2SO_4$, 11·1 mM KCl, 2·2 mM $MgSO_4$, 0·11% Triton X-100, 111·1 µg/ml nuclease-free BSA, 5 U *Pfu* (Stratagene), 0·22 µM antisense primer and 0·11 µM sense primer. Conditions for PCR with *Pfu* were the same as in *Taq* amplification except that cyclic polymerization at 72 °C was extended to 2 min. In all cases a single amplified product was observed after electrophoresis on a 2% agarose gel stained with ethidium bromide.

DNA from *Taq* amplifications was precipitated, digested with *Xba*I or *Pst*I (as appropriate) and cloned into similarly digested pUC18. Alternatively, PCR products were cloned without prior digestion into T-tailed pBluescript (Stratagene), as described by Marchuk *et al.* (1990). DNA from *Pfu* amplifications was precipitated and cloned directly into *Eco*RV-digested pBluescript. Recombinant plasmids were purified with QIAgen microspin columns and sequenced by dye terminator cycle sequencing (Perkin-Elmer) on an Applied Biosystems 373 automated sequencer. Sequences were processed and analysed with the Staden (Dear & Staden, 1992) and Wisconsin GCG (Devereux *et al.*, 1984) software packages.

■ **Control experiments to determine *Taq* and *Pfu* errors during PCR.** One pg ($2{\cdot}6 \times 10^5$ molecules) of a recombinant plasmid containing the amplified P region plus a 4 bp insertion (a total of 514 bp) was used as a template for PCR with both *Taq* and *Pfu* under the conditions described above, except that reverse transcriptase, RNase inhibitor and RNA were omitted. After PCR, DNA was precipitated and cloned into T-tailed (*Taq*) or *Eco*RV-digested (*Pfu*) pBluescript. About 10% of the amplified DNA was used to transform *E. coli* DH5α competent cells (transformation efficiency 2×10^6 colonies per µg supercoiled DNA). Nineteen clones from each amplification were sequenced.

We also estimated the fraction of the plasmid added as the template in the control experiments that 'survived' PCR, and could therefore transform *E. coli* cells, as follows: 100 pg template plasmid was subjected to PCR in the absence of *Taq*, and then used to transform competent cells. A second 100 pg template plasmid was used directly to transform competent cells from the same batch. Plasmid DNA subjected to PCR yielded only 4% as many colonies as were obtained after transformation with plasmids that had not undergone PCR.

Results

Mutation frequencies in VSV samples

Four samples of VSV from infected cultured cells were characterized by RT–PCR with both *Taq* and *Pfu* and subsequent sequencing of cloned DNA. RT–PCR conditions did not differ substantially from those commonly used for both polymerases. For *Taq* amplification of DNA, we used the high-fidelity conditions proposed by Kwiatowski *et al*. (1991). *Pfu* amplifications were performed under the conditions recommended by the supplier, which minimize misincorporations.

Our results have been pooled into two groups, one group comprising samples 0 h, 28 h and 48 h, and the other group containing sample 40 d. The reason for this distinction is that although the population dynamics were comparable, the samples were derived from different initial single plaques, and moreover sample 40 d was expected to be more diverse since the time from initial infection to sampling was much longer.

Sample 0 h, corresponding to the initial diluted plaque, was estimated (according to Steinhauer *et al*., 1989) to have undergone at least two cycles of infection and, therefore, at least four rounds of replication (one replication from negative to positive strand and another from positive to negative genome during each cycle of infection). Since the infection cycle of VSV lasts 8–12 h, samples 28 h and 48 h would have undergone at least 2–3 and 4–6 further cycles, respectively. Virus heterogeneity within and between samples 0 h, 28 h and 48 h was therefore expected to be low, since these samples were taken from a population that originated from a single virus particle. According to the point-mutation frequencies estimated for VSV for base substitutions per site (10^{-3}–10^{-4}; see Domingo & Holland, 1994 and references therein), the expected rate of substitution would be 1–10 substitutions per complete genome (11·2 kbp) per round of replication.

Mutations observed in 40 recombinant clones analysed per sample, 20 from the P region and 20 from the G region, are summarized in Table 1. There were clear differences in the estimated nucleotide diversity, depending on the DNA polymerase used in the analysis. This is best illustrated by considering the results of the P region analysis: between 11

Table 1. Frequencies of nucleotide substitution observed in the G and P regions of VSV in different virus samples determined with *Taq* and *Pfu*

The number of nucleotides sequenced per clone was 514 for the G region, 510 for the P region and 514 for the recombinant plasmid (used as a template in the PCR control amplifications performed to determine error rates of both polymerases).

Amplified region	Sample	Number of polymorphic sites			No. of mutations (replacement/ synonymous)	Nucleotide diversity ($\times 10^3$)†	Frequency of unique variable sites ($\times 10^4$)‡
		Total	Unique	Shared			
Taq							
Plasmid	Control	8	8	0	6/2	1·65 (0·48)	0·27
G	0 h	10	10	0	10/0	1·95 (0·53)	0·32
G	28 h	15	15	0	12/3	2·92 (0·60)	0·49
G	48 h	16	15	1	11/5	3·29 (0·85)	0·49
G	40 d	14	9	5	12/2	4·50 (0·59)	0·29
P	0 h	11	11	0	8/3	2·16 (0·77)	0·36
P	28 h	22	21	1	11/7*	4·31 (0·75)	0·69
P	48 h	21	21	0	9/2*	4·30 (0·90)	0·69
P	40 d	15	13	2	9/4	3·48 (0·70)	0·43
Pfu							
Plasmid	Control	0	0	0	0/0	0·00 (0·00)	< 0·03
G	0 h	0	0	0	0/0	0·00 (0·00)	< 0·03
G	28 h	0	0	0	0/0	0·00 (0·00)	< 0·03
G	48 h	0	0	0	0/0	0·00 (0·00)	< 0·03
G	40 d	5	3	2	3/2	2·12 (0·32)	0·10
P	0 h	0	0	0	0/0	0·00 (0·00)	0·00
P	28 h	1	1	0	0/1	0·20 (0·17)	0·03
P	48 h	2	2	0	2/0	0·39 (0·24)	0·07
P	40 d	2	2	0	1/1	0·39 (0·24)	0·07

*Sum of replacement and synonymous substitutions does not equal the total number of polymorphic sites because some substitutions occurred in the non-coding region.

†Defined as the average number of nucleotide differences per site between all pairs of sequences. Values and standard deviations (in parentheses) were estimated according to Nei (1987).

‡Number of unique polymorphic sites per bp per cycle.

and 22 polymorphic sites were observed when PCR products from samples 0 h, 28 h and 48 h obtained with *Taq* were sequenced, while the sequence analysis of amplified products obtained with *Pfu* revealed 0–2 substitutions for the same samples. Another remarkable fact is that, after pooling sequence data from *Taq* PCR amplification of both genes from samples 0 h, 28 h and 48 h, only 2 of 95 mutations corresponded to shared mutations (i.e. they were present in more than one clone obtained from the same sample), while all the other substitutions were unique (see Table 1). Moreover, none of the unique or shared mutations observed in any of the three samples after amplification with either *Taq* or *Pfu* was ever found in the other two samples. Since the shared substitutions found in the *Taq*-amplified products were not detected in the *Pfu*-amplified products, they were probably *Taq* errors which took place during the initial cycles of amplification, rather than real VSV mutations.

With respect to the nature of the mutations observed with *Taq*, the expected ratio for replacement (nonsynonymous) vs synonymous random substitutions would be 3:1 in unconstrained regions (Li & Graur, 1991). When replacement and synonymous substitutions were compared in the P and G region sequences from samples 0 h, 28 h and 48 h, the observed ratio did not deviate significantly from that expected [$\chi^2 = 0{\cdot}004$; 1 degree of freedom (df); $P = 0{\cdot}95$]. A similar analysis of the mutations detected in *Pfu*-amplified samples was not possible due to the small number of mutations observed (two replacement and one synonymous).

Sample 40 d was expected to be more variable, since it was obtained after 40 days of continuous passages of a population initiated with a single plaque. Similarly to samples 0 h, 28 h and 48 h, more polymorphic sites were found for the two regions analysed from *Taq*-amplified products than in those from *Pfu* amplifications (29 and 7 total mutations, respectively; see Table 1). However, the proportion of shared mutations was higher in sample 40 d than in the other three samples, irrespective of the polymerase: 7 of 29 mutations with *Taq* and 2 of 7 with *Pfu* were shared between clones of sample 40 d compared to 2 of 95 with *Taq* and 0 of 3 with *Pfu* for pooled samples 0 h, 28 h and 48 h. None of the unique mutations of *Taq*-amplified products from sample 40 d was also found in *Pfu*-amplified products. However, 3 of 7 shared mutations detected in the two regions amplified with *Taq* were also found in *Pfu*-amplified products (two of them again as shared polymorphic sites and the other as a unique polymorphic site). The proportion of clones containing the same shared polymorphic site was clearly different for *Taq*-amplified products compared to those amplified by *Pfu*: of 20 clones sequenced, 3, 2 and 4 amplified with *Taq* contained each of the shared mutations, against 9, 3 and 1 amplified with *Pfu*. Such differences in the proportions of shared polymorphic sites between *Taq* and *Pfu* are probably due to the stochastic nature of the sampling. Accordingly, the three shared polymorphic sites detected in sample 40 d with both DNA polymerases, as well as the other four substitutions present as unique polymorphic sites detected only by *Pfu*, are interpreted as real VSV mutations, representing heterogeneity within the virus population. The other four shared polymorphic sites found in the *Taq*-amplified fragments from sample 40 d were found in 3, 3, 2 and 2 of 20 clones sequenced, but they were not found in *Pfu*-amplified PCR products and could therefore be considered as either *Taq* errors or real VSV mutations. A χ^2-test of the frequencies of replacement vs synonymous substitutions in the sequences of both regions from sample 40 d amplified by *Taq* again revealed no statistically significant departure ($\chi^2 = 0$; 1 df; $P = 1$) from the null hypothesis of random substitutions in unconstrained regions. Once again, this analysis could not be performed with the *Pfu* data due to the small number of mutations detected.

Control experiments to determine *Taq* and *Pfu* error rates

By using a recombinant plasmid as a template, we introduced an internal control to measure the error rates for both polymerases: this should be taken into account in the analysis of the virus samples. The *Taq* mutation rate control experiment gave the smallest number of unique polymorphic sites observed for this polymerase ($0{\cdot}27 \times 10^{-4}$ per bp per cycle; see Table 1). Although the recombinant plasmid used as template in the PCR was not digested to prevent molecules 'surviving' that could transform competent cells, an estimate of the proportion of transformation-competent template plasmids remaining after PCR denaturation (see Methods) showed that the number of colonies that could arise from the original 0·1 pg template plasmid after PCR amplification and ligation was negligible. Although we discounted the possibility that the smaller number of mutations observed in our control experiment with *Taq* could be caused by these 'surviving' template molecules yielding transformants, the possibility that a smaller than expected number of mutants could be observed as a result of a reduction in the effective number of PCR cycles, arising from an excess of template, cannot be excluded (Lu *et al*., 1995).

We have compared some of the published substitution error rates per cycle for both *Taq* and *Pfu* with those estimated from our control experiment (Table 2). When different measures of error rate were used, values were converted to error rate per cycle to make them comparable. The error rates used in the comparison were obtained by one of four methodologies: sequencing of recombinant plasmids from a PCR product, sequencing after forward mutation assay with a single cycle of PCR, reversion frequency assay after PCR or denaturing gradient gel electrophoresis (DGGE) after PCR. For *Taq*, these estimates differed by two orders of magnitude (from $2{\cdot}1 \times 10^{-4}$ to $< 2{\cdot}3 \times 10^{-6}$), while the estimated error rate from our control experiment was $2{\cdot}7 \times 10^{-5}$ (see Table 2). For *Pfu*, our estimated substitution error rate ($< 3 \times 10^{-6}$) on a

Table 2. Published substitution error frequencies for *Taq* and *Pfu* expressed as errors per bp per cycle

Reference	Polymerase error rate ($\times 10^4$)*	
	Taq	*Pfu*
This study	0·27	< 0·03
Andre *et al.* (1997)	—	0·0065[a]
Cline *et al.* (1996)	0·08[b]	0·013[b]†
Dunning *et al.* (1988)	0·79	—
Eckert & Kunkel (1990)	0·20[a]	—
Ennis *et al.* (1990)	0·23	—
Ennis *et al.* (1990)	0·21	—
Fucharoen *et al.* (1989)	< 0·023	—
Keohavong & Thilly (1989)	2·1[c]	—
Kwiatowski *et al.* (1990)	0·075	—
Lundberg *et al.* (1991)	0·2[b]	0·016[b]
Martell *et al.* (1992)	0·052	—
Meyerhans *et al.* (1990)	0·091	—
Saiki *et al.* (1988)	0·85	—
Smith *et al.* (1997)	0·27	—

*Error rates were obtained by sequencing of recombinant plasmids from a PCR product except those noted, which were obtained as follows: *a*, by sequencing after forward mutation assay with a single-cycle of PCR; *b*, by reversion assay from a PCR product; *c*, by denaturing gradient gel electrophoresis after PCR; or *d*, by constant denaturant capillary electrophoresis. —, Not determined.
†Observed error frequency per template doubling (td). Calculated as 2^{td} = amount of PCR product/amount of starting target.

VSV sequence is consistent with previous estimates obtained by PCR-based forward mutation assay over either 349 bp (Cline *et al.*, 1996) or 182 bp (Lundberg *et al.*, 1991) of the *lac*I gene: $1{\cdot}3 \times 10^{-6}$ and $1{\cdot}6 \times 10^{-6}$ errors per bp per template duplication, respectively.

Categorizing *Taq* mutations

To have a sufficient number of nucleotide substitutions caused by *Taq*, we pooled the mutations observed in the *Taq* control experiment with those observed in samples 0 h, 28 h and 48 h from the *Taq*-based sequence analysis. As seen before, most, if not all, of the mutations observed in these samples can be assumed to be *Taq* errors. However, sequences from sample 40 d were not considered, in order to avoid inclusion of the genuine VSV mutations that are present in this sample.

A summary of the distribution of *Taq* mutations in this global sample is shown in Table 3. Since it is not possible to distinguish which nucleotide of a given nucleotide pair was misincorporated during DNA amplification (i.e. a change from G to A on one strand cannot be distinguished from a change from C to T occurring on the complementary strand), the twelve possible nucleotide substitutions were grouped into six categories. Distributions of *Taq* errors reported by Dunning *et al.* (1988) and Ennis *et al.* (1990) were also included for comparison. Any possible bias resulting from the nucleotide composition of the target sequence was taken into account by introducing a weighting, achieved by recalculating the number of mutations for a 50% G+C content, before the application of the G-statistic test for heterogeneity (Sokal & Rohlf, 1995). Comparison of our study with that of Dunning *et al.* (1988) reveals an equal distribution of mutations ($G_H = 1{\cdot}08$; 3 df; $P = 0{\cdot}78$). The application of the G-statistic test to the comparison of our data with that of Ennis *et al.* (1990) also confirmed the equal distribution of mutations ($G_H = 0{\cdot}79$; 1 df; $P = 0{\cdot}37$). Since *Taq* substitutions involving template nucleotides A or T are more abundant than those involving G or C, the nucleotide composition of the target used to determine the substitution error rate could affect the observed rate. However, the equal distribution of mutations obtained in studies based on different templates suggests that the *Taq* substitution error rate is not affected significantly by the sequence context.

Discussion

Mutation frequencies in VSV samples

PCR-based sequence analysis of four samples from a VSV population revealed dramatic differences in the total number of observed mutations, depending on the chosen polymerase,

Table 3. Distribution of nucleotide substitutions resulting from *Taq* errors

Since it was not possible to distinguish the strand in which a nucleotide of a given pair was misincorporated during DNA amplification, the twelve possible nucleotide substitutions were grouped into six complementary categories.

Reference	Substitution						% G+C content
	G → A C → T	G → T C → A	G → C C → G	A → G T → C	A → T T → A	A → C T → G	
This study	15	1	0	68	14	4	43
Dunning *et al.* (1988)	4	0	0	11	3	1	40
Ennis *et al.* (1990)	12	0	0	16	0	0	64

Taq or *Pfu*. Nucleotide diversity was certainly overestimated when *Taq* (error-prone) was used in the analysis; this was clearly evident from comparison with *Pfu* (almost free of error) amplification of the same samples. Moreover, in the analysis of *Taq*-amplified sequences from samples 0 h, 28 h and 48 h, it is not possible to discount the possibility that observed mutations could be in reality *Taq* errors introduced during PCR. On the contrary, at least three of the mutations observed in sample 40 d amplified by *Taq* were also found in the *Pfu*-amplified products and are therefore likely to represent real VSV mutations.

Nucleotide diversity was expected to increase from samples 0 h to 40 d, as they were obtained from an evolving population originating from a relatively small number of virus particles derived from a single plaque. However, this tendency was evident only when comparing sample 40 d with any of the other three samples. Mutations observed in the *Pfu*-amplified PCR products are probably real mutations present in the virus population; because of the low number of mutations detected, almost nothing can be said about trends of changes in nucleotide diversity from sample to sample. Nonetheless, the increase in nucleotide diversity from sample 0 h to sample 40 d agrees with expectations.

The fact that only three mutations were observed in *Pfu*-amplified products from samples 0 h, 28 h and 48 h suggests that the reported mutation rate for the VSV genome (10^{-3}–10^{-4} substitutions per nucleotide per replication round), obtained by site-specific mutation analysis of 12 sites susceptible to digestion by RNase T_1 (Steinhauer *et al*., 1989), could be an overestimate. A re-estimation of the VSV error rate by other approaches, involving a larger target sequence, might confirm this suggestion.

Mutation frequencies in *Taq* and *Pfu* control experiments

Even though the high-fidelity PCR conditions proposed by Kwiatowski *et al*. (1991) were used in the present study, the reaction conditions were probably responsible for the nucleotide diversity overestimates obtained from the *Taq*-based analysis. Eckert & Kunkel (1990) examined the conditions required for high-fidelity DNA synthesis by *Taq*. They concluded that 'the fidelity of *Taq* polymerase responds not to the absolute $MgCl_2$ concentration, but rather to the amount of $MgCl_2$ in excess over the total deoxynucleotide triphosphates present during DNA synthesis' and also that 'a base substitution error as low as 10^{-5} can be achieved by using equimolar concentrations of $MgCl_2$ and deoxynucleotide triphosphates in *Taq* polymerase reactions.' These authors also observed that departures from equimolarity to molar ratios of 5 or 20 increased error rates to $2{\cdot}0 \times 10^{-4}$ and $2{\cdot}4 \times 10^{-4}$ substitutions per bp per cycle, respectively (Eckert & Kunkel, 1990). Nevertheless, it is not unusual to see published determinations of virus diversity at different taxonomic levels (including the lowest level of a single infected individual) that have been made by molecular cloning and sequencing after PCR amplification with conditions of comparable or even greater departures from equimolarity. This is not surprising, since low error-rate estimates for *Taq* amplification (see Table 2), some reported in papers focusing on *Taq* mutation rates and others determined as additions to virus variability surveys, are abundant and therefore tend to minimize the importance of PCR conditions on error rates. However, as Smith *et al*. (1997) pointed out, some of these control experiments that report low error rates could be affected by priming of the PCR with an excess of template: this can reduce the effective number of PCR cycles and also increase the survival after PCR of the original template molecules that give rise to transformants. Checking whether the low rates reported in some papers are the result either of efficient transformation of bacteria by surviving template molecules or of a reduced number of PCR cycles is difficult, because this type of information is usually not explicit. Careful measurement of the amount of template used and the adoption of precautions to avoid transformation of competent bacteria by the template (e.g. the use of linear DNA templates or the digestion of circular templates with restriction endonucleases either before or after PCR) should be considered in control experiments. Additional parameters that could affect the accuracy and precision of error-rate estimates (and could therefore give rise to the discrepancies between observed values) are the different methodologies used or the nucleotide composition of the target sequence.

There is an important methodological difference between PCR with DNA templates and RT–PCR with RNA templates that could affect the conditions required for high-fidelity *Taq* amplification. This difference becomes relevant in the analysis of variability of RNA viruses by RT–PCR, cloning and sequencing. The reverse transcription reaction must be performed under relatively high $MgCl_2$ concentrations (e.g. 5 mM for Moloney murine leukaemia virus reverse transcriptase) and, if no cDNA purification step is included, this could alter the $MgCl_2$/dNTPs concentration ratio in the PCR. This is particularly relevant when a good sampling of the RNA population is needed, since large amounts of cDNA should be used in the PCR to minimize the probability of cloning several PCR products that are derived from the same original virus cDNA template (Liu *et al*., 1996). In contrast to *Taq*, high-fidelity amplification by *Pfu* requires Mg^{2+} concentrations above 1·5 mM (according to the manual prepared by Stratagene, the supplier). Therefore, for analyses of virus diversity involving RT–PCR, the use of *Pfu* is preferable to *Taq*, since *Pfu* is affected less by the $MgCl_2$ carried over from the reverse transcription reaction.

Alternatively, other methods could be used to reduce the incidence of artefacts, such as avoiding cloning steps by direct sequencing of PCR products obtained at limiting dilutions. In the case of studies on viruses infecting cultured cells (such as VSV), an alternative to molecular cloning is biological cloning based on direct sequencing of PCR products obtained from

individually isolated plaques. An additional advantage of these two methods is that they avoid more efficiently the artefactual PCR products produced by *in vitro* recombination (Meyerhans *et al*., 1990). In addition, biological cloning avoids the sampling of defective virus particles.

Categorizing *Taq* mutations

The distribution of *Taq* mutations into six categories of nucleotide substitution (see Table 3) appears robust enough for such a distribution to serve as a comparative tool to check whether the distribution of observed changes, especially unique substitutions, in virus samples indicates the presence of *Taq* misincorporations, which would result in overestimation of the virus nucleotide diversity. Such a test could be used generally with the following limitation: the polymerase of the virus under study must have its own characteristic distribution of mutations that must be different from that of *Taq* in order for the test to give a conclusive result.

General conclusions

The analysis of sequence variation in samples with very low sequence diversity is particularly sensitive to sequence errors, irrespective of their source (RT–PCR, sequencing, data-handling), because the signal-to-noise ratio is lower than in samples with higher levels of sequence diversity. Our estimates of nucleotide diversity obtained with *Taq* (error-prone) and *Pfu* (almost error-free) differ to a large extent: spurious conclusions on nucleotide diversity could have been reached if only *Taq* had been used.

In a study of the effect of sequencing errors on evolutionary analysis, Clark & Whittam (1992) showed that the rate of increase in apparent nucleotide divergence between sequences depends on the nucleotide diversity of the sample. In this way, less-diverse sequences appear to increase in diversity at a higher rate with increasing error rates. They also concluded that sequencing errors strongly affect the estimation of parameters such as nucleotide diversity and the reconstruction of gene phylogenies. An example of the effect of *Taq* errors can be observed in the phylogenetic analysis of hepatitis G virus isolates reported by Viazov *et al*. (1997). According to this study, sequences from a single patient, obtained by cloning PCR products, are connected by longer 'branches' and are thus more diverse than virus sequences from different patients from around the world, obtained by direct sequencing of PCR products. This illogical result suggests the presence of *Taq*-induced sequence errors in their data set.

The presence of *Taq*-induced errors in the data of Viazov *et al*. (1997) is also indicated by the result of a χ^2-test of numbers of replacement and synonymous substitutions (not shown). Since *Taq* misincorporations during PCR are distributed randomly in coding sequences, independent of the codon position, the expected ratio of *Taq*-induced errors between replacement and synonymous substitutions is 3:1. This ratio is the same as that observed in unconstrained sequences (e.g. non-coding regions, introns and pseudogenes), where the effects of most mutations can be overcome. This kind of substitution distribution can lead to other erroneous conclusions, for instance on the potential action of positive selection on virus diversity, since replacement/synonymous mutation ratios significantly higher than 1 within coding regions are considered as evidence of positive selection (Endo *et al*., 1996).

Even under adequate PCR amplification conditions, it is obvious that the sample size required to detect significant differences between two given sequence populations is highly dependent on the overall nucleotide diversity of the samples. In experiments designed to distinguish significant differences in nucleotide diversity over time, poor sampling may lead to ambiguous results. For instance, Devereux *et al*. (1997) performed a decade-long evolutionary study of hepatitis C virus in two infected patients by sampling five clones of a 132 nucleotide portion of the NS4 gene each year (a total of 660 nucleotides per year). The oscillatory pattern of diversity over time observed in their study better reflects stochastic behaviour of sampling rather than real changes in the trends of average divergences between years.

In this study we have not made any reference to insertion or deletion events, which were in fact detected (data not shown). This is because in contrast to nucleotide substitutions, which seem to occur independently of the target sequence, we have observed that insertions and deletions are more dependent on the presence of nucleotide 'runs' in the target sequence.

We gratefully acknowledge R. Miralles for her helpful suggestions during the course of this work, and Dr S. F. Elena for his critical reading of the manuscript. This work was supported by grant PM97-0060-C02-02 from DGES to A.M. and by a fellowship from MEC to M.A.B.

References

Andre, P., Kim, A., Khrapko, K. & Thilly, W. G. (1997). Fidelity and mutational spectrum of *Pfu* DNA polymerase on a human mitochondrial DNA sequence. *Genome Research* **7**, 843–852.

Clark, A. G. & Whittam, T. S. (1992). Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution* **9**, 744–752.

Cline, J., Barman, J. C. & Hogrefe, H. H. (1996). PCR fidelity of *Pfu* polymerase and other thermostable DNA polymerases. *Nucleic Acids Research* **24**, 3546–3551.

Dear, S. & Staden, R. (1992). A standard file format for data from DNA sequencing instruments. *DNA Sequence* **3**, 107–110.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

Devereux, H. L., Brown, D., Dusheiko, G. M., Emery, V. C. & Lee, C. A. (1997). Long-term evolution of the 5′UTR and a region of NS4 containing a CTL epitope of hepatitis C virus in two haemophilic patients. *Journal of General Virology* **78**, 583–590.

Domingo, E. & Holland, J. J. (1994). Mutation rates and rapid evolution of RNA viruses. In *The Evolutionary Biology of Viruses*, pp. 161–184. Edited by S. S. Morse. New York: Raven Press.

Dunning, A. M., Talmud, P. & Humphries, S. E. (1988). Errors in the polymerase chain reaction. *Nucleic Acids Research* **16**, 10393.

Eckert, K. A. & Kunkel, T. A. (1990). High fidelity by the *Thermus aquaticus* DNA polymerase. *Nucleic Acids Research* **18**, 3737–3744.

Elena, S. F., Davila, M., Novella, I. S., Holland, J. J., Domingo, E. & Moya, A. (1998). Evolutionary dynamics of fitness recovery from the debilitating effects of Muller's ratchet. *Evolution* **52**, 351–356.

Endo, T., Ikeo, K. & Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution* **13**, 685–690.

Ennis, P. D., Zemmour, J., Russell, D. S. & Parham, P. (1990). Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: frequency and nature of errors produced in amplification. *Proceedings of the National Academy of Sciences, USA* **87**, 2833–2837.

Fucharoen, S., Fucharoen, G., Fucharoen, P. & Fukamaki, Y. (1989). A novel ochre mutation in the β-thalassemia gene of a Thai. Identification by direct cloning of the entire β-globin gene amplified using polymerase chain reactions. *Journal of Biological Chemistry* **264**, 7780–7783.

Holland, J. J., de la Torre, J. C., Clarke, D. K. & Duarte, E. A. (1991). Quantitation of the relative fitness and great adaptability of clonal populations of RNA viruses. *Journal of Virology* **65**, 2960–2967.

Ji, J. P. & Loeb, L. A. (1992). Fidelity of HIV-1 reverse transcriptase copying RNA *in vitro*. *Biochemistry* **31**, 954–958.

Keohavong, P. & Thilly, W. G. (1989). Fidelity of DNA polymerases in DNA amplification. *Proceedings of the National Academy of Sciences, USA* **86**, 9253–9257.

Kwiatowski, J., Skarecky, D., Hernandez, S., Pham, D., Quijas, F. & Ayala, F. J. (1991). High fidelity of the polymerase chain reaction. *Molecular Biology and Evolution* **8**, 884–887.

Li, W. H. & Graur, D. (1991). *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer.

Liu, S. L., Rodrigo, A. G., Shankarappa, R., Learn, G. H., Hsu, L., Davidov, O., Zhao, L. P. & Mullins, J. I. (1996). HIV quasispecies and resampling. *Science* **273**, 415–416.

Logemann, J., Schell, J. & Willmitzer, L. (1987). Improved method for the isolation of RNA from plant tissues. *Analytical Biochemistry* **163**, 16–20.

Lu, M., Funsch, B., Wiese, M. & Roggendorf, M. (1995). Analysis of hepatitis C virus quasispecies populations by temperature gradient gel electrophoresis. *Journal of General Virology* **76**, 881–887.

Lundberg, K. S., Shoemaker, D. D., Adams, M. W., Short, J. M., Sorge, J. A. & Mathur, E. J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from *Pirococcus furiosus*. *Gene* **108**, 1–6.

Marchuk, D., Drumm, M., Saulino, A. & Collins, F. S. (1990). Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *Nucleic Acids Research* **19**, 1154.

Martell, M., Esteban, J. I., Quer, J., Genesca, J., Weiner, A., Esteban, R., Guardia, J. & Gomez, J. (1992). Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *Journal of Virology* **66**, 3225–3229.

Meyerhans, A., Vartanian, J. P. & Wain-Hobson, S. (1990). DNA recombination during PCR. *Nucleic Acids Research* **18**, 1687–1691.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491.

Smith, D. B., McAllister, J., Casino, C. & Simmonds, P. (1997). Virus 'quasispecies': making a mountain out of a molehill? *Journal of General Virology* **78**, 1511–1519.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry*, 3rd edn. New York: W. H. Freeman.

Steinhauer, D. A., de la Torre, J. C. & Holland, J. J. (1989). High nucleotide substitution error frequencies in clonal pools of vesicular stomatitis virus. *Journal of Virology* **63**, 2063–2071.

Viazov, S., Riffelmann, M., Khoudyakov, Y., Fields, H., Varenholz, C. & Roggerndorf, M. (1997). Genetic heterogeneity of hepatitis G virus isolates from different parts of the world. *Journal of General Virology* **78**, 577–581.

Received 17 April 1998; Accepted 4 August 1998