Journal of General Virology (1998), **79**, 321–332. Printed in Great Britain

Short sequences define genetic lineages: phylogenetic analysis of group A rotaviruses based on partial sequences of genome segments 4 and 9

L. Maunula and C.-H. von Bonsdorff

Department of Virology, Haartman Institute, (Haartmaninkatu 3), POB 21, FIN-00014 University of Helsinki, Finland

Genetic diversity in strains of human group A rotaviruses was analysed by phylogenetic methods. The study material comprised 109 serotype G1 or G4 rotavirus samples isolated in Finland during 1986–1990. Parts of the coding regions of rotaviral genome segments 4 and 9, which encode proteins with serotype specificity, the spike protein VP4 (P serotype) and the outer capsid protein VP7 (G serotype), respectively, were sequenced. As determined by analysis of segment 4 sequences all G1 strains and all except one G4 strain showed P[8] specificity, the one being of P[6] specificity. The G1P[8] strains could be further differentiated into four groups based on segment 9 sequences, while G4P[8] strains formed only one group. Type P[8] (G1P[8] and G4P[8]) strains formed two main groups based on segment 4 sequences, suggesting free segregation of segment 4 between these G strains. Most global G1, G4 and P[8] strains in GenBank/EMBL originating from the 1970s to the present co-clustered with these groups, suggesting that the groups exist as relatively stable lineages. No linear accumulation of nucleotide substitutions was detected in strains of one serotype during the study period. Also, the deduced amino acids of the antigenic regions A, B and C of VP7 were nearly conserved within the phylogenetic lineages. Interestingly, only short amino acid sequences were necessary to divide the e-types correctly into phylogenetic lineages. These amino acid signature motifs were located in aa 29–68 of VP7 and aa 121–135 of VP4 of the G1 and P[8] lineages, respectively.

Introduction

Rotavirus is a genus of the family *Reoviridae*, a special feature of which is a segmented dsRNA genome (reviewed in Kapikian & Chanock, 1996). Each of the 11 rotaviral segments encodes at least one protein, six of which are structural. The icosahedral, ca. 70 nm rotavirus particle consists of an inner core of proteins VP1, VP2 and VP3 and the 11 genome segments, an inner capsid formed by protein VP6 and an outer capsid that contains the glycoprotein VP7 and dimeric protein spikes of VP4.

There are several classification systems for rotaviruses based either on analysis of the RNA genome or on antigenic properties of proteins VP4, VP6 or VP7. The 11 RNA segments of rotavirus give rise to a typical RNA migration pattern in PAGE, an electropherotype (e-type). The immunologically distinct rotavirus groups (A, B, C etc.) also have very distinct e-types. In addition, there is a lot of minor variation in the e-types of group A rotaviruses, which are important aetiological agents of gastroenteritis in small children. Gel electrophoresis has been widely used in analysis of diarrhoeal rotavirus samples all over the world, because it rapidly reveals changes in migration of any of the 11 RNA segments. It does not, however, reveal the nature of the mutational event, which may be from a point mutation, recombination, genome rearrangement or segment reassortment (Chanock *et al.*, 1983).

One well-established antigenic classification of rotaviruses, G (glycoprotein) serotyping, is based on identification of epitopes on the VP7 protein, which forms the smooth external surface of the virion. This glycoprotein of 326 amino acids is encoded by segment 9 (1062 bases) (reviewed in Estes & Cohen, 1989). VP7 has nine regions that are variable across serotypes, and three of them, regions A, B and C (aa 86–101, 142–152 and 208–221, respectively), have been shown to be antigenically important (Coulson & Kirkwood, 1991). In humans, 10 G serotypes have so far been identified, G1–G4 being the most common.

Author for correspondence: Leena Maunula.
Fax +358 9 1912 6491. e-mail Leena.Maunula@helsinki.fi

0001-5183

A binary classification system of G and P serotypes has recently been established for rotaviruses, analogous to the classification of influenza viruses into H (haemagglutinin) and N (neuraminidase) subtypes. The P (protease sensitive) serotype of rotavirus is determined by the spike protein VP4. This protein of 775 amino acids in human rotavirus is encoded by segment 4, which is 2359 bp long. In simian rotavirus strain SA-11, VP4 has been shown to be enzymatically cleaved to VP8* and VP5* at amino acid positions arginine-241 or arginine-247, which results in increased virus infectivity. VP5* is more conserved than VP8* with only a short variable region, aa 578–608 (Estes & Cohen, 1989). VP8* has one large variable type-specific region, aa 84–180 (Larralde & Gorziglia, 1992), where many neutralization epitopes are situated (Ruggeri & Greenberg, 1991). Currently, two overlapping classification systems of VP4 are used: P serotypes are determined by neutralizing MAbs, while P genotypes, or P types, are determined by analysis of genome segment 4 sequences. A total of eight P types has been identified in humans (Gorziglia *et al.*, 1988). The most common, type P[8] (or P1A when determined by neutralization), has been found in combination with serotypes G1, G3, G4 and G9. Type P[6] (P2A) was first identified in asymptomatic rotavirus infections of newborns, but later was also found in symptomatic cases associated with serotypes G1–G4 and G9 (Larralde & Flores, 1990; Santos *et al.*, 1994; Ramachandran *et al.*, 1996).

Group A rotaviruses cause seasonal epidemics in countries with a temperate climate in the winter months and many e-types and serotypes occur at the same time during an epidemic (Bishop, 1994). We have previously determined the e-types and G serotypes of 769 rotavirus isolates collected from hospitalized children in the Helsinki metropolitan area during 1986–1990 (Maunula & von Bonsdorff, 1995). During this period two predominant e-types (with more than 100 isolates) and 85 minor e-types (with less than 30 isolates) were detected. One e-type of serotype G1 and one of type G4 predominated over the first and third seasons, respectively. Each season with a predominant e-type was followed by an epidemic with multiple minor e-types. In this paper, sequence variation in rotavirus isolates representing different e-types of the two most commonly found G serotypes, G1 and G4, was analysed by phylogenetic methods. The most variable parts of the two segments, 4 and 9, encoding the antigenically important proteins VP4 and VP7 and defining the P and G serotypes, respectively, were selected for sequencing. Our purpose was to determine and display the genetic relationship between the co-circulating e-types representing these serotypes and to find out to what extent the predominant e-types were related to the minor ones which appeared after them.

Methods

■ **Virus samples.** The study material consisted of 109 rotavirus-positive human stool samples representing 69 different e-types: 85 serotype G1 rotavirus samples (56 e-types) and 24 serotype G4 samples (13 e-types). The material represented most G1 and G4 e-types detected among all samples sent to the Department of Virology, University of Helsinki, for the study of enteric viruses by electron microscopy from children's wards of mainly three metropolitan hospitals during 1986–1990. The e-types and G serotypes of the samples were determined previously (Maunula & von Bonsdorff, 1995). Some isolates of each e-type that remained untypable were sequenced and those determined by analysis of segment 9 sequences to represent serotype G1 or G4 were also included in the study.

■ **RNA extraction, gel electrophoresis and silver staining.** These methods are described in our previous paper (Maunula & von Bonsdorff, 1995). Briefly, the nucleic acid from a rotavirus sample of 10% stool suspension in 0·05 M Tris–HCl/0·1 M NaCl, pH 7·4 with 1 mM $CaCl_2$ was extracted by treatment with phenol and chloroform–isoamyl alcohol, followed by ethanol/acetic acid precipitation. Samples were resolved in a 7·5% polyacrylamide gel (Laemmli, 1970). Silver staining was performed as modified from Herring *et al.* (1982) including fixing in 10% ethanol–0·1% acetic acid, staining in 0·18% $AgNO_3$, brief washing with distilled water, development in 3% NaOH with 8 ml/l formaldehyde and post-fixation with 5% acetic acid–0·5% glycerol.

■ **RNA extraction and RT–PCR.** RNA was extracted from 10% stool suspension with phenol–chloroform treatment and possible inhibitors of RT–PCR were removed by washing the samples while nucleic acids were attached to CF-11 cellulose according to the method of Wilde *et al.* (1990). For RT–PCR the primers used for segment 9 were as previously published (Beg9: 1–28, 5′ GGC TTT AAA AGA GAG AAT TTC CGT CTG G 3′ and End9: 1062–1036, 5′ GGT CAC ATC ATA CAA TTC TAA TCT AAG 3′) with the latter primer biotinylated (Gouvea *et al.*, 1990). For segment 4 the primers used were as published by Gentsch *et al.* (1992) with minor modification (con3mod:11–32, 5′ TGG CTT CGC TCA TTT ATA GAC A 3′, con2: 868–887, 5′ ATT TCG GAC CAT TTA TAA CC 3′, biotinylated). The RT reaction was carried out separately as a 20 µl reaction mixture for 1 h at 37 °C using 50 mM Tris–HCl (pH 8·8) 75 mM KCl, 10 mM DTT, 3 mM $MgCl_2$, 0·5 mM nucleotides (Pharmacia), 0·5 µM primer, 1 U/µl RNasin inhibitor (Promega), 10 U/µl MMLV reverse transcriptase (BRL) and rotavirus RNA that had been pretreated by heating for 5 min in the presence of 37·5% dimethylsulfoxide and chilling on ice for 5 min. For PCR, 5 µl of RT reaction mixture was added to a 100 µl reaction mixture containing 10 mM Tris–HCl (pH 8·3), 50 mM KCl, 1·5 mM $MgCl_2$, 0·001% gelatin, 0·1 mM nucleotides, 0·2 µM each primer and 2·5 U/100µl AmpliTaq DNA polymerase (Perkin Elmer). For segment 9 the PCR programme was as follows: 94 °C 1·0 min, 42 °C 2·0 min and 72 °C 2·0 min (33 cycles) followed by a final 7 min at 72 °C. For segment 4 the procedure was the same except that the annealing temperature was 50 °C. During the extraction and RT–PCR special care was taken to prevent nucleic acid contamination. *Pfu* DNA polymerase (Stratagene), which makes fewer errors than *Taq* polymerase, was used in small-scale experiments to confirm that the point mutations of the samples were authentic.

■ **Direct DNA sequencing.** Avidin-coated beads (IDEXX) were used to trap the biotinylated PCR products and the DNA strands were separated with NaOH (Syvänen *et al.*, 1989) before sequencing by the dideoxynucleotide chain termination method (Tabor & Richardson, 1990). In detail, 40–100 µl biotinylated PCR product was mixed with 10 µl beads and the volume made up to 100 µl with TENT buffer (50 mM NaCl, 40 mM Tris–HCl, 1 mM EDTA and 0·01% Tween 20, pH 7·5). The beads were collected by centrifugation (1–2 min, 3200 ***g***). They were then suspended in 50 µl 50 mM NaOH for 10–15 min at room temperature after which the centrifugation was repeated. After washing

the beads twice with TENT buffer they were suspended in 10 μl of reaction buffer (40 mM Tris–HCl, 20 mM $MgCl_2$ and 50 mM NaCl, pH 7·5) with 1 μM of primer and the annealing reaction was allowed to proceed at 37 °C for 15–30 min. Segment 9 was sequenced partially (663 bp, 100–762, part of the coding region) by using two successive primers – F9 (51–71, 5′ GTA TGG TAT TGA ATA TAC CAC 3′) and F2 (376–392, 5′ GGA TGG CCA ACA GGA TC 3′) (Flores *et al*., 1990) – and segment 4 was sequenced partially (375 bp, 250–624) by using primer S4o (205–221, 5′ GAT GGT CCT TAT CAG CC 3′). The sequencing reaction was performed according to the instructions of the Sequenase version 2.0 DNA sequencing kit (USB) by manual sequencing using [α-^{35}S]dATP as a label for autoradiography. The reaction products were resolved on 6% polyacrylamide–7 M urea thin gels in taurine buffer, pH 8·8 (89 mM Tris base, 29 mM taurine, 0·5 mM EDTA).

■ **Sequence analysis.** Phylogenetic analyses of sequence data were performed using the PHYLIP software package (Felsenstein, 1989), the fastDNAml program (which is derived from the DNAML program of the PHYLIP package) (Olsen *et al*., 1994), and UWGCG programs, especially PILEUP (Devereux *et al*., 1984).

■ **Nucleotide sequence accession numbers.** The sequences described here are available from the GenBank/EMBL database. The accession numbers for segments 4 and 9 of G1 rotavirus strains are Z80234–Z80270 and Z80271–Z80315, respectively, and for segments 4 and 9 of G4 strains Z80316–Z80328 and Z80329–Z80338, respectively (including the G4P[6] strain 314, which is under accession numbers Z80321 and Z80332). Only nonidentical sequences were submitted (see phylogenetic trees in Fig. 1).

Results

Deduced P types

A total of 109 group A rotavirus samples representing 69 G1 or G4 e-types isolated during 1986–1990 was analysed by partial sequencing of segments 4 and 9 (Table 1). Eight isolates of each predominant type and at least one isolate of each minor e-type were selected for sequencing. The P types of the strains were determined by analysis of the partial segment 4 sequences: All samples of serotype G1 and all except one sample of type G4 represented type P[8]. Only one type G4 sample, e-type 314, represented type P[6]. This strain was most probably of foreign origin, because the patient in question had visited Greece just before getting ill.

Phylogenetic analyses

Partial sequences of two segments, 4 and 9, of the rotavirus strains were used in parallel to construct phylogenetic trees (Fig. 1). The serotype G1P[8] and G4P[8] strains were analysed separately. The phylogenetic trees revealed that many e-types that had clearly different overall electrophoretic RNA patterns (examples are shown in Fig. 2) had identical nucleotide sequences for one or both of the segment(s): the 56 e-types divided into only 44 and 37 different sequences of segments 9 and 4, respectively.

The G1P[8] strains that were analysed with the fastDNAml program were distributed into several groups in the phylogenetic trees: segment 9 sequences defined four groups, VP7-G1-1–VP7-G1-4 (Fig. 1*a*) and segment 4 sequences three groups, VP4-P[8]-1, -2 and -3 (Fig. 1*b*). These groups coexisted as rather stable lineages during the 4 year study period. The sequences of the two segments did not define identical phylogenetic trees: there were differences both in branching order and in branching lengths of the trees. Closer scrutiny of the two phylogenetic trees of G1P[8] strains revealed that most e-types in the lineage VP7-G1-1 of segment 9 sequences (Fig. 1*a*) localized into the lineage VP4-P[8]-1 of segment 4 sequences (Fig. 1*b*). Also, most e-types in the lineage VP7-G1-2 localized into this same VP4-P[8]-1 lineage. Actually, the lineage VP7-G1-2 showed a high level of nucleotide identity with VP7-G1-1 (about 96%), although it consistently formed a distinct lineage in the phylogenetic analyses. Most e-types of the distinct lineage VP7-G1-3 localized into a distinct lineage of VP4-P[8]-2. Lineage VP7-G1-1 could be further divided into three sublineages, a, b and c. The lineage VP4-P[8]-1 also formed three sublineages, but the same e-types were distributed differently between them. Interestingly, there were four e-types (e-types 217, 404: 2, 413 and 420) that localized into different VP4-P[8] lineages from the other e-types of their VP7-G1 lineage. They might have been derived through gene reassortment.

The most common combination of the two genome segments in the G1P[8] e-types was that segment 9 was from the lineage VP7-G1-1 and segment 4 from the lineage VP4-

Table 1. Number of group A rotavirus samples (number of different e-types) sequenced in this study

Serotype	Season 1 1986/1987	Season 2 1987/1988	Season 3 1988/1989	Season 4 1989/1990	Total
G1P[8]	24 (12*)	23 (16)	8 (6)	30 (22)	85 (56)
G4P[8]	2 (1)	0	9 (2†)	12 (9)	23 (12)
G4P[6]	0	0	1 (1)	0	1 (1)
Total	26 (13)	23 (16)	18 (9)	42 (31)	109 (69)

* Includes the predominant e-type 101.
† Includes the predominant e-type 305.

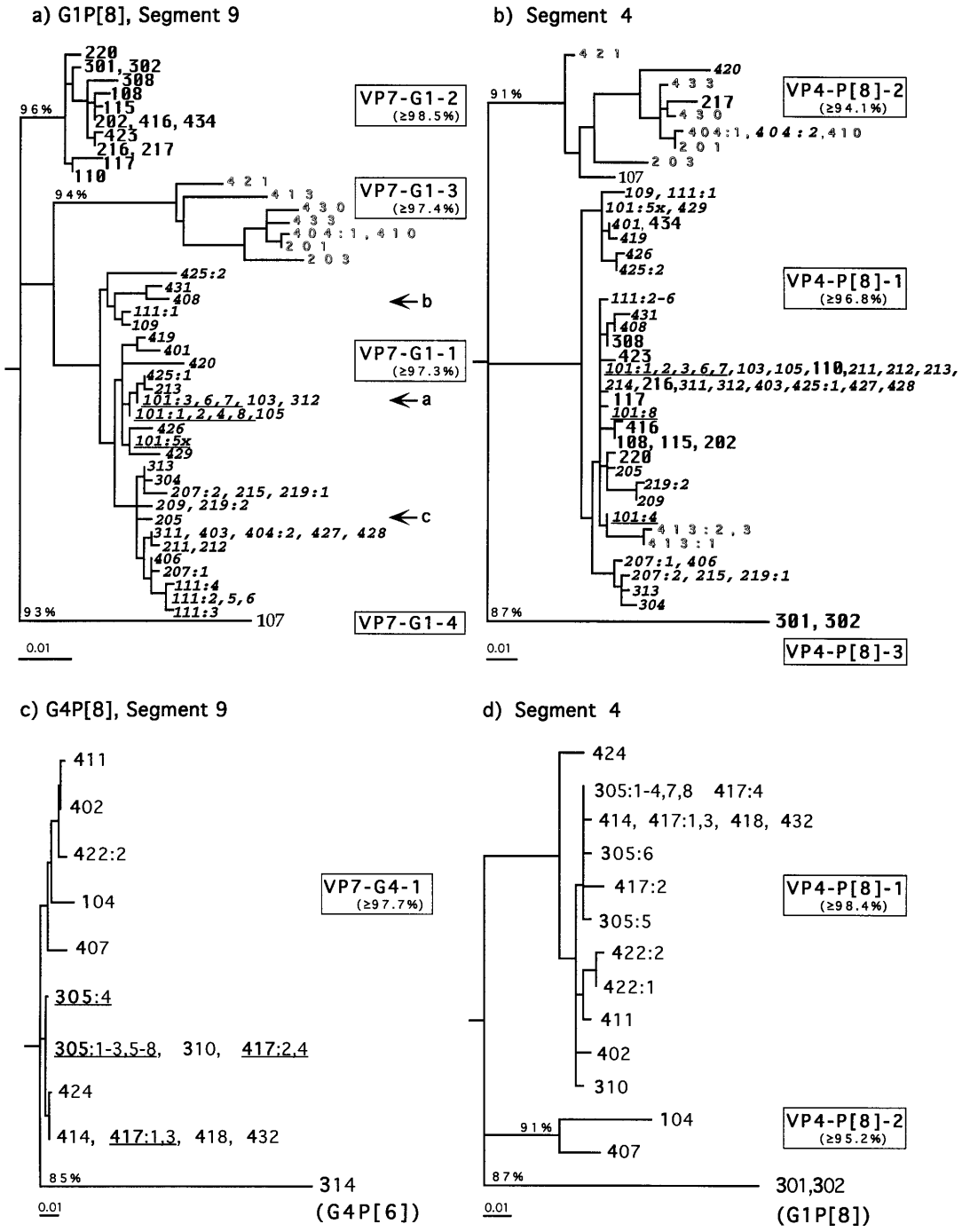

Fig. 1. (*a*, *b*) Phylogenetic trees derived with segment 9 and 4 sequences (nt 100–762 and 250–624, respectively) of G1P[8] rotaviruses using the fastDNAml (maximum likelihood) program. (*c*, *d*) Phylogenetic trees derived with the partial sequences of G4P[8] rotaviruses using the DNAML program in the PHYLIP software package. In (*a*) an identical tree was obtained whether the e-type 107 was taken as an outgroup or no outgroup was named. In (*b*) and (*d*) the e-type 301 (G1P[8]) was taken as an outgroup, and in (*c*) the e-type 314 (type G4P[6]) was used. The nucleotide identity levels within lineages are indicated in parentheses. The minimum identity levels of the predominant e-types (underlined) in comparison with the other lineages are indicated on the branches. Genetic distance is proportional to horizontal branch lengths (bar represents 0·01) : vertical lines are used for graphic representation only. The first digit in the e-type code indicates the season of isolation (1, 1986/87; 2, 1987/88; 3, 1988/89; 4, 1989/90) and the next two digits indicate the identification number in chronological order of appearance: the sequenced isolate is indicated after the colon, if necessary.

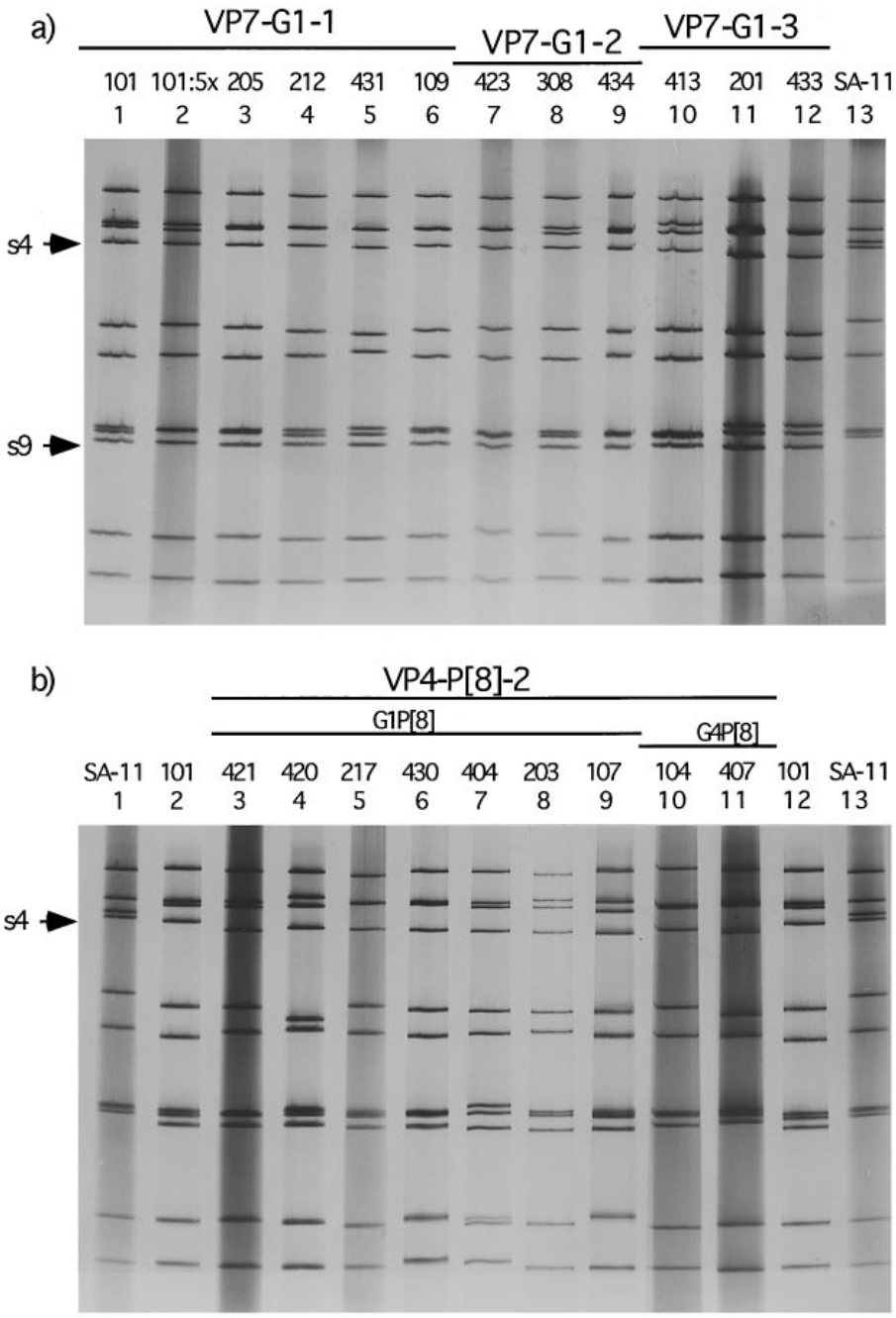

Fig. 2. Selected e-types representing phylogenetic lineages. (*a*) Some e-types of G1 rotaviruses representing lineages VP7-G1-1, -2 and -3 (lanes 1–6, 7–9 and 10–12, respectively). (*b*) E-types of G1P[8] and G4P[8] rotaviruses from the lineages VP4-P8-2 with fast moving segment 4 (lanes 3–9 and 10–11, respectively). For comparison, the predominant G1 e-type 101 is represented in lanes 2 and 12. Segments 4 and 9 are indicated by arrows. Simian rotavirus SA-11 was used as a marker in all gels.

P[8]-1, as in the e-type 101 that predominated in the first epidemic season (isolates underlined in Fig. 1*a*, *b*). Evidently, many of the minor e-types were closely related to the predominant e-type, some of them perhaps descendants of it, but rather than a linear accumulation of nucleotide substitutions starting from the sequence of the predominant e-type, a large cluster with slightly varying sequences was formed around the predominant type. In the lineage VP4-P[8]-1, as many as 14 e-types had identical segment 4 sequences to the predominant e-type and only slightly different segment 9 sequences. Nevertheless, a small proportion of the minor e-types from the other G1 and P[8] lineages were also circulating during all epidemic

Table 2. Nucleotide and amino acid substitutions among isolates with identical e-types

(*a*) E-type 101 of G1P[8] with samples isolated during 2.9.1986–10.8.1987. (*b*) E-type 305 and e-type 417 of G4P[8] identical with it. Samples were isolated during 16.1.1989–19.7.1989 and 11.3.1990–12.6.1990, respectively. (Isolate 101: 5 × turned out to have a slightly different RNA pattern from the others and thus is not included in this table.)

E-type: Isolate	Segment 9 Nucleotide change	VP7 Amino acid change	Segment 4 Nucleotide change	VP4 Amino acid change
(*a*) Comparison to isolates 101: 1, 2 with identical sequences				
101: 3, 6, 7	744; GG<u>G</u> → GG<u>T</u>	Silent	None	–
101: 4	None	–	489; GA<u>T</u> → GA<u>C</u>	Silent
101: 8	None	–	487; <u>G</u>AT → <u>A</u>AT	160; Asp → Asn
(*b*) Comparison to isolates 305; 1, 2, 3, 7, 8, and 417: 4 with identical sequences				
305: 4	617; T<u>C</u>A → T<u>T</u>A	90; Ser → Leu	None	–
305: 5	None	–	340; <u>C</u>AC → <u>T</u>AC	111; His → Tyr
305: 6	None	–	319; <u>G</u>CA → <u>A</u>CA	104; Ala → Thr
417: 1, 3	496; <u>T</u>TG → <u>C</u>TG	Silent	602; A<u>T</u>T → A<u>G</u>T	198; Ile → Ser
417: 2	None	–	300; AA<u>T</u> → AA<u>C</u>	Silent
			343; <u>G</u>TT → <u>A</u>TT	112; Val → Ile
			579; AC<u>T</u> → AC<u>C</u>	Silent

seasons. Representatives of the lineage VP7-G1-2 appeared throughout the whole 4 year study period, but the e-types of the lineage VP7-G1-3 were circulating only during seasons 2 and 4 when no predominant e-type was prevailing. Only one genome rearrangement was detected in the present study, namely one G1P[8] isolate duplicated residues ACA AAT (271–276) of segment 4 (not included in the phylogenetic trees).

Type G4P[8] strains were analysed with the PHYLIP program (Fig. 1*c*, *d*). Only 12 e-types of serotype G4 circulated during the study period, two of them during the third epidemic season, when the e-type 305 predominated, and most others during the fourth season, when a minor e-type (designated 417) that was very similar to the predominant e-type was also detected. Unlike the G1 strains, the G4 strains formed only one lineage, VP7-G4-1, defined by segment 9 sequences (Fig. 1*c*). This may be partly due to the limited time of appearance of the e-types. In this case, it may be that many of the minor e-types were descendants of the predominant e-type. The e-type 314 of different P type, P[6], diverged considerably from the others with respect to the segment 9 sequence (taken as an outgroup of the phylogenetic tree in Fig. 1*c*; segment 4 is not included in Fig. 1*d*). Segment 4 sequences of the G4P[8] strains formed a phylogenetic tree that resembled that of G1P[8] strains with two lineages (VP4-P[8]-1 and -2), which suggests free segregation of segment 4 between these G strains (Fig. 1*d*).

Electropherotypes versus phylogenetic lineages

The strains forming the phylogenetic lineages had a large variability in their overall electrophoretic RNA patterns as illustrated in Fig. 2. Assuming that the ninth segment encodes the VP7 gene in all G1 e-types, as shown for the Wa strain of serotype G1 (Greenberg *et al*., 1981), no clear differences in the migration of this segment were evident between the phylogenetic lineages of VP7-G1 (Fig. 2*a*). In contrast, the migration of segment 4 clearly differed between the two phylogenetic lineages VP4-P[8]-1 and -2, so that the e-types of lineage P[8]-2 had a faster migrating segment 4 than those of lineage P[8]-1 (Fig. 2*b*). The variability of e-types within lineages was due to differences in the mobilities of segments other than 4 and/or 9.

Sequence identity among isolates of the predominant electropherotypes

In general, the isolates representing the same minor e-type had identical sequences. In the few cases where nucleotide differences were detected, co-electrophoresis of the samples revealed that they represented different e-types after all. Some internal sequence variation was found among isolates of the predominant strains (Table 2). Among the eight isolates sequenced representing the e-type 101, three isolates had one nucleotide substitution in segment 9 and two isolates had one (different) substitution in segment 4. At least one of the two segments of all isolates was, however, identical to that of the first isolate. Only one point mutation changed an amino acid: acidic aspartate-160 of VP4 of isolate 8 changed to uncharged asparagine.

The isolates of the predominant e-type 305 of type G4P[8] and the minor e-type 417 from the successive season with an identical RNA migration pattern were analysed (Table 2*b*). As

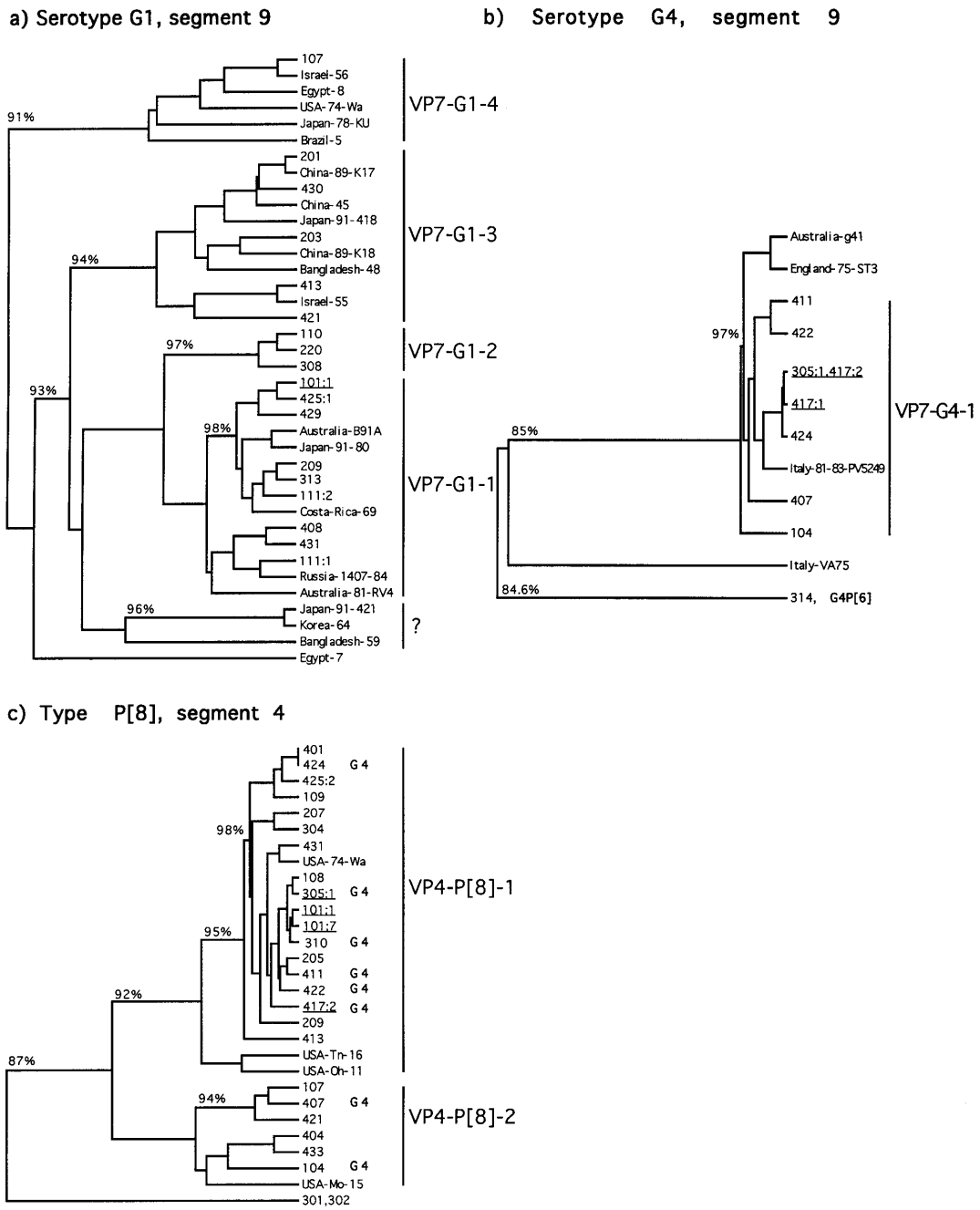

Fig. 3. Dendrograms constructed with segment 9 sequences (nt 100–762) of (*a*) serotype G1 viruses and (*b*) serotype G4 viruses, and (*c*) a dendrogram constructed with segment 4 sequences (nt 250–624) of combined G1P[8] and G4P[8] strains, using the PILEUP program in the UWGCG software package. Only the e-types and global strains essential in forming the clusters were included. In (*c*) the e-types representing serotype G4 are marked; the others represent type G1. The approximate nucleotide identity levels are given at some branch nodes. Accession numbers of the global strains in GenBank/EMBL are as follows. (*a*) Israel-56, U26376; Egypt-8, U26374; USA-74-Wa, K02033; Japan-78-KU, D16343; Brazil-5, U26367; China-89-K17, D16320; China-45, U26371; Japan-91-418, D16327; China-89-K18, D16319; Bangladesh-48, U26364; Israel-55, U26375; Australia-B91A, M93006; Japan-91-80, D16325; Costa-Rica-69, U26369; Russia-84-1407, S83903; Australia-81-RV4, M64666; Japan-91-421, D16326; Korea-64, U26378; Bangladesh-59, U26366; Egypt-7, U26373. (*b*) Australia-g41, A01321; England-75-ST3, X13603; Italy-81-83-PV5249, M86490; Italy-VA75, M86833. (*c*) USA-74-Wa, L34161; USA-Tn-16, U26763; USA-Oh-11, U26757; USA-Mo-15, U26755.

expected, the sequences shared identity to the extent that one isolate of the latter e-type had segment 4 and 9 sequences identical to the first isolates of the e-type 305. A few substitutions were, however, detected in the other isolates of e-type 417: two of them had one substitution in both segments and one had three substitutions in segment 4. Interestingly,

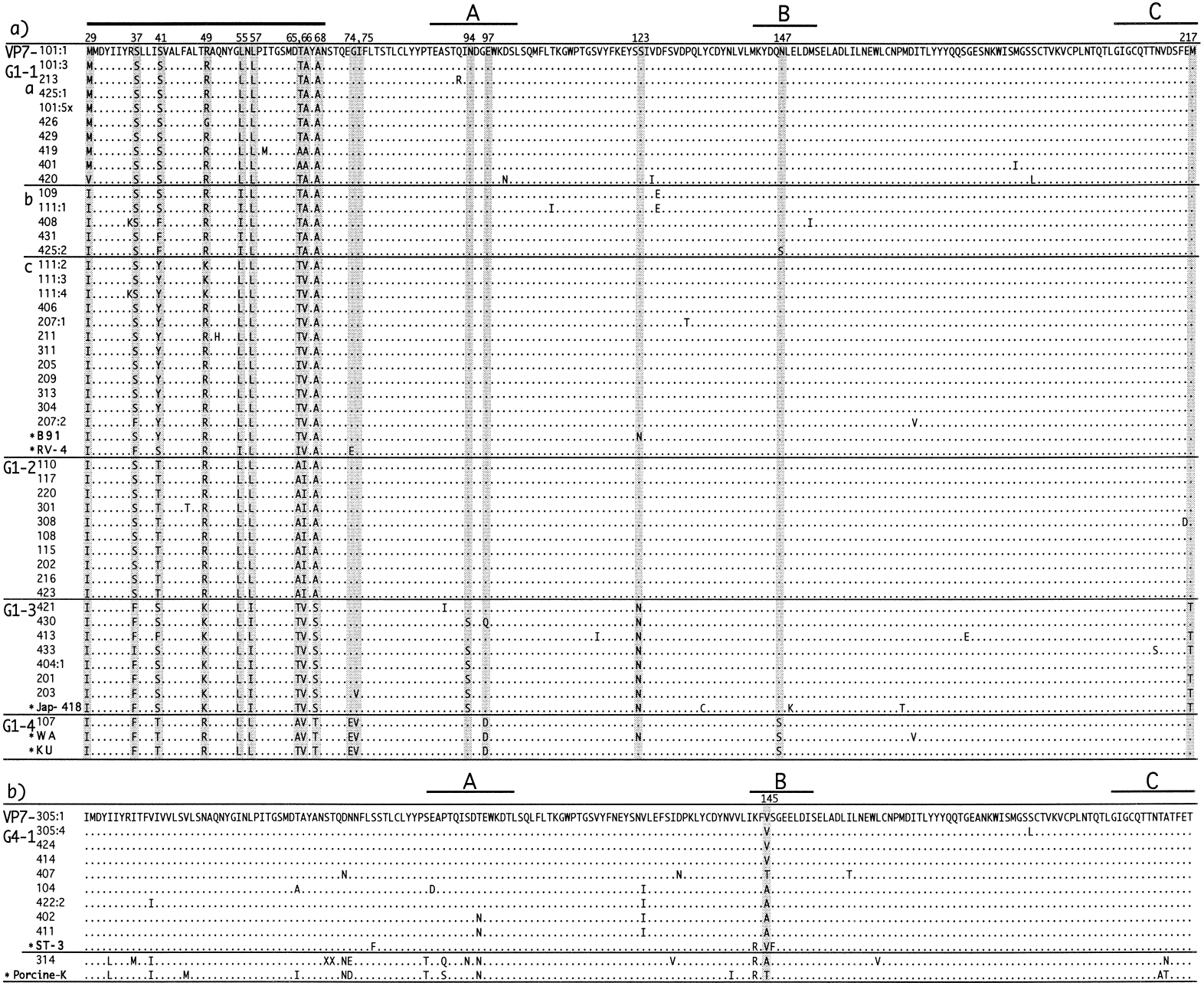

Fig. 4. Alignment of VP7 amino acid sequences (aa 29–217) of (*a*) G1 strains and (*b*) G4 strains. Antigenic sites A, B and C and the phylogenetic lineage-defining sequences are indicated by horizontal bars. All amino acids showing lineage or sublineage specificity are shaded. The global strains B91 and RV-4 from Australia, 418 from Japan, Wa and KU are included in (*a*) and the ST-3 and porcine K strains (X58439) in (*b*).

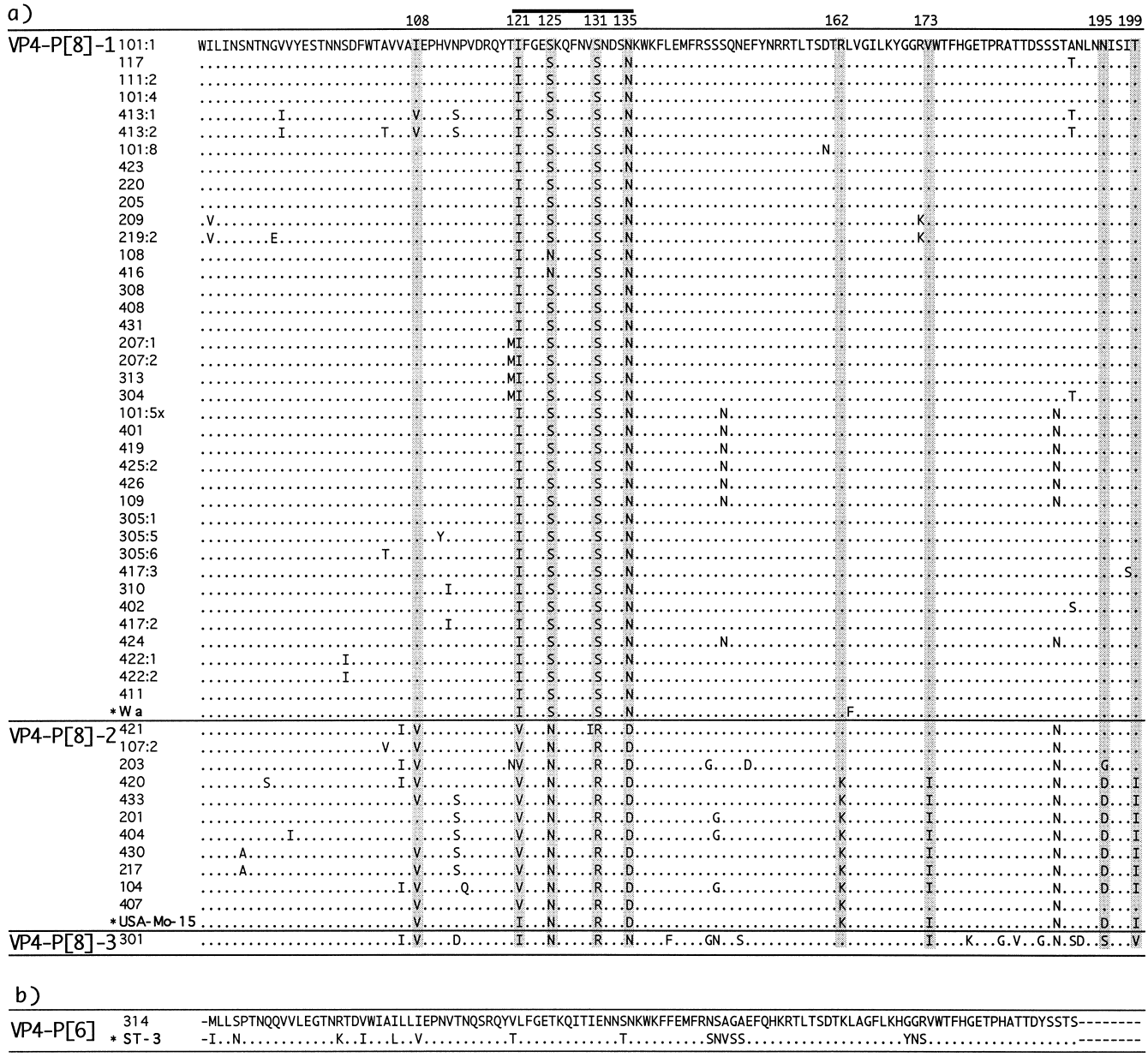

Fig. 5. Alignment of VP4 amino acid sequences of (*a*) aa 81–199 of P[8] lineages VP4-P8-1, -2 and -3 together with global strains Wa and USA-Mo-15 and (*b*) of aa 82–191 of the type G4P[6] e-type 314 and ST-3. The phylogenetic lineage-defining sequences are indicated by a horizontal bar. All amino acids showing lineage specificity are shaded.

most nucleotide substitutions in segment 4 resulted in a amino acid changes, one of which changed basic histidine-111 to uncharged tyrosine.

Global comparison of nucleotide sequences

The nucleotide sequences determined in this study were compared to the data available in GenBank/EMBL using the UWGCG PILEUP program. In all dendrograms, our strains and global strains were mixed without any geographical clustering and global strains localized mainly into the same lineages as the strains of our study (Fig. 3). In the dendrogram defined by segment 9 sequences of G1 strains (Fig. 3*a*) the predominant e-type 101 clustered together with strains from Australia, Japan, Costa Rica and Russia. Interestingly, the Australian strain B91A was a predominant strain in Melbourne at the beginning of the 1990s (Palombo *et al.*, 1993), which suggests that this VP7-G1-1 lineage has prevailed in different parts of the world in recent years. The Wa strain, isolated as early as 1974, co-localized with the e-type 107 into the lineage VP7-G1-4. The few global G4 sequences thus far available localized into the same G4 lineage as our strains with the exception of the Italian strain VA-75 (Fig. 3*b*). When a dendrogram was constructed

with segment 4 sequences from all our P[8] strains and those from GenBank/EMBL, the division into two clusters remained (Fig. 3*c*). The G1 and G4 strains were mixed in both clusters to the extent that two e-types (401 and 424) representing different G types had identical segment 4 sequences. Taken together, these phylogenetic analyses conformed to the idea of rather stable VP7 and VP4 lineages, the earliest samples originating from the 1970s.

Alignment of the deduced amino acid sequences

The deduced amino acid sequences of VP7 and VP4 proteins were aligned for the different e-types (Figs 4 and 5). When the antigenic regions A, B and C of VP7 were examined (Fig. 5*a*), it became evident that these regions were very similar in lineages VP7-G1-1 and -2 with only single amino acid substitutions. Most e-types of the lineage G1-3 had serine in position 94 of region A instead of asparagine, and threonine-217 in region C instead of methionine. The only e-type 107 of lineage G1-4 had aspartate-97 in region A instead of glutamate, and serine-147 in region B instead of asparagine. Also, global strains Wa and KU were identical to the e-type 107 in these positions. These lineages might have slightly different antigenic properties as compared to the lineages G1-1 and -2, because it has been reported that amino acids 94 and 147 are important in further division of G1 strains into monotypes by a panel of MAbs (Coulson & Kirkwood, 1991). Also, amino acid substitutions in all these four positions have been reported to occur in mutants selected with MAbs exhibiting homo- and/or heterotypic neutralizing activity (reviewed in Kapikian & Chanock, 1996).

Interestingly, the alignment of the amino acid sequences revealed short sequences of VP4 and VP7 that generally defined the same lineages as were derived from the phylogenetic analyses of partial nucleotide sequences of segments 4 and 9 (Figs 4*a* and 5*a*). Among the VP7 sequences of the G1 strains these lineage-defining amino acids concentrated in the beginning of the sequenced area (Fig. 4). Nine amino acids (29, 37, 41, 49, 55, 57, 65, 66 and 68) formed an identification code for each lineage or sublineage, e.g. sublineage VP7-G1-1a had a code of MSSRLLTAA and lineage VP7-G1-4 had a code of IFTRLLAVT, with one replacement accepted. Most global strains (some aligned in Fig. 5*a*) also conformed to these codes, with the exception of RV-4, which perhaps represents a new sublineage of VP7-G1-1. As expected, among the VP7 sequences of the G4 strains only single amino acid substitutions throughout the sequence were seen (Fig. 4*b*). The P[8] lineage signature motifs were concentrated into the sequence 121–135 (Fig. 5*a*). Four amino acids at positions 121, 125, 131 and 135 formed a code of ISSN, VNRD and INRN for the lineages VP4-P[8]-1, -2 and -3, respectively. In addition to these strictly conservative amino acid positions within lineages, several less conservative positions (108, 162, 173, 195 and 199) were observed throughout the sequenced area of VP4.

The only G4P[6] e-type (314) was compared to other G4 (segment 9) or P[6] (segment 4) sequences in GenBank/EMBL, but no close relatives were found. Interestingly, the VP7 amino acid sequence of e-type 314 was more similar to the porcine G4 strain K (X58439; amino acid identity 93·7%) than to human G4 strains (Fig. 4*b*). The closest human strain, our e-type 305, had an amino acid identity of 92·3% and the ST-3 strain (type G4P[6]) an amino acid identity of 90·5% with strain K. The VP4 protein of the e-types 314 had the most similar sequence to that of ST-3 strain with an amino acid identity of 85·6% and similarity of 91·0% (Fig. 5*b*).

Discussion

Rotaviruses are well suited for analysis of the evolution of RNA viruses. In the gut, they replicate readily to high titres which allows genetic analysis of virus obtained directly from patient samples. With their segmented genome, rotaviruses have ample possibility to reassort, since numerous rotavirus strains co-circulate and double infections often occur (Gouvea *et al*., 1995). This study concerns rotavirus strains only from hospitalized children, which might bias strain selection. However, the distribution of rotavirus serotypes in hospitalized children has been reported to be relatively similar to that in children attending day-care centres during the same period (O'Ryan *et al*., 1990).

The genetic variation of group A rotaviruses within serotypes G1 and G4 was analysed by sequencing parts of two genes encoding the outer capsid proteins VP4 and VP7. The P types of both these G serotypes turned out to be very homogeneous, all but one representing type P[8] determined by sequence analysis. The phylogenetic studies revealed that the isolates of serotype G1 and P[8] could be further divided into several groups or lineages. Earlier, Xin *et al*. (1993) suggested that segment 9 sequences of rotavirus strains from Japan and China could be divided into three G1 subtypes. In our analysis, their samples localized into our groups VP7-G1-1, -3 and -4. In contrast to the G1 serotype, only one lineage of serotype G4 circulated during the study period. One explanation for the almost continuous presence of serotype G1 (Woods *et al*., 1992) might be the fact that it has several lineages.

The immense genetic variability of rotaviruses expressed in e-types is still an enigma for rotavirologists (Anon., 1990). The analysis of nucleotide sequences of two segments, 4 and 9, revealed that many different e-types had one or both of these segments identical. Based on this observation, it is obvious that the other RNA segments show genetic variability. Indeed, proteins VP6 (Greenberg *et al*., 1983) and VP2 (Taniguchi *et al*., 1986) are known to comprise antigenically different subgroups. Further, it has been reported that segment 5 shows considerable sequence variation among human G types (Dunn *et al*., 1994). In fact, most of the RNA segments of rotaviruses may allow division into phylogenetic lineages in future studies.

In the current study, reassortment was suggested for four e-types. Natural reassortment has been reported for many RNA viruses with a segmented genome, like influenza viruses (Peng *et al*., 1994) and bunyaviruses (Henderson *et al*., 1995). Also, for rotaviruses several reports suggesting reassortment *in vivo* have been published (e.g. Mascarenhas *et al*., 1989; Midthun *et al*., 1987). *In vitro*, reassortment between two rotaviruses has been shown to be quite efficient once they infect the same cell (reviewed in Ramig & Ward, 1991). Also, it has been shown that the event is more efficient the more closely related the strains are. Thus, it would be very likely to happen between lineages of one serotype. In our study, type G1P[8] and G4P[8] strains were mixed in P[8] lineages suggesting frequent reassortment between them. Genome rearrangement seemed to be a rare event in RNA segments 4 and 9 in our study. This conforms to the observation that rearrangements occur mainly in the genes encoding nonstructural proteins (Desselberger, 1996).

Short peptides were found to differentiate rotavirus strains into generally the same groups as the phylogenetic analysis of nucleotide sequences. The sequence that defined the G1 lineages was located at the beginning of the coding region of VP7; this includes regions VR-2–VR-4 that are also variable across rotavirus serotypes (Green *et al*., 1989). It also includes the signal H2 with the cleavage site and its last amino acids, aa 65–68, belong to a conservative hydrophilic region that includes a potential glycosylation site (aa 69–71). Amino acid signature sequence motifs specific for genotypes of another RNA virus have been reported recently (Örvell *et al*., 1997). The authors detected a triplet signature motif of three successive amino acids in the small hydrophobic (SH) protein gene of mumps virus. It could be speculated that some (steric?) constraints would limit the amino acid substitutions allowed in these positions and thus explain the finding.

We thank Anssi Mörttinen for excellent electron microscopy, Alexander Plyusnin, Matti Kaartinen and Heli Piiparinen for valuable advice in PCR and sequencing and Leena Kinnunen, Heikki Lehväslaiho, Kari Asikainen and Pekka Salonen for help in phylogenetic analyses and computer graphics. In addition, we thank Antti Vaheri for critical reading of the manuscript.

References

Anon. (1990). Puzzling diversity of rotaviruses. *Lancet* **335**, 573–575.

Bishop, R. F. (1994). Natural history of human rotavirus infections. In *Viral Infections of the Gastrointestinal Tract*, 2nd edn, pp. 131–167. Edited by A. Z. Kapikian. New York: Marcel Dekker.

Chanock, S. J., Wenske, E. A. & Fields, B. N. (1983). Human rotaviruses and genome RNA. *Journal of Infectious Diseases* **148**, 45–50.

Coulson, B. S. & Kirkwood, C. (1991). Relation of VP7 amino acid sequence to monoclonal antibody neutralization of rotavirus and rotavirus monotype. *Journal of Virology* **65**, 5968–5974.

Desselberger, U. (1996). Genome rearrangements of rotaviruses. *Advances in Virology* **46**, 69–95.

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

Dunn, S. J., Cross, T. L. & Greenberg, H. B. (1994). Comparison of the rotavirus nonstructural NSP1(NS53) from different species by sequence analysis and Northern blot hybridization. *Virology* **203**, 178–183.

Estes, M. K. & Cohen, J. (1989). Rotavirus gene structure and function. *Microbiological Reviews* **53**, 410–449.

Felsenstein, J. (1989). PHYLIP – phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.

Flores, J., Sears, J., Schael, I. P., White, L., Garcia, D., Lanata, C. & Kapikian, A. Z. (1990). Identification of human rotavirus serotype by hybridization to polymerase chain reaction-generated probes derived from a hyperdivergent region of the gene encoding outer capsid protein VP7. *Journal of Virology* **64**, 4021–4024.

Gentsch, J. R., Glass, R. I., Woods, P., Gouvea, V., Gorziglia, M., Flores, J., Das, B. K. & Bhan, M. K. (1992). Identification of group A rotavirus gene 4 types by polymerase chain reaction. *Journal of Clinical Microbiology* **30**, 1365–1373.

Gorziglia, M., Green, K., Nishikawa, K., Taniguchi, K., Jones, R., Kapikian, A. Z. & Chanock, R. M. (1988). Sequence of the fourth gene of human rotaviruses recovered from asymptomatic or symptomatic infections. *Journal of Virology* **62**, 2978–2984.

Gouvea, V. & Brantly, M. (1995). Is rotavirus a population of reassortants? *Trends in Microbiology* **3**, 159–162.

Gouvea, V., Glass, R. I., Woods, P., Taniguchi, K., Clark, H. F., Forrester, B. & Fang, Z.-Y. (1990). Polymerase chain reaction amplification and typing of rotavirus nucleic acid from stool specimens. *Journal of Clinical Microbiology* **28**, 276–282.

Green, K. Y., Hoshino, Y. & Ikegami, N. (1989). Sequence analysis of the gene encoding the serotype-specific glycoprotein (VP7) of two new human rotavirus serotypes. *Virology* **168**, 429–433.

Greenberg, H. B., Kalica, A. R., Wyatt, R. G., Jones, R. W., Kapikian, A. Z. & Chanock, R. M. (1981). Rescue of noncultivatable human rotavirus by gene reassortment during mixed infection with ts mutants of a cultivatable bovine rotavirus. *Proceedings of the National Academy of Sciences, USA* **78**, 420–424.

Greenberg, H. B., McAuliffe, V., Valdesuso, J., Wyatt, R., Flores, J., Kalica, A., Hoshino, Y. & Singh, N. (1983). Serological analysis of the subgroup proteins of rotavirus, using monoclonal antibodies. *Infection and Immunity* **39**, 91–99.

Henderson, W. W., Monroe, M. C., Jeor, S. C. S., Thayer, W. P., Rowe, J. E., Peters, C. J. & Nichol, S. T. (1995). Naturally occurring Sin Nombre virus genetic reassortants. *Virology* **214**, 602–610.

Herring, A. J., Inglis, N. F., Ojeh, C. K., Snodgrass, D. R. & Menzies, J. D. (1982). Rapid diagnosis of rotavirus infection by direct detection of viral nucleic acid in silver-stained polyacrylamide gels. *Journal of Clinical Microbiology* **16**, 473–477.

Kapikian, A. Z. & Chanock, R. M. (1996). Rotaviruses. In *Fields Virology*, 3rd edn, pp. 1657–1708. Edited by B. N. Fields, D. M. Knipe & P. M. Howley. Philadelphia: Lippincott–Raven.

Laemmli, U. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.

Larralde, G. & Flores, J. (1990). Identification of gene 4 alleles among human rotaviruses by polymerase chain reaction-derived probes. *Virology* **179**, 469–473.

Larralde, G. & Gorziglia, M. (1992). Distribution of conserved and specific epitopes on the VP8 subunit of rotavirus VP4. *Journal of Virology* **66**, 7438–7443.

Mascarenhas, J. D., Linhares, A. C., Gabbay, Y. B., Mendez, E. de F. R., Lopez, S. & Arias, C. F. (1989). Naturally occurring serotype 2/subgroup II rotavirus reassortants in northern Brazil. *Virus Research* **14**, 235–240.

Maunula, L. & von Bonsdorff, C.-H. (1995). Rotavirus serotypes and electropherotypes in Finland from 1986 to 1990. *Archives of Virology* **140**, 877–890.

Midthun, K., Valdesuso, J., Hoshino, Y., Flores, J., Kapikian, A. Z. & Chanock, R. M. (1987). Analysis by RNA–RNA hybridization assay of intertypic rotaviruses suggests that gene reassortment occurs in vivo. *Journal of Clinical Microbiology* **25**, 295–300.

Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeck, R. (1994). fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Computer Applications in the Biosciences* **10**, 41–48.

Örvell, C., Kalantari, M. & Johansson, B. (1997). Characterization of five conserved genotypes of the mumps virus small hydrophobic (SH) protein gene. *Journal of General Virology* **78**, 91–95.

O'Ryan, M. L., Matson, D. O., Estes, M. K., Bartlett, A. V. & Pickering, L. K. (1990). Molecular epidemiology of rotavirus in children attending day care centers in Houston. *Journal of Infectious Diseases* **162**, 810–816.

Palombo, E. A., Bishop, R. F. & Cotton, R. G. H. (1993). Intra- and inter-season genetic variability in the VP7 gene of serotype 1 (monotype 1a) rotavirus clinical isolates. *Archives of Virology* **130**, 57–69.

Peng, G., Hongo, S., Muraki, Y., Sugawara, K., Nishimura, H., Kitame, F. & Nakamura, K. (1994). Genetic reassortment of influenza C viruses in man. *Journal of General Virology* **75**, 3619–3622.

Ramachandran, M., Das, B. K., Vij, A., Kumar, R., Bhambal, S. S., Kesari, N., Rawat, H., Bahl, L., Thakur, S., Woods, P. A., Glass, R. I., Bhan, M. K. & Gentsch, J. R. (1996). Unusual diversity of human rotavirus G and P genotypes in India. *Journal of Clinical Microbiology* **34**, 436–439.

Ramig, R. F. & Ward, R. L. (1991). Genomic segment reassortment in rotaviruses and other reoviridae. *Advances in Virus Research* **39**, 163–207.

Ruggeri, F. M. & Greenberg, H. B. (1991). Antibodies to the trypsin cleavage peptide VP8′ neutralize rotavirus by inhibiting binding of virions to target cells in culture. *Journal of Virology* **65**, 2211–2219.

Santos, N., Gouvea, V., Timenetsky, M. C., Clark, H. F., Riepenhoff-Talty, M. & Garbarg-Chenon, A. (1994). Comparative analysis of VP8* sequences from rotaviruses possessing M37-like VP4 recovered from children with and without diarrhoea. *Journal of General Virology* **75**, 1775–1780.

Syvänen, A.-C., Aalto-Setälä, K., Kontula, K. & Söderlund, H. (1989). Direct sequencing of affinity-captured amplified human DNA application to the detection of apolipoprotein E polymorphism. *FEBS Letters* **258**, 71–74.

Tabor, S. & Richardson, C. C. (1990). DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Journal of Biological Chemistry* **265**, 8322–8328.

Taniguchi, K., Urasawa, T. & Urasawa, S. (1986). Reactivity pattern to human rotavirus strains of a monoclonal antibody against VP2, a component of the inner capsid of rotavirus. *Archives of Virology* **87**, 135–141.

Wilde, J., Eiden, J. & Yolken, R. (1990). Removal of inhibitory substances from human fecal specimens for detection of group A rotaviruses by reverse transcriptase and polymerase chain reactions. *Journal of Clinical Microbiology* **28**, 1300–1307.

Woods, P. A., Gentsch, J., Gouvea, V., Mata, L., Simhon, A., Santosham, M., Bai, Z., Urasawa, S. & Glass, R. I. (1992). Distribution of serotypes of human rotavirus in different populations. *Journal of Clinical Microbiology* **30**, 781–785.

Xin, K.-Q., Morikawa, S., Fang, Z.-Y., Mukoyama, A., Okuda, K. & Ushijima, H. (1993). Genetic variation in VP7 gene of human rotavirus serotype 1 (G1 type) isolated in Japan and China. *Virology* **197**, 813–816.

Received 2 September 1997; Accepted 21 October 1997