Journal of NeuroVirology, 7: 339–344, 2001
ISSN 1355–0284/01 $12.00+.00

Epidemiology Report

Predicted amino acid sequences for 100 JCV strains

Christopher L Cubitt,[1] Xiaohong Cui,[1] Hansjürgen T Agostini,[2] Vivek R Nerurkar,[3] Iris Scheirich,[3] Richard Yanagihara,[3] Caroline F Ryschkewitsch,[1] and Gerald L Stoner[1]

[1]Neurotoxicology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA; [2]Department of Ophthalmology, University of Freiburg, Freiburg, Germany; and [3]Retrovirology Research Laboratory, Pacific Biomedical Research Center, University of Hawaii at Manoa, Honolulu, Hawaii, USA

DNA sequence variation between JCV genotypes is confined largely to noncoding intergenic regions and introns. Nevertheless, evidence suggests that the amino acid sequence variations among the 8 genotypes of JCV can influence the potential for neurovirulence of the virus. In the current study, the amino acid sequences for 100 JCV genomes were translated and grouped into genotype families. Subtype consensus sequences were determined and the type-specific amino acid sequence variants were identified. *Journal of NeuroVirology* (2001) **7,** 339–344.

Keywords: amino acid sequence; genotype; JC virus; polyomavirus

JC virus (JCV) is a slowly replicating, small double-stranded DNA virus that is ubiquitous in the human population. JCV is the causative agent of progressive multifocal leukoencephalopathy (PML) in immunocompromised patients. For the diagnosis of PML, sequences from the JCV genome are amplified from CSF samples or brain biopsies from patients suspected of having the disease. Besides the pathological consequences of this virus, JCV-DNA is detectable in the urine of 20–80% of the human population and is increasingly used as a marker for anthropological studies, as the phylogenetics of JCV correlates with the known genetic and linguistic history of humans (Sugimoto *et al*, 1997; Stoner *et al*, 2000). JCV strains can be divided into at least 7 distinct genotypes based on phylogenetic analysis of partial genome sequences of 610 bp including the VT-intergenic region (Sugimoto *et al*, 1997) or full genome sequences (Jobes *et al*, 1998). The different genotypes of JCV correlate with human populations in different geographic regions of the world. Europeans and Americans of European descent typically have Type 1 or Type 4 JCV (Agostini *et al*, 2001). JCV Types 2A and 2B are Asian genotypes. Type 2A is also found in Native Americans, and Type 2B is a minor type in Eurasians. An additional JCV genotype found in Asian populations is Type 7. Africans have the Type 3 and 6 variants of JCV (Agostini *et al*, 1998; Chima *et al*, 2000), and African-Americans retain mainly the Type 3 varient (Chima *et al*, 2000).

Autopsy studies of AIDS patients have shown the frequency of PML varies from a low of 0.8% in Brazil to a high of 10% in Italy (Chimelli *et al*, 1992; Kuchelmeister *et al*, 1993). Recent epidemiological evidence suggests there are positive and negative correlations of JCV genotype with the incidence of PML in AIDS patients. In a retrospective study conducted in the United States, the prevalence of Type 2B was found to be significantly more frequent in PML brain than in the urine samples of a control cohort (Agostini *et al*, 2000). In a recent study conducted in France, the investigators identified a significant negative correlation of Type 4 with PML, indicating that Type 4 may have a lower propensity of causing PML in AIDS patients (Dubois *et al*, 2001). However, whether the correlation of JCV genotype and incidence of PML is due to a phenotypic difference in JCV subtypes, or is due to differences in the population in which the subtypes are endemic remains to be determined.

Whereas most DNA sequence polymorphisms between genotypes are silent mutations, others of these variations in DNA sequence result in protein

Address correspondence to Gerald L. Stoner, Ph.D., National Institutes of Health, Bldg. 36, Room 4A-27, MSC-4126, 36 Convent Drive, Bethesda, MD 20892-4126, USA. E-mail: stonerg@ninds.nih.gov

Received 7 March 2001; revised 18 March 2001; accepted 19 April 2001.

sequence differences. In the current study, the amino acid sequences for 100 JCV genomes were determined and separated according to genotype family. Amino-acid sequence variants were then tabulated.

Results

With the availability of DNA sequences for 100 JCV genomes, we first used this data to analyze the conservation of DNA sequence among all known JCV genotypes. Prior to a PILEUP alignment, the ~267-bp NCRR was removed from all sequences leaving only the coding region for subsequent analysis. The GCG program SIMILARITYPLOT was used to determine the running average of the similarity among all 100 DNA sequences using a sliding window length of 10 nucleotides. The average similarity of the aligned sequences was >90% similar across most of the coding region of the genome except for 3 areas (Figure 1A). Three areas with <90% average similarity among all genomes occurred in the V-T intergenic region, the large T intron between small T and exon 2 of large T, as well as the intergenic region between agnoprotein and VP2. The only coding region in which the average similarity score dropped below 90% was the 3′-region of the agnoprotein gene.

The predicted amino acid sequences were next derived from the 100 genome sequences. All amino acid sequences for the early JCV proteins were aligned and used to derive a consensus. The consensus sequences for large and small T antigens were 688 and 172 amino acids in length, respectively (Figure 1B). After grouping the amino acid sequences according to subtype, a consensus sequence for each subtype was determined. All subtype groups contained 3 or more members except for Types 2D1 and 3B, which contained 2 members. Amino acid positions that did not match the over all T/t antigen consensus have been compiled in Figure 1C. The alternately spliced T′ (T-prime) variants T'_{165}, T'_{136}, and T'_{135} would include the amino acid variant at position 29 (Trowbridge and Frisque, 1995). T'_{165} would additionally have the amino acid mutations found at positions 662, 667, and 670.

Amino acid sequences for the late (capsid) JCV proteins were aligned and the consensus sequence derived (Figure 2). The consensus amino acid (aa) lengths are 71 aa (agnoprotein), 354 aa (VP1), 344 aa (VP2), and 225 aa (VP3).

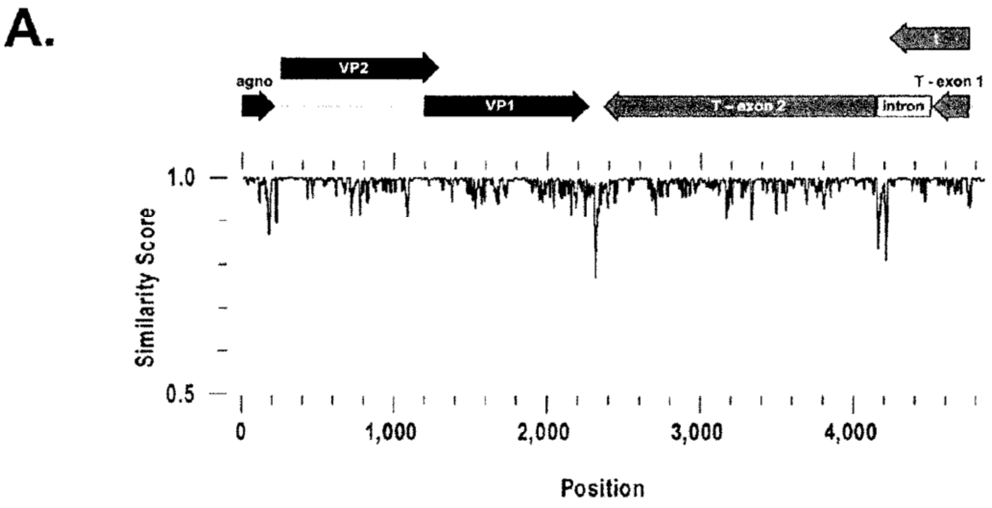
Discussion

Full genome sequencing of the JC virus substantially increases the number of phylogenetically informative sites and more adequately resolves relationships between the JCV genotypes. In addition, examination of sequence variations throughout the genome could explain potential biological and pathological differences between genotypes.

Similarity analysis of 100 JCV genomes without the regulatory region revealed that most sequence heterogeneity occurs in noncoding regions, whereas the coding regions exhibited greater than 90% similarity among sequences. Analysis of JCV sequence differences within coding regions and noncoding regions could offer clues to the mechanisms behind the observed pathogenic variations between genotypes. We have previously proposed type-specific differences in amino acid residues as a basis for the altered neurovirulence associated with JCV Type 2B and Type 4 (Agostini *et al*, 1999; Agostini *et al*, 2000). The hydrophilic amino acid Gln residue at position 301 of large T antigen (Figure 1C) differentiates Type 2B from other Type 2 subtypes and Types 3, 6, 7, and 8, which have a hydrophobic Leu at this position. Because position 301 is two residues before the first zinc-coordinating Cys residue (Cys_{303}), this change to a hydrophobic residue may alter the local environment of and affect the stability or function of the large T antigen zinc finger motif. Other sequences may also be involved, however, as Type 4 strains, which are less prevalent in PML brain, are not distinguished from Type 2B in the zinc-finger motif. In Type 4 strains, an amino acid sequence change at position 164 of the predicted VP1 gene is changed from the basic residue Lys to a neutral Thr as in Type 3, but at position 332, Type 3 has a neutral Gln and Type 4 retains the acidic Glu. The net result of these changes in this part of VP1 is a net charge change of −1 in Type 4 strains.

The accurate diagnosis of PML and characterization of the virus presents several challenges. Among these challenges are the risk of false negatives, false positives, and the risk of PCR-induced errors. Falsely negative PCR results can occur when PCR primers do not sufficiently anneal to regions of the genome due to polymorphisms in the target sequence. In the JCV genome, the V-T intergenic region and the noncoding region of the large T intron exhibited the highest level of polymorphism (Figure 1A). PCR detection of various JCV genotypes can be improved by the careful selection of PCR primers that are complementary to more conserved regions of the genome or by using degenerate primers representing more than one nucleotide at a typing site. Heterogeneity in the 3′-end of the template primer binding site may be of little consequence for primer annealing.

False detection of JCV sequence PCR signal from samples is frequently the result of contamination during handling of specimens. PCR of JCV is especially prone to extraneous signal because of the high level of PCR amplification needed to detect the low levels of viral DNA in CSF or latently infected tissue. The Mad-1/Mad-4 strains of JCV are frequently used in laboratories and because of this, JCV-DNAs amplified from clinical samples that contain the Mad-1 or Mad-4 type regulatory region should be carefully examined. To identify whether a strain of the Mad-1 regulatory region configuration is authentic original



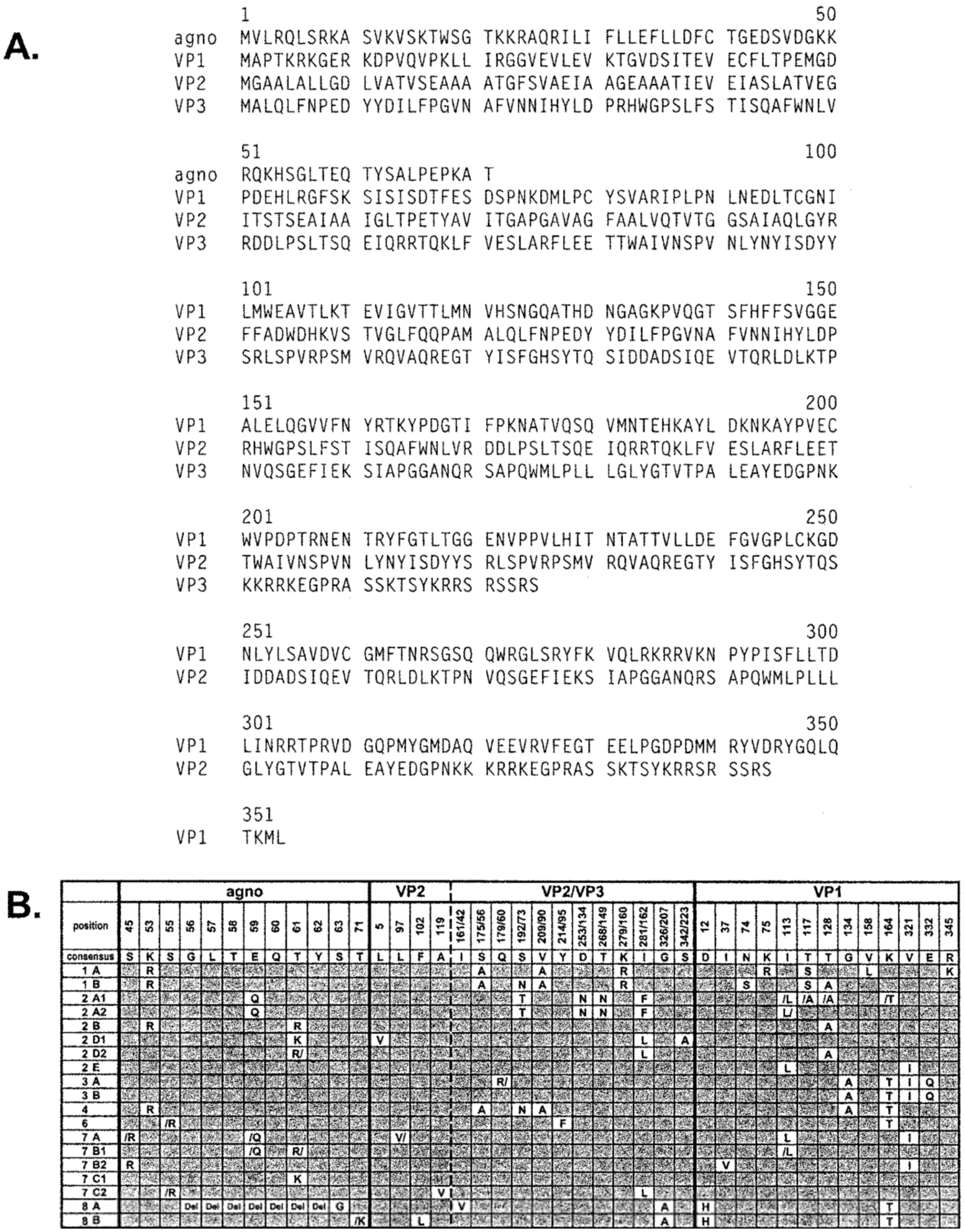
B.

```
      1                                                   50
T     MDKVLNREES MELMDLLGLD RSAWGNIPVM RKAYLKKCKE LHPDKGGDED
t     MDKVLNREES MELMDLLGLD RSAWGNIPVM RKAYLKKCKE LHPDKGGDED

      51                                                 100
T     KMKRMNFLYK KMEQGVKVAH QPDFGTWNSS EVPTYGTDEW ESWWNTFNEK
t     KMKRMNFLYK KMEQGVKVAH QPDFGTWNSS EVGCDFPPNS DTLYCKEWPN

      101                                                150
T     WDEDLFCHEE MFASDDENTG SQHSTPPKKK KKVEDPKDFP VDLHAFLSQA
t     CATNPSVHCP CLMCMLKLRH RNRKFLRSSP LVWIDCYCFD CFRQWFGCDL

      151                                                200
T     VFSNRTVASF AVYTTKEKAQ ILYKKLMEKY SVTFISRHGF GGHNILFFLT
t     TQEALHCWEK VLGDTPYRDL KL

      201                                                250
T     PHRHRVSAIN NYCQKLCTFS FLICKGVNKE YLFYSALCRQ PYAVVEESIQ

      251                                                300
T     GGLKEHDFNP EEPEETKQVS WKLVTQYALD TKCEDVFLLM GMYLDFQENP

      301                                                350
T     LQCKKCEKKD QPNHFNHHEK HYYNAQIFAD SKNQKSICQQ AVDTVAAKQR

      351                                                400
T     VDSIHMTREE MLVERFNFLL DKMDLIFGAH GNAVLEQYMA GVAWIHCLLP

      401                                                450
T     QMDTVIYEFL KCIVLNIPKK RYWLFKGPID SGKTTLAAAL LDLCGGKSLN

      451                                                500
T     VNMPLERLNF ELGVGIDQFM VVFEDVKGTG AESRDLPSGH GISNLDCLRD

      501                                                550
T     YLDGSVKVNL ERKHQNKRTQ VFPPGIVTMN EYSVPRTLQA RFVRQIDFRP

      551                                                600
T     KAYLRKSLSC SEYLLEKRIL QSGMTLLLLL IWFRPVADFA AAIHERIVQW

      601                                                650
T     KERLDLEISM YTFSTMKANV GMGRPILDFP REEDSEAEDS GHGSSTESQS

      651                                      688
T     QCSSQVSEAS GADTQEHCTY HICKGFQCFK KPKTPPPK
```

C.

	T																				t		
position	29	155	192	202	233	280	281	301	354	364	408	470	474	479	493	533	555	653	662	667	670	29	83	121
consensus	V	R	G	H	F	D	T	L	I	E	E	M	E	T	S	S	R	S	A	H	Y	V	G	R
1 A						E		Q			D							F		N	F			
1 B						E		Q			D									N	F			
2 A1										/Q						/A		F/						
2 A2																		F						
2 B								Q																
2 D1							I																	
2 D2																		F/						
2 E			S															C						
3 A																	/T						S	
3 B																								
4	I					E		Q			D							F		N		I		
6					L/	E								/S	N									K
7 A																								
7 B1									/L															
7 B2									L															
7 C1										Q			D											
7 C2		/K		/Q	L				L															
8 A														S					T					
8 B												I												

Figure 1 (**A**) Similarity plot of 100 JCV genome sequences. The similarity score is the average similarity among all aligned sequences at each position in the alignment, using a sliding window length of 10. Lower similarity scores denote more polymorphic regions of the genome. (**B**) Consensus sequences of all JCV early proteins large T (T) and small T (t) antigens. (**C**) Polymorphisms of JCV large and small T proteins arranged by type. Genotypes different from the consensus are indicated by an amino acid letter designation within a box. Amino acid positions with a forward slash (/) before the amino acid letter designation specifies that this amino acid represents 50% or less of the constituents in the type or subtype group at that position. A forward slash after the letter designation specifies that the amino acid is a major constituent of the subtype group for that position.



A.

```
     1                                                  50
agno MVLRQLSRKA SVKVSKTWSG TKKRAQRILI FLLEFLLDFC TGEDSVDGKK
VP1  MAPTKRKGER KDPVQVPKLL IRGGVEVLEV KTGVDSITEV ECFLTPEMGD
VP2  MGAALALLGD LVATVSEAAA ATGFSVAEIA AGEAAATIEV EIASLATVEG
VP3  MALQLFNPED YYDILFPGVN AFVNNIHYLD PRHWGPSLFS TISQAFWNLV

     51                                                100
agno RQKHSGLTEQ TYSALPEPKA T
VP1  PDEHLRGFSK SISISDTFES DSPNKDMLPC YSVARIPLPN LNEDLTCGNI
VP2  ITSTSEAIAA IGLTPETYAV ITGAPGAVAG FAALVQTVTG GSAIAQLGYR
VP3  RDDLPSLTSQ EIQRRTQKLF VESLARFLEE TTWAIVNSPV NLYNYISDYY

     101                                               150
VP1  LMWEAVTLKT EVIGVTTLMN VHSNGQATHD NGAGKPVQGT SFHFFSVGGE
VP2  FFADWDHKVS TVGLFQQPAM ALQLFNPEDY YDILFPGVNA FVNNIHYLDP
VP3  SRLSPVRPSM VRQVAQREGT YISFGHSYTQ SIDDADSIQE VTQRLDLKTP

     151                                               200
VP1  ALELQGVVFN YRTKYPDGTI FPKNATVQSQ VMNTEHKAYL DKNKAYPVEC
VP2  RHWGPSLFST ISQAFWNLVR DDLPSLTSQE IQRRTQKLFV ESLARFLEET
VP3  NVQSGEFIEK SIAPGGANQR SAPQWMLPLL LGLYGTVTPA LEAYEDGPNK

     201                                               250
VP1  WVPDPTRNEN TRYFGTLTGG ENVPPVLHIT NTATTVLLDE FGVGPLCKGD
VP2  TWAIVNSPVN LYNYISDYYS RLSPVRPSMV RQVAQREGTY ISFGHSYTQS
VP3  KKRRKEGPRA SSKTSYKRRS RSSRS

     251                                               300
VP1  NLYLSAVDVC GMFTNRSGSQ QWRGLSRYFK VQLRKRRVKN PYPISFLLTD
VP2  IDDADSIQEV TQRLDLKTPN VQSGEFIEKS IAPGGANQRS APQWMLPLLL

     301                                               350
VP1  LINRRTPRVD GQPMYGMDAQ VEEVRVFEGT EELPGDPDMM RYVDRYGQLQ
VP2  GLYGTVTPAL EAYEDGPNKK KRRKEGPRAS SKTSYKRRSR SSRS

     351
VP1  TKML
```

B.

	agno											VP2				VP2/VP3											VP1												
position	45	53	55	56	57	58	59	60	61	62	63	71	5	97	102	119	161/42	175/56	179/60	192/73	209/90	214/95	253/134	268/149	279/160	281/162	326/207	342/223	12	37	74	75	113	117	128	134	158	164	321	332	345
consensus	S	K	S	G	L	T	E	Q	T	Y	S	T	L	L	F	A	I	S	Q	S	V	Y	D	T	K	I	G	S	D	I	N	K	I	T	T	G	V	K	V	E	R
1 A		R																A			A				R							R		S			L				K
1 B		R																A		N	A				R						S			S	A						
2 A1							Q													T			N	N		F								/L	/A	/A			/T			
2 A2							Q													T			N	N		F								L/								
2 B		R							R																											A						
2 D1									K				V														L		A													
2 D2									R/																			L										A						
2 E																																			L						I		
3 A																			R/																				A		T	I	Q	
3 B																																							A		T	I	Q	
4		R																A		N	A																	A		T			
6			/R																				F																		T			
7 A	/R						/Q						V/																						L						I		
7 B1							/Q		R/																										/L								
7 B2	R																													V									I		
7 C1									K																																	
7 C2			/R												V										L															
8 A				Del	Del	Del	Del	Del	Del	Del	G							V										A		H									T			
8 B												/K			L													A		H									T			

Figure 2 (**A**) Consensus sequences of JCV late proteins: agnoprotein, VP1, VP2, and VP3. (**B**) Polymorphic sites of JCV agnoprotein, VP1, VP2, and VP3. Genotypes different from the consensus are indicated by an amino acid letter designation within a box. Amino acid positions with a forward slash (/) before the amino acid letter designation specifies that this amino acid represents 50% or less of the constituents in the type or subtype group at that position. A forward slash after the letter designation specifies that the amino acid is a major constituent of the subtype group for that position. The vertical dashed line indicates where VP3 begins to overlap with VP2 starting at position 120 of VP2 and continuing in the same frame until the termination of both proteins at position 344 of VP2.

Mad-1, 2 sites on the genome can be examined for rare mutations that differentiate Mad-1 from other Type 1A JCV strains. At position 2311 in the VP1 gene, Mad-1 has G rather than T, whereas at position 3134 in the T-antigen gene, Mad-1 has T rather than C. If, on the other hand, the Mad-1 type regulatory region is being generated anew, then the characteristic viral regulatory region should not have these unusual coding region polymorphisms.

Corruption of PCR amplified JCV sequences is another frequent challenge to be overcome. Random point mutations in PCR-amplified regulatory region products may reflect PCR-induced error, and should not be considered evidence for unique strains, particularly in cloned products. Commercially available DNA polymerases vary greatly in their fidelity (rate of misincorporation). Enzymes such as Pfu that have 3′-5′ exonuclease proofreading activity have 10-fold lower rate of misincorporation than Taq polymerases. It is interesting to note that polymorphisms in the JCV genome that result in truncations or deletions in viral proteins are exceedingly rare. To date, we have only observed 1 genotype (Figure 2B, Type 8A) and one JCV isolate within the Type 4 family with an amino acid deletion in the agnoprotein of urinary sequences (Jobes *et al*, 1999; Deckhut, unpublished data). JCV DNA sequences in two PML brains showed capsid protein (VP1) deletions (Stoner and Ryschkewitsch, 1995).

The characterization of JCV sequences promises to be beneficial for identifying conserved areas of the genome to help maximize the sensitivity of PCR diagnostics and for the exploration of the relationship of JCV strains and relative risk of PML. In addition, DNA sequence analysis of the amplified products followed by genotype assignment can help to distinguish between clinical strains and cloned laboratory contaminants.

Methods

Computer analyses of DNA and protein sequences were accomplished using the suite of programs available in version 10 of the Wisconsin GCG package (Genetics Computer Group, Pharmacopeia, Inc). One hundred JCV genomes were aligned using the PILEUP alignment algorithm. Next, the origin and noncoding regulatory region (NCRR) of the genomes were removed, leaving the protein-coding region of the genome (~4854 bp). The large T intron was removed from the sequences and the open-reading frames were then selected for translation resulting in the amino acid sequences for the major proteins encoded by the JCV genome: the agnoprotein, VP1-3, and large T and small T antigens.

The protein sequences were divided into types and subtypes and aligned using the PC program OMIGA (Genetics Computer Group, Pharmacopeia, Inc). An amino acid consensus sequence for each subtype was determined by examining the amino acids at each position (see Figure 1 legend). If a group contained only 1 member with a polymorphic amino acid, the variant amino acid was dropped and not considered in the consensus.

Acknowledgements

We would like to thank Joshua Hvidding for expert technical assistance.

References

Agostini HT, Deckhut AM, Jobes DV, Girones R, Schlunck G, Prost MG, Frias C, Pérez-Trallero E, Ryschkewitsch CF, Stoner GL (2001). Genotypes of JC virus in East, Central and Southwest Europe. *J Gen Virol* **82:** 1221–1231.

Agostini HT, Jobes DV, Chima SC, Ryschkewitsch CF, Stoner GL (1999). Natural and pathogenic variation in the JC virus genome. In: *Recent Research Development in Virology*. Pandalai SG (ed). Trivandrum, India: Transworld Research Network, pp 683–701.

Agostini HT, Ryschkewitsch CF, Baumhefner RW, Tourtellotte WW, Singer EJ, Komoly S, Stoner GL (2000). Influence of JC virus coding region genotype on risk of multiple sclerosis and progressive multifocal leukoencephalopathy. *J NeuroVirol* **6(Suppl 2):** S101–S108.

Agostini HT, Ryschkewitsch CF, Stoner GL (1998). Complete genome of a JC virus genotype Type 6 from the brain of an African American with progressive multifocal leukoencephalopathy. *J Hum Virol* **1:** 267–272.

Chima SC, Ryschkewitsch CF, Fan KJ, Stoner GL (2000). Polyomavirus JC genotypes in an urban United States population reflect the history of African origin and genetic admixture in modern African Americans. *Hum Biol* **72:** 837–850.

Chimelli L, Rosemberg S, Hahn MD, Lopes MBS, Barretto Netto M (1992). Pathology of the central nervous system in patients infected with the human immunodeficiency virus (HIV): A report of 252 autopsy cases from Brazil. *Neuropathol Appl Neurobiol* **18:** 478–488.

Dubois V, Moret H, Lafon ME, Brodard V, Icart J, Ruffault A, Guist'hau O, Buffet-Janvresse C, Abbed K, Dussaix E, Ingrand D (2001). JC virus genotypes in France: Molecular epidemiology and potential significance for progressive multifocal leukoencephalopathy. *J Infect Dis* **183:** 213–217.

Jobes DV, Chima SC, Ryschkewitsch CF, Stoner GL (1998). Phylogenetic analysis of 22 complete genomes of the human polyomavirus JC virus. *J Gen Virol* **79:** 2491–2498.

Jobes DV, Friedlaender JS, Mgone CS, Koki G, Alpers MP, Ryschkewitsch CF, Stoner GL (1999). A novel JC virus variant found in the Highlands of Papua New Guinea has a 21-base pair deletion in the agnoprotein gene. *J Hum Virol* **2:** 350–358.

Kuchelmeister K, Gullotta F, Bergman M, Angeli G, Masini T (1993). Progressive multifocal leukoencephalopathy (PML) in the acquired immunodeficiency syndrome (AIDS). A neuropathological study of 21 cases. *Pathol Res Pract* **189:** 163–173.

Stoner GL, Jobes DV, Fernandez Cobo M, Agostini HT, Chima SC, Ryschkewitsch CF (2000). JC virus as a marker of human migration to the Americas. *Microbe Infect* **2:** 1905–1911.

Stoner GL, Ryschkewitsch CF (1995). Capsid protein VP1 deletions in JC virus from two AIDS patients with progressive multifocal leukoencephalopathy. *J NeuroVirol* **1:** 189–194.

Sugimoto C, Kitamura T, Guo J, Al-Ahdal MN, Shchelkunov SN, Otova B, Ondrejka P, Chollet JY, El-Safi S, Ettayebi M, Grésenguet G, Kocagöz T, Chaiyarasamee S, Thant KZ, Thein S, Moe K, Kobayashi N, Taguchi F, Yogo Y (1997). Typing of urinary JC virus DNA offers a novel means of tracing human migrations. *Proc Natl Acad Sci USA* **94:** 9191–9196.

Trowbridge PW, Frisque RJ (1995). Identification of three new JC virus proteins generated by alternative splicing of the early viral mRNA. *J NeuroVirol* **1:** 195–206.