Evolution of human polyomavirus JC

John N. Hatwell† and Paul M. Sharp

Institute of Genetics, University of Nottingham, Queen’s Medical Centre, Nottingham NG7 2UH, UK

More than 20 near full-length genome sequences have been reported for human polyomavirus JC (JCV). These have previously been classified into seven genotypes, and additional subtypes, which exhibit geographical associations. One of these genotypes, Type 4, has been suggested to be a recombinant of Types 1 and 3. We have investigated the pattern of diversity, and evolutionary relationships, among these sequences. In direct contradiction of a recent report, we found that different phylogenetic methods gave consistent results for the phylogenetic relationships among strains. The single known strain representing Type 5 was shown to be a mosaic of sequences from Types 2 and 6, although whether this recombination occurred in vivo or in vitro is not clear. In contrast, there was no substantial evidence that Type 4 strains are recombinant; rather they seem to be simply divergent examples of Type 1. On the assumption that the major genotypes of JCV diverged with human populations, the rate of synonymous nucleotide substitution was estimated to be around $4 \times 10^{-7}$ per site per year, about 10 times higher than a previous estimate for primate polyomaviruses.

Introduction

JC virus (JCV) is a polyomavirus that infects a large fraction of the human population (Shah, 1990). JCV infection appears harmless in the majority of individuals, but can cause the lethal disease progressive multifocal leukoencephalopathy (PML) in immunocompromised individuals (Shah, 1990). JCV was first described in 1971 (Padgett et al., 1971), but has attracted more attention since the AIDS epidemic became apparent. Several studies have aimed to assess the global diversity among JCV strains, and to understand the evolutionary relationships among strains infecting individuals from different populations. It has been concluded that there are several distinct clades within the phylogeny of JCV (referred to as ‘Types’), and that these Types have distinctive geographical distributions (Guo et al., 1996; Agostini et al., 1997b). Using the nomenclature of Jobes et al. (1998), with that of Sugimoto et al. (1997) in parentheses, JCV Type 1 (Type A subtype EU) is found in Europe, Type 2 (Type B subtypes B1 and MY) is found in Asia, and Types 3 (Type B subtype Af2) and 6 (Type C subtype Af1) are found in Africa. It seems that JCV is normally transmitted only as a result of prolonged exposure. For example, North Americans are generally infected with viruses reflecting their ethnic origin, and native Americans have Type 2 viruses, perhaps as result of the initial invasion of the Americas by people who crossed the Bering straits. Thus it has been suggested that evolutionary analyses of JCV strains, as well as providing insights into the biology of the virus, may also shed light on prehistoric human migrations (Sugimoto et al., 1997; Agostini et al., 1997b).

JCV has a circular double-stranded DNA genome just over 5 kb long. Until recently, the only full-length genome sequences available were those for the prototypic (Type 1) strain Mad-1 (Frisque et al., 1984), and GS/B and GS/K (Loeb & Dörries, 1988), the latter being two very similar Type 2 strains from a single individual. Most diversity studies have focused on a particular short region of the genome, about 610 bp in length, encompassing the convergent 3′ ends of the VP1 and large T antigen genes, and 69 bp of intergenic sequence. However, recently H. T. Agostini, G. L. Stoner and colleagues have determined full-length genome sequences from a variety of strains (Agostini et al., 1996, 1997a, 1998a, b), and used these to investigate the phylogeny of JCV. Their most recent analysis, of 22 complete JCV genomes, resulted in some very surprising conclusions (Jobes et al., 1998). Three different phylogenetic algorithms were used: the unweighted
pair group method with arithmetic means (UPGMA), the neighbour-joining method (NJ) and the maximum parsimony method (MP). Apparently very different results were obtained from the three methods. For example, a clade comprising JCV Type 1 and Type 4 strains was found (i) to be the earliest diverging lineage using UPGMA, (ii) to be the most recently derived lineage using NJ, and (iii) to span the root of the tree, so that all other strains were derived from within Type 1, using MP. While it is well known that alternative methods can produce different evolutionary trees (Li, 1997), this degree of discrepancy among results seems unprecedented, and potentially undermines any conclusions that might be drawn from phylogenetic studies of JCV.

Another apparent inconsistency among the results reported by Jobes et al. (1998) concerned the phylogenetic position of strain 402 (representing Type 4). It was stated that the MP method separated 402 from the Type 1 clade, whereas NJ and UPGMA did not. Since it had previously been suggested that strain 402 is most likely a recombinant of Types 1 and 3 (Agostini et al., 1996), Jobes et al. (1998) concluded that the MP method was effective in detecting this mosaicism, whereas the UPGMA and NJ methods were not. However, the phylogenetic analyses were performed on a single complete genome sequence alignment, and none of these three methods are expected to be able to distinguish recombinant sequences because they take no account of the relative positions of variable sites within the alignment. Indeed, if it is suspected that a dataset contains mosaic sequences, phylogenetic analyses should be performed using nonmosaic subregions from the alignment.

In the light of these curious results, we have analysed the diversity and evolutionary relationships among full-length JCV sequences. A variety of methods was used to search for any recombinant sequences. These revealed one, previously unidentified, strain (X01) that is clearly a mosaic of sequences from different major Types, but we found no substantive evidence suggesting that strain 402 is a recombinant. Importantly, our results on the phylogeny of JCV strains directly contradict those of the earlier study (Jobes et al., 1998) in many respects.

Methods

**JCV sequences.** All near full-length JCV genome sequences were obtained from the GenBank/EMBL/DDBJ database (GenBank release 112) using the ACNUC retrieval software (Gouy et al., 1985). Among the 22 sequences analysed previously (Jobes et al., 1998) JCV strains 309 and 310 were not included in the present study because they have not been deposited in the database; those sequences are extremely similar to strain 308 (Agostini et al., 1997a) and their omission is most unlikely to have affected our results. Strain GS/K was not used in the previous study, but was included here. The accession numbers for all sequences were given in Jobes et al. (1998), except for GS/B (AF004350) and GS/K (AF004349).

**Comparative analyses.** DNA sequences were aligned using CLUSTAL W (Thompson et al., 1994). The extent of variation among sequences was examined in diversity plots. The observed sequence difference, between one sequence and a variety of others, was calculated for windows of 500 sites, moved in steps of 50. Numbers of nonsynonymous and synonymous nucleotide substitutions per site, with correction for multiple hits, were estimated by the method of Li (1993).

Phylogenetic relationships among nucleotide sequences were estimated by the unweighted pair group method with arithmetic means (UPGMA), and the neighbour-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML) methods (see Li, 1997). The UPGMA and NJ methods were applied to distances between pairs of sequences estimated by Kimura’s 2-parameter method (Kimura, 1980). The NJ method (Saitou & Nei, 1987) was implemented using CLUSTAL W. To assess the reliability of branching orders within the phylogenetic trees obtained, 10000 bootstrap replicates (Felsenstein, 1985) were performed. The UPGMA, MP and maximum likelihood methods were implemented using the NEIGHBOR, DNAPARS and DNAML programs from the PHYLIP package (Felsenstein, 1992).

**Recombinant analyses.** A number of approaches were used to identify any putative recombinant sequences, and localize breakpoints within them. First, NJ phylogenetic analyses were performed for nine segments of 1000 sites, moved in steps of 500 along the alignment. Sequences with significantly discordant positions in different phylogenies are potential recombinants. Second, a ‘phylogenetic profile’ was calculated and plotted using the algorithm of Weiller (1998). In this method, the distances from one sequence to each of the others are calculated for windows along the alignment, and a correlation coefficient is calculated between the vectors of distances for two adjacent windows. This is performed for numerous pairs of adjacent windows, and the values plotted against position in the alignment. Here a window size of 100 bp was used, moved in steps of 1 bp. Similar profiles are calculated for each sequence, and can all be plotted on a single diagram. Unusually low correlation coefficients identify potential recombinant sequences.

Third, after putative recombinant sequences had been identified, a ‘compatibility matrix’ (Jakobsen & Eastal, 1996) was used to examine the distribution of phylogenetically informative sites supporting the placement of a mosaic sequence within alternative clades. Informative sites are plotted along the two axes of the matrix, and the cells within the matrix are coloured white if the two sites concur with respect to the phylogenetic partition among the sequences, or black if they do not. Blocks of sequence with discordant phylogenetic histories, implying recombination, are then readily visually identified. Finally, four sequence informative site breakpoint analysis was performed using the maximum chi-square approach (Robertson et al., 1995). The heterogeneity between the distribution of informative sites supporting two alternative phylogenies, on either side of a breakpoint within the alignment, was assessed by a 2 × 2 chi-square. This was calculated for all possible breakpoints along the alignment, and that yielding the highest chi-square value was retained. The significance of the chi-square values was assessed by permutation tests in which the same set of informative sites were randomly shuffled.

Results

**JCV genome sequence diversity.** Twenty-one (near) full-length JCV genome sequences were obtained from the DNA sequence database. The circular genome contains a nonprotein-coding regulatory region around the origin of replication that appears to be hypervariable within infected individuals due to rearrangements (Loeber & Dörries, 1988). Excluding that region (which,
anyway, is unavailable for 13 of the sequences) yielded an alignment of 4856 sites starting at the initiation codon of the agnoprotein gene (Fig. 1). Only three gaps were required in the alignment: at each of two sites a single sequence contained one extra nucleotide, and at a third site two (closely related) sequences shared a single nucleotide deletion. All three sites were in untranslated regions. The overall extent of nucleotide sequence divergence ranged from 1/4854 (= 0.02%) between strains 225 and 226 to 2.60% between strains 224 and 402.

The pattern of nucleotide sequence divergence in different regions of the genome was examined using diversity plots comparing, for example, the GS/B (Type 2) sequence to each of seven other strains representing various Types (Fig. 1). The more divergent sequences, e.g. GS/B versus Mad-1 (Type 1) or 601 (Type 6), differed at about 2% of sites across most of the genome. In one region, covering the 3' half of the VP1 gene plus the intergenic region between VP1 and the large T antigen gene, divergence increased to about 4%. In other regions, notably at the ends of the alignment, and where the VP2 and VP1 genes overlap in different reading frames, divergence decreased to about 1%. In general, the plots with different overall levels of divergence moved in parallel, and plots using other sequences were similarly broad. These results suggest that the entire genome sequence is useful for phylogenetic studies, but also confirm that the region around site 2100 (Fig. 2). The exceptional regions were the ends of the alignment, where (as expected; Weiller, 1998), the values for all sequences were low. The region with low correlations at the left end was quite extended, due to low sequence divergence in that area (Fig. 1). The exceptional sequence was strain X01, where the plot dipped to very low values in the region around site 2100 (Fig. 2). Plots for other sequences did not dip in this region, strongly suggesting that the X01 sequence has a recombinant origin.

The mosaic nature of X01 was also evident in diversity plots (Fig. 1). Across the first half of the alignment, strains GS/B and X01 exhibited a level of divergence typical of sequences belonging to different Types, but in the second half that divergence dropped to zero (Fig. 1). Closer inspection revealed that X01 is identical to GS/B from site 2247 to the end of the alignment (site 4856). In the region 1–2246 X01 and GS/B differed at 36 sites (1.60%), and X01 was most similar to strain 601. In the region 1–2246 X01 and 601 differed at only 10 sites (0.45%), whereas across the region 2247–4856 they differed at 55 sites (2.11%). Thus the X01 sequence appears to be a mosaic of Type 2 (GS/B-like) and Type 6 (601-like) sequences. This was also apparent from phylogenetic trees based on different segments of the alignment (not shown).

The breakpoint for this recombination was mapped by examining the distribution of phylogenetically informative sites in a four sequence alignment containing the putative recombinant, representatives of the two ‘parental’ lineages (i.e. GS/B and 601) and an outgroup (Robertson et al., 1995); here we used Mad-1 (Type 1). A compatibility matrix (Jakobsen & Easteal, 1996) clearly illustrated the nonrandom distribution of sites placing X01 with GS/B or with 601 (Fig. 3). Among the 40 informative sites, 16 adjacent sites in one half of the genome (from 670 to 2087) were all mutually compatible, as were 20 adjacent sites in the other half (from 2249 to 4765); however, these two blocks were mutually incompatible. Maximum chi-square analysis (Robertson et al., 1995) placed the most likely breakpoint between sites 2087 (where X01 and 601 are both G, while all other sequences are T) and 2111 (where X01 and all Type 2 sequences are C, while 601 and all Type 1 sequences are T); the distribution of informative sites around this breakpoint was extremely nonrandom (P < 10^-3). This location leaves one contradictory informative site (at 2246) within the 3' region, whereas if the breakpoint were located after site 2246 there would be two contradictory informative sites within the immediate 5' region (Fig. 3). There is also a single contradictory informative site (182) at the extreme 5' end of the alignment. Recalling that the genome is circular, there must be two breakpoints in X01; the second could be between sites 182 and 670, or within the (unavailable) noncoding region linking the two ends of the alignment.

Demonstration of recombination requires that different regions of a sequence be shown to have significantly different phylogenetic histories. Therefore, phylogenetic analyses were performed separately for the genomic regions 1–2100 and 2101–4856. The only significant difference between the trees concerned the position of X01 (Fig. 4). As expected, in the first region X01 was closely related to 601, whereas in the second X01 clustered with GS/B, in both cases with very high bootstrap values indicating strong statistical support for the alternative phylogenetic histories. The topology of trees produced by the maximum likelihood method differed from these NJ trees (Fig. 4) only with respect to the order of branching among Type 2 strains, and showed exactly the same positions for X01.
In contrast, we could find no substantial evidence that strain 402 is mosaic. Strain 402 is closely related to Type 1 strains (Fig. 4), but is thought to contain a fragment of Type 3-like sequence within the VP1 gene (Agostini et al., 1996; Jobes et al., 1998). There was no signal that 402 is mosaic in the phylogenetic profile (Fig. 2), even after strain X01 was removed from the comparisons (not shown). The suggestion that 402 is mosaic was made on the basis of a cluster of four sites within the VP1 gene where 402 contains nucleotides considered to be diagnostic of Type 3 rather than Type 1 (Fig. 2 in Agostini et al., 1996). The variable sites in this region are shown in Fig. 5; the four sites previously identified are numbered 1543, 1594, 1595 and 1684. It is evident that, at site 1543, the C in strain 402 is shared by all other sequences except Type 1, suggesting a C-to-G substitution during the recent origin of Type 1. At site 1595, the A seen in strain 402 is shared by two of the Type 3 strains, and all of the Type 2 and 7 strains, but not by another Type 3 strain (311), suggesting that there have been at least two substitutions at this site during the divergence of JCV strains. Similarly, the C at site 1684 in strain 402 is shared not only by Type 3 strains but also by 601 and X01, again suggesting that there have been at least two substitutions. Only one of the four sites (C at 1594) is uniquely shared by 402 and the Type 3 strains, and the simplest explanation appears to be that this similarity was also caused by parallel mutations, and not by recombination. In a phylogenetic analysis based on the complete genome excluding this region (1543–1684) strain 402 occupied the same position, close to but clearly outside the Type 1 cluster (not shown). Finally, there was no strong evidence that any of the other sequences in the alignment are mosaic.

**Phylogeny of JC virus genomes**

The phylogenetic relationships among 20 JCV sequences, i.e. excluding the mosaic sequence X01, were investigated. Phylogenetic algorithms generally produce unrooted trees and require either knowledge that one strain or clade represents an outgroup, or an assumption of approximately equal rates of evolution across the tree, in order to locate the position of the root (the most ancestral point) of the tree. The nearest known outgroup for JCV is BK virus (BKV), another human polyomavirus. However, BKV nucleotide sequences differ from JCV by more than 20%, compared to the maximum difference
Evolution of human polyomavirus JC

Fig. 2. Phylogenetic profiles of the JCV sequences, using the method of Weiller (1998). The plots show the magnitude of the correlation between vectors of distances (from each sequence to all others) from adjacent windows of 100 sites. The profiles for strains X01 and 402 are highlighted in red and turquoise. The positions of coding sequences are indicated below the plot.

within JCV of 2.6% noted above. This indicates that BKV sequences are too divergent to be used effectively as an outgroup to root the JCV tree. Therefore, we rooted JCV trees at their midpoint, in effect assuming that the various strains have diverged from their common ancestor at roughly similar rates.

The phylogeny obtained by NJ analysis is shown in Fig. 6. Within this tree the major Types represented by multiple strains (i.e. Types 1, 2 and 3) each formed monophyletic groups supported by high bootstrap values. ML analysis gave an identical topology. MP analysis produced four equally parsimonious trees, differing from each other and from the NJ tree only with respect to the branching order within the Type 2A clade. Thus, these three approaches yielded consistent results, since the branching order within Type 2A was not clearly resolved by any method. In contrast, the UPGMA tree differed more substantially, in clustering strain 230 with the Type 2B strains rather than the Type 2A strains. The UPGMA method assumes a constant molecular clock, and this can lead to errors in phylogenetic reconstruction when evolutionary rates vary among lineages. Therefore, the UPGMA tree is probably the least reliable, although it should be noted that this difference between the UPGMA results and those from the other three methods was minor compared to the apparent variation reported by Jobes et al. (1998).

The trees obtained here closely resemble that previously produced by UPGMA analysis by Jobes et al. (1998). At first glance, the NJ phylogeny obtained here (Fig. 6) seems quite different from that obtained using the same algorithm, and an almost identical set of sequences, by Jobes et al. (1998; their Fig. 2A). However, on closer inspection it can be seen that the topologies of the two trees are very similar, and it is largely in the position of the root that the two trees differ. For reasons that were not given, Jobes et al. rooted their NJ tree at the common ancestor of Type 2A, at the position marked by the lower circle in our Fig. 6. Similarly, the phylogenies obtained here seem quite different from the tree obtained by MP analysis by Jobes et al. (1998; their Fig. 2C, D). Again, however, this can be seen to be result of the position in which those authors rooted their MP tree, within the Type 1 clade of the common ancestor of Mad-1 and 124 (marked by the upper circle in our Fig. 6). While the precise position of the root of the JCV tree is not known, it seems most likely that it lies somewhere on a branch between the major Type clades. There
are no grounds for placing the root within any of the major clades, and clearly no reason for placing the root at very different positions in the results obtained from different algorithms.

Apart from the different locations of the root, our NJ tree differed from that of Jobes et al. (1998) with respect to the position of the Type 2B clade. Whereas all Type 2 strains formed a clear clade (supported in 92% of bootstraps) in Fig. 6, the previous analysis (after re-positioning the root) placed Type 2B strains outside a clade comprising the other Type 2 strains, Type 3 strains and Tai-3. We repeated the NJ analysis including X01, and found a similar result to that of Jobes et al. (1998), indicating that their anomalous position for the Type 2B lineage was a consequence of including the mosaic X01 sequence in the analysis. Thus, the inclusion of a recombinant sequence not only gave a false impression of the phylogenetic position of that sequence, but also distorted the positions of other sequences.

### Nucleotide substitution rates among JCV strains

To characterize the rate and pattern of divergence among JCV strains in more detail, we estimated separately the numbers of synonymous (Ks) and nonsynonymous (Ka) substitutions per site for each of the six genes, compared among various strains (Table 1). Normally, synonymous substitutions are expected to be effectively neutral, and so reflect the underlying rate of mutation. Then Ks values are expected to be similar among different genes, and the ratio of Ka/Ks reflects the action of natural selection on nonsynonymous mutations. For JCV the Ka/Ks ratio for the agnoprotein gene was unusually high, but this was due to low Ks values rather than high Ka values. This may indicate that silent sites within the agnoprotein gene are under additional constraint. It has been suggested that the homologous region in the polyomavirus simian virus (SV)40 contains attenuator sequences (Goldring et al., 1992). However, there may also be a sampling effect here, as the agnogene is very short; similar sized regions with no synonymous variations among strains were seen within other genes. The very low Ka/Ks values for the genes encoding VP1 and the T antigens imply strong constraint on the sequences of those proteins.

The mean Ks between strains from different major Types, was calculated as the average across Agno, Vp1, Vp2 and LTAg, weighting for the number of sites in each gene; Vp3 and StAg were excluded because of their substantial overlap with Vp2 and LTAg, respectively. The value obtained was around 0.08 substitutions per synonymous site. It has been suggested that the closely related polyomaviruses BKV (from humans) and SV40 (from Old World monkeys) co-evolved with their hosts (Soeda et al., 1980). This allowed Yasunaga & Miyata...
Evolution of human polyomavirus JC

Fig. 4. Phylogenetic relationships among 21 JCV sequences for (A) sites 1–2100, (B) sites 2101–4856. Numerical values within the tree indicate the percentage of 1000 bootstrap replicates in which the clade to the right appeared; only values greater than 70% are shown. Type designations for the strains are indicated at the right; strains 601 and Tai-3 are the single representatives of Types 6 and 7, respectively.

Fig. 5. Variable sites within a fragment of the VP1 gene. Sites identified by Agostini et al. (1996) as being Type 3-like in the 402 sequence are indicated by * above the alignment. Within this region sequences are identical for strains 308 and 312, for strains 225 and 226, for strains 228 and 229, for strains GS/B, GS/K, 223 and 227, and for strains 601 and X01.

(1982) to estimate the rate of synonymous substitution in primate polyomaviruses as $3.8 \times 10^{-8}$ substitutions per site per year. If JCV has been evolving at that rate, the divergence between the major Types would be estimated to have occurred around 1 Myr ago. This is a surprising result, because it seems that divergences among human populations have occurred within the last 200,000 years (von Haeseler et al., 1996). On the other hand, if it is assumed that the major Types of JCV split with their human hosts, and if we take the emergence of Homo sapiens sapiens from Africa at roughly 100,000 years ago, then the rate of synonymous substitution in JCV would be estimated as $4 \times 10^{-7}$ synonymous substitutions per site per year.
Table 1. Nucleotide substitution rates among JCV types

For the six genes, the estimated numbers of synonymous (K_s), nonsynonymous (K_a) and overall average (K) nucleotide substitutions per site, the ratio of K_a/K_s and the number of codons compared (L), are presented as averages of comparisons among three Type 1 strains, five Type 2A strains and the Type 6 strain.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Type 1 vs Type 2A (n = 15)</th>
<th>Type 1 vs Type 6 (n = 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>K</td>
</tr>
<tr>
<td>Agno</td>
<td>71</td>
<td>0.014</td>
</tr>
<tr>
<td>Vp2</td>
<td>344</td>
<td>0.020</td>
</tr>
<tr>
<td>Vp3</td>
<td>225</td>
<td>0.028</td>
</tr>
<tr>
<td>Vp1</td>
<td>354</td>
<td>0.033</td>
</tr>
<tr>
<td>LTAg</td>
<td>688</td>
<td>0.026</td>
</tr>
<tr>
<td>StAg</td>
<td>172</td>
<td>0.016</td>
</tr>
</tbody>
</table>

Fig. 6. Phylogenetic relationships among complete genome sequences of 20 nonrecombinant JCV strains. The two nodes that have been circled are referred to in the text. See Fig. 4 for more details.

Discussion

A recent phylogenetic analysis of full-length JCV sequences (Jobes et al., 1998) reached some extraordinary conclusions. Those authors found, and discussed at some length, striking differences among phylogenetic trees obtained using three different methods of analysis. In absolute contrast, we have used four different phylogenetic algorithms (including the three used previously) and found that all methods yielded very similar results. In retrospect, it can be seen that the differences reported by Jobes et al. (1998) arose entirely as an artefact of placing the root of each tree in a different location. Unless otherwise directed, programs producing NJ and MP trees necessarily choose some arbitrary position (often dictated by the input order of sequences) for a root in order to display the results. Among the trees reported by Jobes et al. (1998), that from UPGMA analysis is the most similar to our results, simply because UPGMA assumes a constant molecular clock and so automatically places the root of the tree at its midpoint. However, this should not be taken as a recommendation of the UPGMA method: its clock assumption is often violated, and the method can then give erroneous results. Indeed, we found that while three methods gave very similar results, UPGMA yielded a slightly different phylogeny, probably due to rate differences among virus lineages.

It was suggested by Jobes et al. (1998) that one reason for the differences among the results of different analyses was because one sequence (strain 402) was a recombinant, and that one method (MP) was more useful than others in detecting that mosaicism. However, we found no substantial evidence that strain 402 is recombinant. Again, the apparent differences in the position of strain 402 were entirely due to the different rootings of the trees from different methods. Furthermore, none of the methods previously employed would be expected to distinguish recombinant sequences.

We used a variety of methods to detect recombinant viruses, and found that one strain, X01, has a clearly mosaic sequence. Jobes et al. (1998) renamed that strain as 501, because they considered it to be the prototype of a new genotype, Type 5. Our results indicate that strain X01 is a mosaic of sequences from Types 2 and 6, and that therefore there is no basis for (this) Type 5, and in light of this it seems more appropriate to retain the label X01 for this strain. Since this was the only recombinant strain detected, the question arises as to where and how the mosaic sequence of X01 arose. Almost half
of the X01 genome sequence is identical to a sequence (GS/B) from Germany (Loeber & Dörries, 1988), while the other half is most similar to strain 601 from an African American (Jobes et al., 1998); strain X01 was obtained from an American with European parents (Agostini et al., 1998b). The identity between part of the X01 sequence and GS/B, compared to a difference (albeit small) between two sequences (GS/B and GS/K) obtained from a single PML patient, might point to X01 being a mosaic sequence arising from some laboratory artefact.

However, there are other instances of extremely closely related sequences from distinct origins, such as strains 228 and 229, from a native American in New Mexico, and a European American in California, respectively (Agostini et al., 1998b). Also, the relatively slow rate of evolution of JCV (see below) implies that identical sequences might be found among viruses that have been separated for thousands of years. However the X01 sequence was generated, it was important to identify it as mosaic because including it in the phylogenetic analysis distorted the phylogenetic positions of other sequences.

Our analysis of full-length genome sequences, excluding the recombinant sequence, should provide the best basis on which to classify strains into Types and subtypes. Three Types (1–3) represented by multiple strains were clearly distinguished, although Types 2 and 3 are much closer to each other than to Type 1 (Fig. 6). The single representative of Type 4, strain 402, is no more divergent from Type 1 strains than strains within Type 2 are from each other; for consistency, it seems that strain 402 is better classified within Type 1. The single representative of Type 6, strain 601, is highly divergent from other subtypes, and so warrants being placed in a separate genotype. Strain Tai-3 has been tentatively designated as Type 7 (Jobes et al., 1998). Here Tai-3 was found to lie outside the clade currently defined as Type 2, but even closer to Type 2 than is Type 3 (Fig. 6); it could be designated as a distinct Type, but (in the absence of Types 4 and 5) the numbering is moot. A number of Type 2 subtypes have been proposed (Agostini et al., 1998b; Jobes et al., 1998). There seems no good basis for separating strains 228 and 229 (previously termed Type 2C) from other Type 2A strains, but strain 230 (previously termed Type 2D) is phylogenetically distinct from Types 2A and 2B (Fig. 6).

It has been suggested that the various clades of JCV have co-evolved with human populations (Sugimoto et al., 1997; Agostini et al., 1997b, 1998a; Guo et al., 1998). If it can be assumed that the major genotypes of JCV diverged when human populations migrated, then the rate of synonymous substitution in JCV has been around 4 \times 10^{-7} substitutions per site per year. This is around ten times faster than previously suggested for primate polyomaviruses (Yasunaga & Miyata, 1982), possibly reflecting saturation of substitutions in the earlier analysis which compared much more divergent viruses (BKV and SV40). Our rate estimate is about four orders of magnitude lower than in viruses using RNA-dependent polymerases for replication, such as influenza viruses (Fitch et al., 1991) or the retrovirus HIV-1 (Li et al., 1988), but about two orders of magnitude higher than synonymous substitution rates in host (primate) nuclear genes (Li et al., 1987). This rate estimate for JCV also appears to be higher than that for another group of DNA viruses, the alphaherpesviruses. McGeoch et al. (1995) estimated the common ancestor of herpes simplex viruses 1 and 2 (HSV-1 and HSV-2) at 8.5 Myr ago, while Dolan et al. (1998) calculated the average Ks between HSV-1 and HSV-2 as 0.47 substitutions per site; that would imply a rate of 2.8 \times 10^{-8} substitutions per site per year. Thus DNA viruses appear to have comparatively slow evolutionary rates. Presumably the precise rate depends on the numbers of rounds of replication the viruses undergo, per unit time.

We are grateful to Liz Bailes for her assistance with computer analyses.

References


Received 16 November 1999; Accepted 31 January 2000