Virus ‘quasispecies’: making a mountain out of a molehill?

Donald B. Smith, Jane McAllister, Carmela Casino and Peter Simmonds

Department of Medical Microbiology, University of Edinburgh, Medical School, Teviot Place, Edinburgh EH8 9AG, UK

Introduction

‘Science begins with myth, and with the criticism of myths.’
Karl Popper

The study of virus variation covers a wide range of phenomena. At one extreme are the comprehensive differences that allow viruses to be arranged taxonomically into different families, genera and species. Less marked differences provide the basis for grouping viruses into serotypes, genotypes or phenotypes. In turn, these different groupings each comprise viruses that are not identical to each other but vary between different infected individuals and are classed as distinct isolates. The focus of this review is on a still finer level of variation, that existing within the population of viruses co-infecting a single individual, often referred to as a ‘quasispecies’.

The term ‘quasispecies’ has most commonly been used to describe the heterogeneous population structure of human immunodeficiency virus (HIV), other lentiviruses and hepatitis C virus (HCV), but it has also been applied to picornaviruses, rhabdoviruses, coronaviruses, bunyaviruses, hepadnaviruses, reoviruses, arenaviruses and bunaviruses, as well as plant tobamoviruses and bromoviruses. Increasingly, there has been a tendency to explain certain features of virus natural history as a consequence of the heterogeneity of virus populations, which is itself viewed as a result of the error-prone replication of virus genomes (reviewed in Holland et al., 1992; Domingo et al., 1994, 1996). However, we argue here that the investigation of virus ‘quasispecies’ has often been marred by inconsistency in terminology and the introduction of artefactual substitutions during PCR amplification of virus sequences. We discuss the implications of this hypothesis, review alternative methods for the investigation of virus heterogeneity, and suggest that some explanations that relate the existence of virus ‘quasispecies’ to aspects of virus natural history may be logically flawed.

What are ‘quasispecies’?

One difficulty with discussing virus ‘quasispecies’ is that the word has been used in a variety of different ways. The concept of quasispecies was introduced to describe replicons in the early stages of the evolution of life, and in virology it is used to describe a population of viruses that share a common origin but which have distinct genomic sequences as a result of mutation, drift and the impact of selection. A single individual might therefore be infected with many distinct variants that together are a ‘quasispecies’, or form a ‘quasispecies distribution’. Confusingly, ‘quasispecies’ has sometimes been used to refer to each of the distinct virus sequences that together make up the quasispecies (Enomoto et al., 1994; Moribe et al., 1995; Gonzalez-Peralta et al., 1996; Sherman et al., 1996), to describe reassortants of viruses with segmented genomes (Gouvea & Brantly, 1995), to distinguish an isolate from that present in other infected individuals (Delwart et al., 1995) or to describe the unrelated variants in an individual infected from multiple sources (Toyoda et al., 1996). Here we use ‘quasispecies’ to describe the whole population of phylogenetically related variants observed within a single infected individual, although we note that such ‘quasispecies’ usually will not have reached equilibrium as required for mathematical models of the behaviour of quasispecies under different conditions (Eigen, 1996).

Relationship between mutation rate and heterogeneity

An important distinction is that between the mutation rate of a virus polymerase, and the diversity of a virus population. The relation between these is not straightforward, but depends on the direction and strength of selection, the number of cycles of copying of the genome in each replication cycle and the number of replication cycles. In the absence of selection, the heterogeneity at each nucleotide position will increase with each copying of the nucleic acid by an increment approximately equal to the error rate. Eventually, substitutions will become saturated and if there are no mutational biases, heterogeneity will approach 0\(\pm 75\). With positive selection, heterogeneity will increase more quickly and certain substitutions may eventually become fixed, while at sites subject to negative selection, heterogeneity will be lower. A factor that can also reduce virus diversity is the occurrence of bottlenecks in population size due to virus transmission by small numbers of infectious particles, or resulting from selection in the infected host for cell tropism, resistance to antivirals or resistance to immune-mediated neutralization. Hence, the level of heterogeneity of the virus population within a particular sample is not simply...
related to the mutation rate of the virus replicase, but is dependent upon many different aspects of the natural history of infection.

**Measurement of mutation rates and heterogeneity**

Estimates of polymerase error rates have been made from the frequency of misincorporation during *in vitro* copying of homopolymeric templates, although much higher error rates are observed on such artificial templates than on nucleic acids encoding functional proteins. Such *in vitro* error rates can still be much higher than the rates observed *in vivo* (Mansky & Temin, 1995), and so it is more accurate to estimate mutation rates from the sequence analysis of single isolated plaques or virus genomes, although this is very labour intensive. In a limited number of studies the frequency of nucleotide substitution has been observed to be less than $2.1 \times 10^{-6}$/nt/replication cycle for poliovirus, $1.5 \times 10^{-5}$ for influenza virus (Parvin et al., 1986), more than $10^{-4}$ for a single nucleotide site in vesicular stomatitis virus (Steinhauer & Holland, 1986), and $0.48-3.4 \times 10^{-5}$ for various retroviruses (Leider et al., 1988; Mansky & Temin, 1995; Kim et al., 1996) (and references therein).

An indirect way of assessing the extent of variation within a virus population is to measure the frequency of a selectable phenotypic marker. This has been done for a variety of markers such as resistance to antibody-mediated neutralization, resistance to antiviral substances, for properties such as host range or phenotype, or from the reversion frequency of nonsense or temperature-sensitive mutants. Where such studies have been done on plaque-purified virus stocks the frequency of mutants has generally been between $10^{-3}$ and $10^{-6}$, and a similar range is obtained if figures are corrected for

<table>
<thead>
<tr>
<th>Study</th>
<th>Virus</th>
<th>Region</th>
<th>No. of clones</th>
<th>Length</th>
<th>No. of PCR cycles</th>
<th>No. of variants</th>
<th>No. of sporadics</th>
<th>Frequency of sporadics*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martell et al. (1992)</td>
<td>HCV</td>
<td>5' NCR</td>
<td>20</td>
<td>177</td>
<td>35</td>
<td>9</td>
<td>9</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NS3</td>
<td>20</td>
<td>240</td>
<td>35</td>
<td>11</td>
<td>12</td>
<td>0.71</td>
</tr>
<tr>
<td>Martell et al. (1994)</td>
<td>HCV</td>
<td>5' NCR</td>
<td>17</td>
<td>232</td>
<td>35</td>
<td>7</td>
<td>4</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5' NCR</td>
<td>16</td>
<td>232</td>
<td>35</td>
<td>5</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5' NCR</td>
<td>12</td>
<td>232</td>
<td>35</td>
<td>4</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2</td>
<td>19</td>
<td>243</td>
<td>35</td>
<td>19</td>
<td>31</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2</td>
<td>21</td>
<td>243</td>
<td>35</td>
<td>17</td>
<td>21</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E2</td>
<td>12</td>
<td>235</td>
<td>35</td>
<td>11</td>
<td>14</td>
<td>1.42</td>
</tr>
<tr>
<td>Lu et al. (1995)</td>
<td>HCV</td>
<td>5' NCR</td>
<td>18</td>
<td>240</td>
<td>30</td>
<td>3</td>
<td>2</td>
<td>0.15</td>
</tr>
<tr>
<td>Murakawa et al. (1992)</td>
<td>HCV</td>
<td>NS5a</td>
<td>27</td>
<td>271</td>
<td>35</td>
<td>11</td>
<td>10</td>
<td>0.43</td>
</tr>
<tr>
<td>Smith et al. (1997a)</td>
<td>HCV</td>
<td>NS5b</td>
<td>10</td>
<td>222</td>
<td>60</td>
<td>3</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NS5b</td>
<td>7</td>
<td>222</td>
<td>60</td>
<td>5</td>
<td>3</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NS5b</td>
<td>9</td>
<td>222</td>
<td>60</td>
<td>4</td>
<td>2</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NS5b</td>
<td>7</td>
<td>222</td>
<td>60</td>
<td>4</td>
<td>3</td>
<td>0.32</td>
</tr>
<tr>
<td>Okamoto et al. (1992)</td>
<td>HCV</td>
<td>complete</td>
<td>3</td>
<td>9412</td>
<td>35</td>
<td>3</td>
<td>37</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>complete</td>
<td>3</td>
<td>9412</td>
<td>35</td>
<td>3</td>
<td>169</td>
<td>1.71</td>
</tr>
<tr>
<td>Vartanian et al. (1992)</td>
<td>HIV</td>
<td>tat</td>
<td>106</td>
<td>219</td>
<td>35</td>
<td>31</td>
<td>21</td>
<td>0.38†</td>
</tr>
<tr>
<td>Najera et al. (1995)</td>
<td>HIV</td>
<td>pol</td>
<td>21</td>
<td>321</td>
<td>70</td>
<td>6</td>
<td>10</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pol</td>
<td>20</td>
<td>321</td>
<td>70</td>
<td>19</td>
<td>11</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pol</td>
<td>21</td>
<td>321</td>
<td>70</td>
<td>20</td>
<td>15</td>
<td>0.33</td>
</tr>
<tr>
<td>Viazov et al. (1997)</td>
<td>HGV</td>
<td>NS5b</td>
<td>6</td>
<td>354</td>
<td>70</td>
<td>6</td>
<td>11</td>
<td>0.73</td>
</tr>
<tr>
<td>and personal</td>
<td></td>
<td>NS5b</td>
<td>5</td>
<td>354</td>
<td>70</td>
<td>5</td>
<td>8</td>
<td>0.65</td>
</tr>
<tr>
<td>Plyusnin et al. (1994)</td>
<td>Hantavirus</td>
<td>S segment</td>
<td>55</td>
<td>550</td>
<td>30</td>
<td>28</td>
<td>27</td>
<td>0.3</td>
</tr>
</tbody>
</table>

* Frequency of sporadic substitutions per nucleotide sequenced per cycle of PCR × 10⁻⁴
† Data for sporadic amino acid substitutions; the frequency of sporadic nucleotide substitutions is assumed to be 50% higher.
‡ Data for particular amino acid substitutions previously associated with antiviral resistance; the substitution frequency is calculated assuming that each amino acid replacement is produced by a single nucleotide substitution at one position in the codon.
the number of different mutations that can produce the selected phenotype (Smith & Inglis, 1987).

The bulk of information about virus heterogeneity derives from the sequence analysis of multiple cloned copies of virus genomes derived from a single infected individual, and for viruses such as HCV and HIV there is considerable interest in the clinical implications of the extent and nature of the heterogeneity observed. However, in this case the virus mutation rate cannot be easily inferred since there are many unknown parameters such as the diversity of the infecting source, the number of replication cycles, the strength and direction of selection at different sites, and the history of population growth and decline. An example of this type of data comes from the sequence analysis of HCV from an individual recently infected through blood transfusion (Martell et al., 1992). Of 20 clones of a 240 nt fragment of the NS3 gene, nine were identical, two had the same two substitutions, while the remaining nine were distinct (Table 1). Similar results have been obtained for other regions of the HCV genome or between the multiple clones used to obtain a complete HCV genome sequence, or for subgenomic regions of other viruses (Table 1).

**Artefactual sources of variation**

One difficulty with interpreting this kind of nucleotide sequence data is that almost all virus nucleotide sequence information obtained since 1988 has derived from templates produced by PCR amplification of virus DNA or of virus-specific cDNA following reverse transcription. Both reverse transcriptase (RT) and Thermus aquaticus DNA polymerase (Taq) have relatively high error rates, and for Taq this is compounded by the multiple cycles of copying during PCR. Any errors produced during RT–PCR amplification of the virus genome will be retained amongst plasmid clones of the amplification product and could therefore be interpreted as virus heterogeneity. In passing, it should be noted that errors introduced during manual or automatic sequence analysis can occur at the level of 0.5–1% (Koop et al., 1993). Although such errors are removed when a consensus sequence is generated from several independent clones, this method obviously cannot be applied to the analysis of individual variants in a virus population, and so great care must be taken in the preparation and checking of sequence data.

The error rate of retrovirus RT enzymes has been estimated at $10^{-3}$ on homopolymeric templates, while on complex RNA templates the error rate falls to about $10^{-4}$. Substitutions introduced during reverse transcription cannot be distinguished from pre-existing heterogeneities in the virus population. Error rates for Taq can be estimated from the reversion frequency of defined mutant plasmid templates following copying in vitro, and range from $0.2–2 \times 10^{-4}$ errors/bp (Lundberg et al., 1991; Barnes, 1992). Similar estimates are obtained from sequence analysis of clones derived by PCR from a template of defined sequence. For example, amongst 28 clones of the human HLA-DPβ gene (239 nt), 17 introduced substitutions were observed (Saiki et al., 1988), giving an error rate after 30 cycles of PCR of $0.85 \times 10^{-4}$/nt/cycle. Another estimate is provided by comparison of full-length genome copies of HIV-1 generated by 25 cycles of PCR in which 13/9000 positions differed because of PCR errors, equivalent to an error rate of $0.6 \times 10^{-4}$ (Salminen et al., 1995). We have also estimated the error rate of Taq by amplifying cDNA for the HCV core gene at a dilution where only 1 out of 10 replicate PCR reactions was positive. The PCR product from this limiting dilution is expected to derive from a single cDNA molecule, and consistent with this, all the substitutions observed were sporadic, occurring in only one of the clones sequenced. After nested PCR (50 amplification cycles), eight substitutions were detected amongst 16 clones of 359 nt, giving an error rate of $0.27 \times 10^{-4}$.

Assuming no selection against mutants, the expected frequency of mutations amongst PCR products after n cycles is the mutation rate $n$, which for 30 cycles of PCR would be $0.6–6 \times 10^{-3}$/nt. For clones of a PCR product of 300 bp this is equivalent to each clone containing $0.2–2$ artefactual substitutions, approximately two-thirds of which are expected to be nonsynonymous. Most artefactual substitutions will be sporadic (i.e. found in only one of the multiple clones sequenced) since an error occurring after the first few cycles of PCR will represent a minority of the final PCR product, while independent substitutions at the same nucleotide site are also unlikely to be observed unless the number of clones sequenced is very large. In contrast, substitutions that are present in more than one clone are likely to represent segregating polymorphisms actually present in the virus population.

Significantly, sporadic substitutions within the range expected for Taq errors have been observed amongst cloned virus sequences in several published studies, but this heterogeneity has been interpreted as indicating diversity of the virus population (Table 1). The frequency of sporadic substitution/nt sequenced/cycle of PCR in these studies ranges from $0.15–1.7 \times 10^{-4}$, similar to the error rate of Taq during amplification of homogeneous nucleic acid targets ($0.27–0.85 \times 10^{-4}$) or by functional assays ($0.2–2 \times 10^{-4}$), suggesting that most sporadic substitutions are indeed artefacts.

**Conflict with previous control experiments**

The conclusion that cloned DNA contains a high frequency of artefactual substitutions is at odds with some previous observations. For example, in one study (Martell et al., 1992) cloned DNA for the HCV 5′ NCR was amplified by PCR for 35 cycles and recloned. Only one substitution was observed amongst 23 clones of 237 nt, giving a frequency of $5.2 \times 10^{-6}$/nt/amplification cycle. In another study of the same region (Lu et al., 1995), one substitution was detected amongst 30 clones of 278 nt after amplification of diluted plasmid DNA.
for 30 cycles, giving a substitution frequency of $3 \times 10^{-8}$. Amplification of a plasmid containing HIV tat sequences for 35 cycles produced only two errors amongst 20 clones of 314 bp, giving a substitution frequency of $9.1 \times 10^{-6}$ (Meyerhans et al., 1989). These frequencies of substitution are 3–50-fold lower than those observed in other measurements of the error rate of Taq described above. Two different reasons can be suggested to account for these discrepant findings. First, if reactions are primed with a large amount of template, saturation may be reached after only a few cycles and the effective number of amplification cycles will then be fewer than supposed. For example, 30 productive cycles of PCR should produce amplification by a factor of $10^6$, but for PCR of a 300 bp fragment yielding 1 µg this would only be achieved if the input of DNA was 1 fg or about 3000 molecules. If reactions contain larger quantities of target DNA such as 10 pg (Lu et al., 1995), the effective number of cycles would be only 17.

A second explanation of the low apparent error rate of Taq DNA polymerase in some studies is that the plasmid DNA used as the target for amplification might survive the cycles of denaturation and itself give rise to transformants. This possibility was tested by making dilutions of a plasmid containing the HCV core gene and amplifying samples by PCR using primers specific for core. Transformation of competent cells with the products of the PCR reactions revealed that just as many transformants were observed using an equivalent dilution of plasmid DNA not subjected to PCR. No substitutions were observed amongst a 383 nt sequence in 11 different transformants from the PCR reactions, as expected for plasmid DNA replicated by a proof-reading bacterial DNA polymerase with an error rate of about $5 \times 10^{-10}/nt$. Since the transformation efficiency of competent cells is typically $10^8–10^9/µg$ of plasmid DNA, as little as 1 pg of DNA could produce transformants outnumbering those obtained by ligation of an amplified fragment into vector DNA. Hence, studies using plasmid DNA to infer the rate of artefactual substitution amongst PCR products might underestimate the number of artefacts unless survival of the target is prevented by restriction at multiple sites, and carry over is minimized by size-fractionation and purification of the amplified product.

Methods of avoiding artefactual substitutions

Studies of sequence heterogeneity within infected individuals therefore need to take precautions to ensure that artefacts are not introduced during the amplification of virus genomes. One approach might be to limit the extent to which Taq errors are introduced during amplification and this can be done by minimizing the number of PCR cycles, for example by avoiding nested PCR, or by optimizing cycling parameters and conditions (Eckert & Kunkel, 1990).

Another potential approach is to use thermostable polymerases that have a lower error rate than Taq. Examples include the DNA polymerases from Thermococcus litoralis (Vent, New England Biolabs), Therma brockianus (Thb, NBL Gene Sciences) or an N-terminal deletion of Taq (Barnes, 1992), all of which have error rates 30–50% of that of Taq in direct comparisons. More striking is the 90% reduction in error rate reported for the DNA polymerase Py (Stratagene) isolated from Pyrococcus furiosus (Lundberg et al., 1991). Note that although the number of artefactual substitutions occurring during PCR can be reduced by using polymerases with higher fidelities, errors produced during reverse transcription of virus RNA will still occur at a frequency of about $10^{-4}$.

A more stringent way of avoiding artefactual errors is to avoid the analysis of cloned sequences and instead to directly sequence PCR products derived by amplification of a single target molecule (Simmonds et al., 1990). In this case, an error produced during PCR will represent a minority of the population of amplification products and will not be detected against the majority of templates that have no mutation at that position. Even if an error occurs during the first PCR cycle, half of the templates will still have the original nucleotide and an ambiguity should be observed on the sequencing gel. In order to confirm the efficacy of this method, we have made a direct comparison of the level of heterogeneity observed amongst HCV sequences derived from cloned PCR products with that observed between PCR products obtained at limiting dilution and sequenced directly. For three different HCV-infected individuals, the rate of sporadic substitution was higher amongst the cloned products than amongst PCR products obtained at limiting dilution and sequenced directly (Fig. 1). Furthermore, most of the sporadic substitutions in the cloned products were nonsynonymous, as expected if they were artefacts produced during PCR. The difference between the frequency of sporadic substitutions amongst cloned PCR products compared with that amongst direct sequences ranged from 0.16–1.3 $\times 10^{-4}/nt$ cycle of amplification, similar to the range of estimates for the error rates of Taq. Similarly, the observation that about 10% of cloned HIV provirus DNAs were defective for tat, which makes up only 2.3% of the virus genome, might be taken to suggest that 98.5% of provirus was incapable of replication (Vartanian et al., 1992). However, much lower frequencies of defective viruses (<50%) are observed by direct sequence analysis of the env gene from individual provirus molecules obtained by PCR amplification at limiting dilution (Balfe et al., 1990). The extent of virus heterogeneity can also be overestimated when gel mobility techniques are used to screen multiple clones for variant sequences. For example, temperature gradient gel electrophoresis (TGGE) can detect single nucleotide differences between a target sequence and a master sequence, and has been used to assess the degree of heterogeneity in infected individuals (Lu et al., 1995). However, since a large proportion of clones are expected to contain substitutions introduced during PCR, many of the variants detected by this technique could be artefacts. Sporadic artefacts
Fig. 1. Frequency of sporadic substitutions assessed by different methods. Virus RNA from three HCV-infected individuals (A, B and C) was purified and amplified by reverse transcription and nested PCR using primers 2174 (5’ TTCATCCAYGTRCASC- CRAACCA 3’, positions 1667–1645, numbered from the AUG initiation codon) and 2173 (5’ CAYGNATGGCNTGGGAYA- TGATG 3’, positions 946–969) for the first round. For the second round of PCR, primers were 8914 (5’ CCGGATCCGGGT- GCCTCAGTTGGAGTCCTGGCGGGC 3’ positions 1048–1074) and 2070 (5’ GGAATTCGTGAARCARTACACYGGRCRCCANAC 3’, positions 1529–1504) for cloning, and 2172 (5’ CCGGATCCATGATGMTNAAYTGGTCNCC 3’, positions 964–983) and 588 (5’ GGYGSGTARTGCCAGCARTANGG 3’, positions 1468–1450) for limiting dilution (30 cycles for each round of PCR). PCR products were either sequenced after cloning into the BamHI and EcoRI sites of pUC18 using standard methods, or sequenced directly from reactions set-up at limiting dilution for virus cDNA in which one primer was biotinylated, from which single strands were purified on streptavidin-coated magnetic beads (Dynabeads, Dynal). The frequency of sporadic substitutions in each group of sequences was calculated as the total number of sporadic nucleotide or amino acid substitutions/(number of clones or single molecules × sequence length × number of PCR cycles).

can only be distinguished from genuine segregating polymorphisms by sequence analysis of variants.

**Defining virus heterogeneity**

The argument of the previous sections is that some of the variation observed in virus populations is artefactual. A consequence of this is that variation between clones of virus genomes generated by PCR should be discounted unless heterogeneity is greater than expected from the error rate of *Taq* (0.2–2 × 10⁻⁴/nt/cycle). Unless sporadic substitutions occur at a level higher than that expected from *Taq* errors (between 0.2 and 2 substitutions per clone for a 300 nt fragment after 30 cycles of PCR), or measures have been taken to avoid the incorporation of *Taq* errors, a description of heterogeneity in a virus population should be limited to segregating polymorphisms. Given this re-definition, the extent of heterogeneity in virus populations may be significantly less than previously supposed.

An example of how this re-definition affects the interpretation of virus heterogeneity is provided by sequence analysis of multiple clones of the *pol* gene of HIV from individuals who had not been treated with antiviral drugs (Najera et al., 1995). Two sporadic substitutions were observed that produced amino acid substitutions expected to alter the sensitivity of the virus to antiviral drugs. However, a total of 32 different amino acid substitutions known to produce drug resistance were monitored in this study, and the observed frequency of substitution at these sites (0.32 × 10⁻⁴/nt/cycle) was similar to the overall frequency of sporadic substitutions in the rest of the *pol* gene amongst these clones (mean of 0.35 × 10⁻⁴, Table 1). Since this frequency is within the range expected from *Taq* errors these data do not support the assertion that drug-resistant mutants are present at a high frequency prior to treatment.

Another example of misinterpretation is provided by the ‘hypervariable’ regions of HCV and HIV-1 in which non-synonymous substitutions are relatively common. Although these regions typically contain segregating amino acid polymorphisms, sequence analysis of multiple clones often reveals numerous sporadic amino acid substitutions. No distinction is usually made between the segregating and sporadic poly-
morphisms, and the observation that most clones differ has been interpreted as evidence for the selection of neutralization-resistant variants. However, for the 25 amino acid hypervariable region at the N terminus of the HCV E2 envelope glycoprotein, the number of sporadic nonsynonymous substitutions in several studies is within the range expected from the error rate of Taq (Table 2), suggesting that many of these represent artefactual substitutions rather than minor antigenic variants. This possibility is consistent with the observation in some studies that peptides bearing sporadic substitutions were as well-recognized by patient serum as the consensus sequence (Kato et al., 1993; Yoshioka et al., 1996).

### Table 2. Frequency of sporadic amino acid substitutions in the HCV hypervariable region

<table>
<thead>
<tr>
<th>Study</th>
<th>No. of samples</th>
<th>Clones per sample</th>
<th>Total sequences</th>
<th>PCR cycles</th>
<th>Expected sporadics*</th>
<th>Observed sporadics†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moribe et al. (1995)</td>
<td>9</td>
<td>8–11</td>
<td>87</td>
<td>80</td>
<td>0.8–8</td>
<td>3.9</td>
</tr>
<tr>
<td>Okada et al. (1992)</td>
<td>9</td>
<td>10</td>
<td>90</td>
<td>75</td>
<td>0.75–7.5</td>
<td>3.7</td>
</tr>
<tr>
<td>Okamoto et al. (1992)</td>
<td>3</td>
<td>8–10</td>
<td>27</td>
<td>35</td>
<td>0.32–3.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Odeberg et al. (1995)</td>
<td>1</td>
<td>14</td>
<td>14</td>
<td>55</td>
<td>0.9–9</td>
<td>10</td>
</tr>
<tr>
<td>Higashi et al. (1993)</td>
<td>12</td>
<td>5–11</td>
<td>96</td>
<td>80</td>
<td>0.64–6.4</td>
<td>5.8</td>
</tr>
<tr>
<td>Weiner et al. (1993)</td>
<td>5</td>
<td>3–10</td>
<td>29</td>
<td>70</td>
<td>0.4–4</td>
<td>1.2</td>
</tr>
<tr>
<td>Nakazawa et al. (1994)</td>
<td>11</td>
<td>5</td>
<td>55</td>
<td>70</td>
<td>0.35–3.5</td>
<td>3.4</td>
</tr>
<tr>
<td>Sekiya et al. (1994)</td>
<td>4</td>
<td>9–11</td>
<td>39</td>
<td>70</td>
<td>0.69–6.9</td>
<td>2</td>
</tr>
<tr>
<td>Gretch et al. (1996)</td>
<td>5</td>
<td>10–18</td>
<td>73</td>
<td>70</td>
<td>2.4–24</td>
<td>4.4</td>
</tr>
</tbody>
</table>

* The expected number of sporadic amino acid substitutions is calculated as the error rate of Taq (0.2–2 × 10⁻⁴) × length of HVR (75 nucleotides) × number of PCR cycles × number of clones sequenced × proportion of sporadics expected to produce amino acid substitutions (2/3)/number of samples.
† Mean number of sporadic amino acid substitutions amongst the set of sequences obtained for individual samples.
‡ This sequence data covers 58 aa (174 nt).

Do ‘quasispecies’ explain features of virus natural history?

Although it appears that many studies of virus heterogeneity have included some artefactual substitutions, there is no doubt that virus populations within infected individuals can be heterogeneous. For HCV, sporadic substitutions were observed amongst directly sequenced products at a frequency higher than expected from errors during reverse transcription (Fig. 1), and there are numerous examples of segregating polymorphisms that are unlikely to represent artefacts. Virus populations are therefore not homogeneous but consist of variants that may differ from the population average at one or more positions, and so could be described as forming a ‘quasispecies distribution’.

In passing, it should be noted that this description is not quantitative, since no distinction is made between sites of extreme variation, such as the hypervariable regions of the envelope proteins of HCV and HIV, and the relatively infrequent variation observed elsewhere in their genomes. Taking this to an extreme, mutations occur and accumulate in all organisms, and so in a trivial sense all species could be considered as ‘quasispecies’ since they are comprised of individuals that deviate to some degree from the population average. A more useful definition would include a measurement of heterogeneity per site, or a parameter describing the shape of the frequency distribution of distinct variants.

A recurring difficulty with some explanations relating the existence of virus quasispecies to aspects of virus natural history is that a causal relationship is often doubtful. For example, current dogma relates the persistent nature of infection with hepatitis C virus to the existence of virus quasispecies and the emergence of antigenic variants driven by immune selection. The quasispecies distribution is attributed to the observation that RNA polymerases lack a proof-reading activity. However, the same argument should apply to all RNA viruses, and yet only a minority of RNA viruses establish persistent infections. Even viruses closely related to HCV such as pestiviruses and flaviviruses that have similar amino acid sequences, genome organizations and virion structures do not generally produce persistent infections, nor do other RNA viruses that also infect the liver (hepatitis A and E). Hence, the persistence of HCV infections must result from something other than the error-prone replication of the RNA-dependent RNA polymerase of HCV. Indeed, it may be the persistence of HCV that provides the opportunity for divergence within an infected individual and the existence of variants could be a consequence of persistence rather than a cause.

A similar flaw in logic occurs where the absence of proof-reading of RNA polymerases is used to explain the existence of hypervariable regions within the genomes of HIV and HCV.
Hypervariable regions could be produced if the rate of mutation was greater in these regions, or if selective pressures were relaxed or accentuated. The basal error rate of the virus polymerase is of secondary importance in all of these explanations.

Finally, several features of RNA viruses (antigenic variation, genotypic diversity, immune escape) have been attributed to the lack of proof-reading of RNA-dependent RNA polymerases. However, the mutation rate of DNA viruses does not seem to be significantly different from that of RNA viruses since the frequency of variants resistant to monoclonal antibodies or bearing defined nucleotide substitutions is similar between the two classes (Smith & Inglis, 1987). If there are generic differences between RNA and DNA viruses, these may actually reflect differences in the periodicity of virus replication or the nature of selective forces rather than a difference in the rate of mutation. As an example, the genomes of some DNA viruses are larger than those of any RNA viruses, giving them the potential for greater complexity in their virion structure and in their interactions with the host cell.

The close relation between virus natural history and the extent of virus diversity can also be seen from comparisons between pairs of human viruses that are closely related but display very different patterns of diversity. Whereas the average difference between env sequences from different subtypes of HIV-1 is about 30%, and longitudinal studies suggest a rate of change $0.5-1%$ per year (Balfe et al., 1990), the most divergent and geographically distant isolates of human T-lymphotropic virus type I (HTLV-I) differ by only 7% (Bastian et al., 1993). Both HIV-1 and HTLV-I are retroviruses with similar genome organizations whose replication is dependent upon a virus-encoded reverse transcriptase. The error rate of the HIV-1 reverse transcriptase ($3 \times 10^{-5}$) is not dramatically different from that of other retroviruses (Mansky & Temin, 1995), and so the lower rate of substitution observed for HTLV-I may therefore reflect a reduced number of replication cycles per year and/or different selective constraints and mode of transmission, rather than a higher fidelity of replication.

Another example is provided by HCV and the flavivirus most closely related to it, hepatitis G virus (HGV/GBV-C). HCV types 1–6 differ by more than 30% over their complete genome sequences while subtypes, of which more than 70 are currently known from different parts of the world, differ by more than 20%. In contrast, five different isolates of HGV/GBV-C from North America, Africa and China differ by less than 14% and show a relatively strong bias against non-synonymous substitution (Smith et al., 1997b), apparently lacking a hypervariable region corresponding to that at the N terminus of the HCV E2 gene.

In neither of these examples is there evidence that the observation of extreme variability in a virus population is the result of an unusually error-prone virus replica. Instead, it may be virus natural history that determines the speed with which substitutions spread through a virus population, since this is much more dependent on the strength and direction of selection than on the underlying rate of mutation (Coffin, 1995). The extent of heterogeneity observed within an infected individual may therefore be a consequence of virus natural history, and the idea that virus heterogeneity can account for contrasting features of different viruses may be an example of putting the cart before the horse.

References


Higashi, Y., Kakumu, S., Yoshioka, K., Wakita, T., Mizokami, M., Ohba,


