was performed by using the TREE-PUZZLE program version 5.2 (Strimmer & von Haeseler, 1996). ML genetic distances and nucleotide-substitution statistical parameters were estimated under the selected TN93+Γ substitution model with an initial neighbour-joining tree and then the best ML tree was reconstructed with these optimized parameters by using the quartet-puzzling method in the TREE-PUZZLE program. Fifty thousand and one hundred thousand quartet-puzzling steps were performed in constructing trees from the 19 genomic and 43 genomic region sequences, respectively. Sequence similarities and observed distances were calculated by using the Old Distance program in the GCG software package. A nucleotide sequence PLOTSIMILARITY plot across the genome (100 nt window) was generated by using the PLOTSIMILARITY program in the GCG software package. As the sequences of genotype 1a viruses were over-represented, the plot was generated by using six sequences from each clade, including members from each genotype. Nucleotide sequence substitution-rate analysis was carried out with PILEUP (GCG package), fastDNAml (version 1.2.2) and DNArates (version 1.1.0), employing default parameters. To detect recombination, phylogenetic analysis of sequences on either side of putative break points was conducted by using TREE-PUZZLE with the same parameter settings as were used in the genomic sequence analysis. Recombination was also analysed by using the sequence recombination-detection programs TOPALi (Milne *et al.*, 2004), RIP 2.0 (Recombination Identification Program; http://hivweb.lanl.gov/RIP/RIPsubmit.html) and the four-cluster likelihood mapping analysis in the TREE-PUZZLE program.

RESULTS

Genomic sequences and comparisons

A representative rubella virus phylogenetic tree based on the standard E1 gene window recommended by the WHO (nt 8291–9469) and containing the reference viruses for each genotype and the ten viruses for which the genomic sequence has been determined is shown in Fig. 1. As the genomic sequences were from genotype 1a and 2A viruses, the genomic sequences of nine representative viruses from six additional genotypes were determined (Table 1). Among these 19 viruses, with three exceptions, the genomes were 9762 nt in length and consisted (5′–3′) of a 40 nt 5′ UTR, a 6351 nt NSP-ORF, a 120 nt junction region, a 3192 nt SP-ORF and a 59 nt 3′ UTR. All three exceptions were in the junction region: the genome of one of the genotype 1B viruses (GUZ_GER92) was 9760 nt in length because it had a deletion of 2 nt at positions 6480–6481 (between the end of the NSP-ORF and the SG RNA start site) and the genomes of both genotype 2B viruses were 9761 nt in length because they had a deletion of 1 nt at position 6422 (between the SG RNA start site and the start of the SP-ORF). Fig. 2 shows a similarity plot across 12 genomic sequences proportionally representing all eight genotypes. Overall variability averaged approximately 7 % and was roughly comparable across the genome, with the exceptions of the 5′-terminal approximately 400 nt, a region encoding the methyl/guanylyltransferase (MT) domain within the P150 gene that exhibited variability of approximately 4 %, and a region of the P150 gene encompassing nt 2100–2400, the HVR, in which local variability peaked at up to 18 %. Pairwise observed genomic distances between viruses in different genotypes (see Supplementary Table S3, available in JGV Online) ranged from 2.0 to 8.7 %. The range of pairwise observed distances for genomic regions (the five genes and domains within P150 and P90) and the 3′ *cis*-acting element (3′CAE) is given in Table 2. Maximal observed distances of these regions were shown to range from 8.29 to 11.44 %, with the exception of the MT domain (5.24 %) and the HVR (21.18 %). Given this limited degree of variability across most of the genome, it was not suprising that 78 % of the nucleotides in the genome were invariant across the 19 sequences (Table 2).

The additional sequences contributed by this report greatly expand information on the genomic diversity of rubella viruses and, therefore, we took the opportunity to calculate a number of evolutionary parameters (Table 2). The maximum pairwise genetic distance (see Supplementary Table S3, available in JGV Online) was 14.78 substitutions in 100 sites, greater than the largest observed distance (8.74 observed substitutions in 100 sites). The maximal genetic distances for the genomic regions (the five genes, domains within P150 and P90 and the 3′CAE) ranged from 13.97 to 23.0 substitutions in 100 sites, with the exception of the MT domain (6.77 substitutions in 100 sites) and the HVR (35.85 substitutions in 100 sites). The transition/transversion site parameter (K) was 7.04 for the entire genome and ranged among genomic regions from 4.5 (HVR) to 13.35 (3′CAE). The pyrimidine and purine (Y/R) transition parameter of entire genome was 2.7 and varied among the genomic regions from 0.93 (X domain of the P150 gene) to 3.7 (C gene). To test whether the observed distance simply underestimates the genetic distance or whether substitution saturation has been reached, the pairwise number of transitions and transversions was plotted as a function of the calculated genetic distance (by using DAMBE; Xia & Xie, 2001), with the result that both transitions and transversions increased linearly with genetic distance, with the number of transitions being higher than transversions (data not shown). Neither reached a plateau, indicating that substitution saturation had not occurred.

Among the 19 genomic sequences, 78 % of the nucleotides were invariant. Not surprisingly, the parameter of rate heterogeneity, α, was 0.22 for the entire genome and varied between 0.19 and 0.33 across the genomic regions, with the exception of the HVR, within which $\alpha = 1.35$. These small α values indicated a strong substitution-rate heterogeneity among nucleotide sites across most of the genome (i.e. more than three-quarters of the nucleotides remained constant, whilst fewer than one-quarter exhibited variability). Within the HVR, 46 % of the nucleotides were variable.

Phylogenetic analysis

ML phylogenetic trees constructed from the complete genomic sequences, as well as from the NSP- and SP-ORFs, are displayed in Fig. 3. As in the E1-based tree, the six clade 2 sequences formed a clear, consistent branching pattern with high support values in all three trees, indicating that

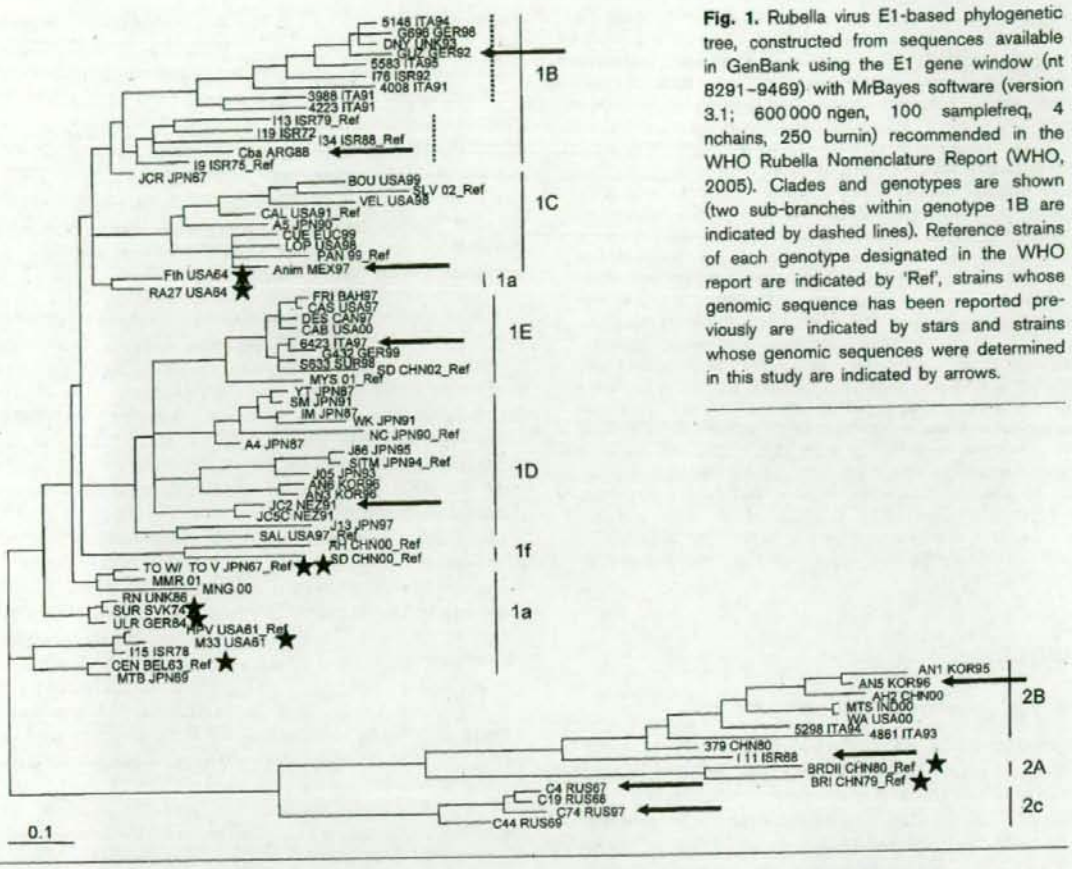

Fig. 1. Rubella virus E1-based phylogenetic tree, constructed from sequences available in GenBank using the E1 gene window (nt 8291–9469) with MrBayes software (version 3.1; 600 000 ngen, 100 samplefreq, 4 nchains, 250 burnin) recommended in the WHO Rubella Nomenclature Report (WHO, 2005). Clades and genotypes are shown (two sub-branches within genotype 1B are indicated by dashed lines). Reference strains of each genotype designated in the WHO report are indicated by 'Ref', strains whose genomic sequence has been reported previously are indicated by stars and strains whose genomic sequences were determined in this study are indicated by arrows.

genotypes 2A and 2B are related more closely to each other than to genotype 2c. In clade 1, the groupings of genotypes 1B, 1C, 1D and 1E on the three trees were consistent: the two genotype 1B sequences formed a branch, as did the individual genotype 1D and 1E sequences, whilst the individual genotype 1C sequence extended from the baseline with no relative relationship to other genotypes. On the genomic and SP-ORF trees, the eight genotype 1a sequences grouped into four pairs, indicated in Fig. 3 as a1 (TO-w and TO-v; a wild-type parent and the attenuated vaccine derived from it), a2 (Fth and RA27/3; both isolated in the north-eastern USA in 1964), a3 (CEN and M33; isolated from Europe and the USA in 1961–1962) and a4 (SUR and ULR; both isolated from eastern Europe in 1974 and 1984). Interestingly, on the NSP-ORF tree, M33 separated from CEN (a3) and clustered with the a4 grouping. We also constructed trees by using the sequences of the genes and regions within the NSP-ORF and SP-ORF (data not shown), with the result that they had the same general topology as the ORF-generated trees. The exception was the HVR-generated tree, on which each of the clade 1 viruses formed an individual branch, apart from the a1 and M33-SUR groupings, which were preserved.

Extensive phylogenetic analysis has not been done previously using sequences from 5′ regions of the genome. To do so, trees were constructed from the sequences of the non-structural protease (NP) region (nt 3035–3973; Fig. 2), the junction region and adjacent sequences (JR; nt 6351–6829, which includes the 3′ end of the P90 gene, the UTR between the ORFs and the 5′ end of the C gene) and the E1 gene [nt 8731–9469, the recommended window for routine genotyping (WHO, 2005)] of 43 viruses representing eight genotypes. ML phylogenetic trees constructed from these sequences are shown in Fig. 4(a). Clustering of viruses on the three trees was similar, with the exception of a group of seven genotype 1B viruses that formed a single branch on the NP tree and two branches (one of five and one of two viruses) on the JR tree, but did not form a cluster on the E1 tree. The single-nucleotide deletion at nt 6422 in the junction region detected in the genome sequence of the two genotype 2B viruses was confirmed in the JR sequences determined from three additional genotype 2B viruses, and the 2 nt deletion at nt 6480–6481 detected in the genome sequence of one of the two genotype 1B viruses was discovered in the JR sequences determined from four of the five additional genotype 1B viruses. Interestingly, this 2 nt

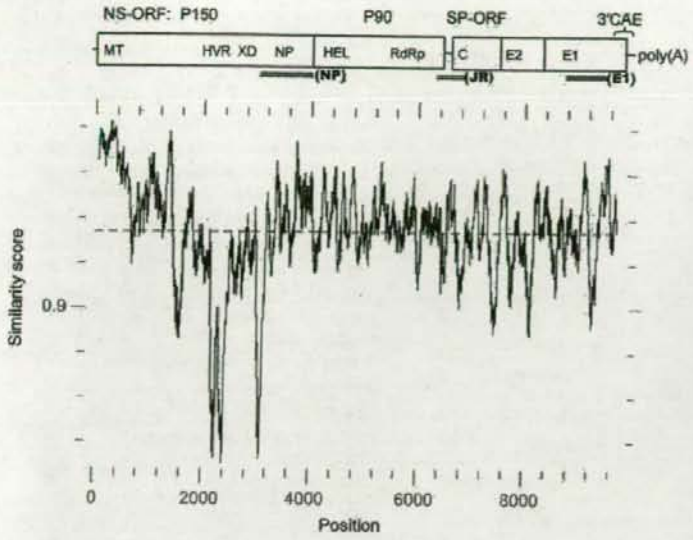

Fig. 2. Rubella virus genome nucleotide-similarity plot. The plot was generated from 12 genome sequences proportionally representing the two clades and eight genotypes by using the PLOTSIMILARITY program in the GCG software package with a 100 nt window. A genomic map is overlaid above the plot that shows the two ORFs, five genes, regions/domains within the NS-ORF (MT, methyl/guanylyltransferase; HVR, hypervariable region; XD, X domain; NP, non-structural protease; HEL, helicase; RdRp, RNA-dependent RNA polymerase or replicase) and the 3′ *cis*-acting elements (3′CAE). The regions from which sequences were determined from 43 viruses for comprehensive phylogenetic analysis (see Fig. 5) are also denoted (NP, non-structural protease; JR, junction region; E1, WHO standard E1 window).

Table 2. Maximum-likelihood parameters calculated for regions of the rubella virus genome

Region	G+C content (mol%)	No. sites	No. invariant sites (%)	Observed distance (range)*	Genetic distance (range)*	K*	Y/R*	α*
Genome†	69.5	9762	7611 (78.0)	0.22–8.74	0.22–14.78	7.04	2.7	0.22
P150†	71.8	3903	3033 (77.7)	0.26–9.3	0.26–15.89	5.97	2.25	0.23
P150: MT†	67.4	210	189 (90.0)	0–5.24	0–6.77	9.9	3.05	0.23
P150: HVR†	81.1	321	174 (54.2)	2.49–21.18	0.31–35.85	4.5	1.78	1.35
P150: XD†	74.2	507	377 (74.4)	0–11.44	0–23	7.2	0.93	0.24
P150: NP†	72.7	900	700 (77.8)	0.33–10	0.34–17.04	6.31	2.4	0.24
P90†	67.2	2448	1952 (79.7)	0.12–8.29	0.12–13.97	10.67	3.08	0.21
P90: HEL†	66.8	756	596 (78.8)	0.4–8.73	0.4–16.9	9.95	3.39	0.26
P90: RdRp†	66.9	1563	1250 (80.0)	0–8.83	0–15.3	11.06	2.74	0.19
C†	72.8	900	697 (77.4)	0.11–9.44	0.11–17.59	7.26	3.7	0.26
E2†	71	846	625 (73.9)	0–10.05	0–15.86	6	3.13	0.33
E1†	66.2	1446	1153 (79.7)	0.14–9.2	0.14–15.49	7.6	2.82	0.19
3′CAE†	64.9	307	244 (79.5)	0–8.79	0–14.67	13.35	2.98	0.2
NP‡	73	939	670 (71.4)	0–10.44	0–16.69	5.98	2.35	0.28
JR‡[c]	69.9	479	334 (69.7)	0–10.67	0–16.66	5.43	2.69	0.32
E1‡	67.5	739	546 (73.9)	0–10.42	0–21.06	7.74	3.42	0.21

*Parameters were calculated by using the TREE-PUZZLE program with setting of the TN93 substitution model with gamma discrete distribution-rate heterogeneity. K, Transitions/transversions; Y/R, pyrimidine transitions (T↔C)/purine transitions (A↔G); α, rate-heterogeneity shape parameter. .

†Parameters were calculated from the indicated regions by using the 19 genomic sequences. P150: nt 41–3943; MT: methyltransferase, nt 230–439; HVR: hypervariable region, nt 2120–2440; XD: X domain, nt 2492–2998; NP: non-structural protease, nt 3041–3940; P90: nt 3944–6391; HEL: helicase region, nt 4043–4798; RdRp: RNA-dependent RNA polymerase region, nt 4826–6388; C, nt 6512–7411; E2: nt 7412–8257; E1: nt 8258–9703; 3′CAE: 3′ *cis*-acting elements, nt 9456–9762.

‡Parameters were calculated from the sequences of the NP (nt 3035–3973), JR (nt 6351–6829) and E1 (nt 8731–9469) regions determined for 43 viruses.

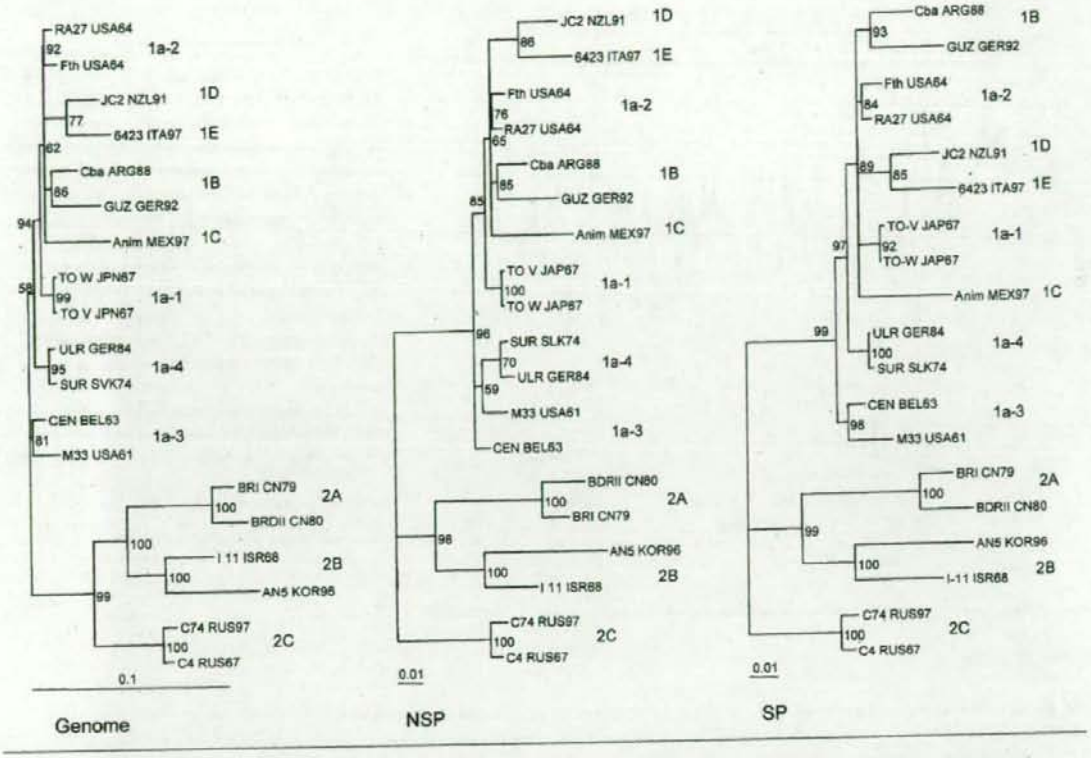

Fig. 3. Phylogenetic trees of genomic, NS-ORF and SP-ORF sequences. Trees were constructed by using TREE-PUZZLE (version 5.2) with 50 000 puzzle steps; reliability values are indicated on each node. Genotypes are denoted, including four subclusters of genotype 1a (1a-1 to 1a-4). Genetic distance (substitutions in 100 nt) calculated by using the TN93+Γ substitution model is indicated by the bar below each tree.

deletion did not co-segregate with the JR sequence of these viruses, as it was present in three members of the five-virus genotype 1B branch and both members of the two-virus genotype 1B branch [marked by an asterisk in Fig. 4(a)]. Evolutionary parameters calculated from these larger and more genotypically representative sequence sets, shown in Table 2, were similar to those calculated by using the smaller genome sequence set.

Detection of genomic recombination among genotype 1B viruses

The lack of co-segregation of the genotype 1B JR sequences and the 2 nt deletion led us to hypothesize that a recombinational event had occurred in this region of the genome of these viruses, at or downstream of the deletion. To test this hypothesis, we expanded the sequence determined upstream into the 3′ end of the P90 gene (the RdRp domain) and employed several software programs designed to detect recombination events, as well as secondary phylogenetic analysis, to detect putative break points. The RIP program predicted a break point at nt 6555, within the 5′ end of the C gene (which begins at nt 6512). As shown in Fig. 4(b), this prediction is supported by trees of sequences up- and downstream of this break point. Similar to the NP tree in Fig. 4(a), all seven genotype 1B sequences form a branch on the tree constructed by using sequences upstream of this break point (nt 5720–6554). On the tree constructed by using sequences downstream of the break point (nt 6555–6814), two sequences (TOM_UNK86 and 0754_GER92) were on a branch distinct from the other genotype 1B sequences, similar to the JR tree in Fig. 4(a). We thus conclude that a recombinational event occurred at or near this site during the evolution of these viruses.

DISCUSSION

The goal of this study was to expand the rubella virus genomic sequence database to include viruses in the majority of the currently defined genotypes. Whilst ten genomic sequences had been reported previously, only two of the ten currently defined genotypes were represented. This study added nine genomic sequences representing an additional six genotypes, encompassing the most widely

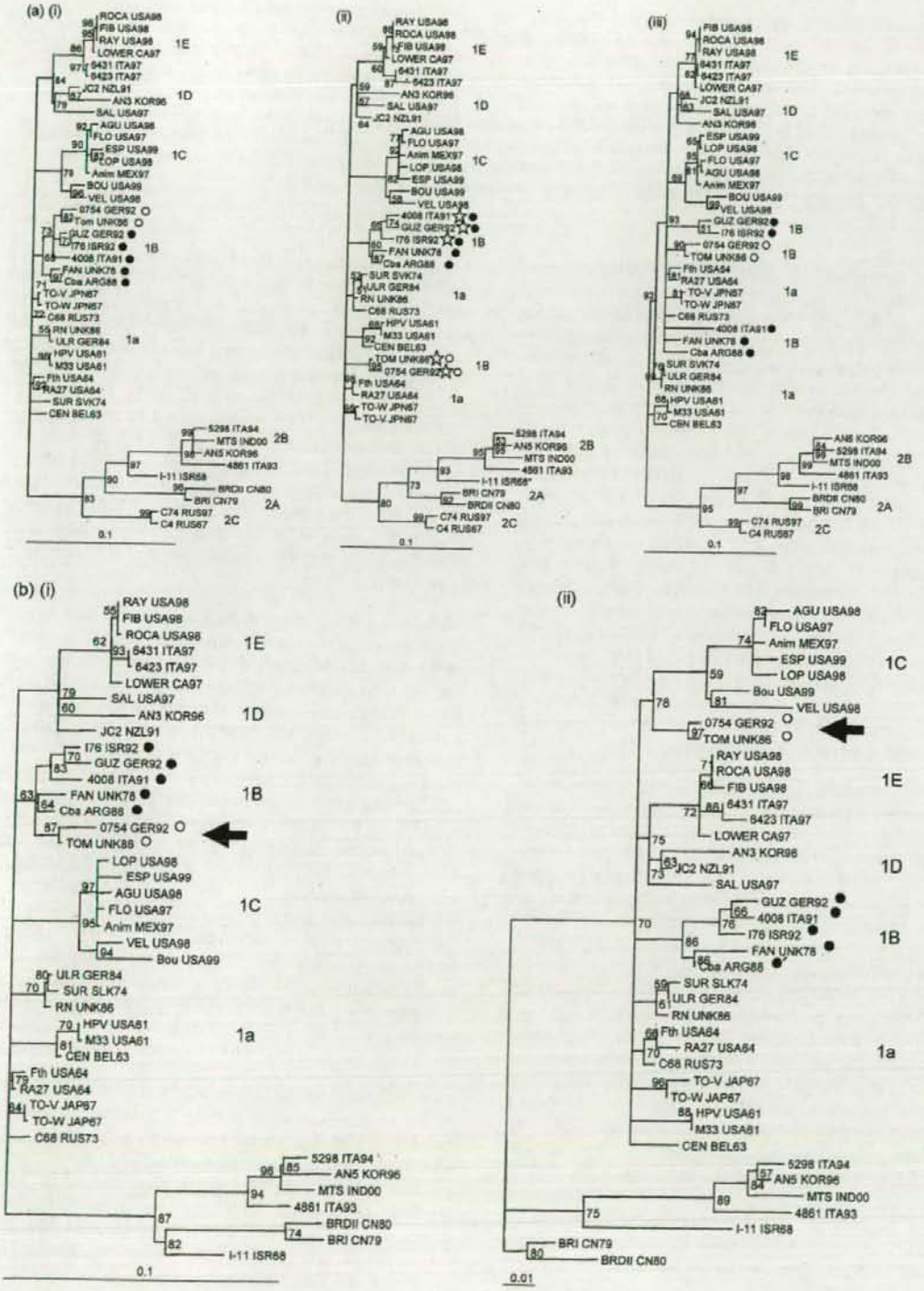
(a) (i)
ROCA USA98
FIB USA98
RAY USA98
LOWER CA97
6431 ITA97
6423 ITA97
JC2 NZL91
AN3 KOR96
SAL USA97
AGU USA98
FLO USA97
ESP USA99
LOP USA98
Anim MEX97
BOU USA99
VEL USA98
0754 GER92
Tom UNK86
GUZ GER92
I76 ISR92
4008 ITA91
FAN UNK75
Cba ARG88
TO-V JPN67
TO-W JPN67
C68 RUS73
RN UNK86
ULR GER84
HPV USA61
M33 USA61
Fth USA64
RA27 USA64
SUR SVK74
CEN BEL63
5298 ITA94
MTS IND00
AN5 KOR96
4861 ITA93
I-11 ISR68
BRDII CN80
BRI CN79
C74 RUS97
C4 RUS67
1E
1D
1C
1B
1a
2B
2A
2C
0.1
(ii)
RAY USA98
ROCA USA98
FIB USA98
LOWER CA97
6431 ITA97
6423 ITA97
AN3 KOR96
SAL USA97
JC2 NZL91
AGU USA98
FLO USA97
Anim MEX97
LOP USA98
ESP USA99
BOU USA99
VEL USA98
4008 ITA91
GUZ GER92
I76 ISR92
FAN UNK78
Cba ARG88
SUR SVK74
ULR GER84
RN UNK86
C68 RUS73
HPV USA61
M33 USA61
CEN BEL63
TOM UNK86
0754 GER92
Fth USA64
RA27 USA64
TO-W JPN67
TO-V JPN67
5298 ITA94
AN5 KOR96
MTS IND00
4861 ITA93
I-11 ISR68
BRI CN79
BRDII CN80
C74 RUS97
C4 RUS67
1E
1D
1C
1B
1a
2B
2A
2C
0.1
(iii)
FIB USA98
ROCA USA98
RAY USA98
6431 ITA97
6423 ITA97
LOWER CA97
JC2 NZL91
SAL USA97
AN3 KOR96
ESP USA99
LOP USA98
FLO USA97
AGU USA98
Anim MEX97
BOU USA99
VEL USA98
GUZ GER92
I76 ISR92
0754 GER92
TOM UNK86
Fth USA64
RA27 USA64
TO-V JPN67
TO-W JPN67
C68 RUS73
4008 ITA91
FAN UNK78
Cba ARG88
SUR SVK74
ULR GER84
RN UNK86
HPV USA61
M33 USA61
CEN BEL63
AN5 KOR96
5298 ITA94
MTS IND00
4861 ITA93
I-11 ISR68
BRDII CN80
BRI CN79
C74 RUS97
C4 RUS67
1E
1D
1C
1B
1a
2B
2A
2C
0.1
(b) (i)
RAY USA98
FIB USA98
ROCA USA98
6431 ITA97
6423 ITA97
LOWER CA97
SAL USA97
AN3 KOR96
JC2 NZL91
I76 ISR92
GUZ GER92
4008 ITA91
FAN UNK78
Cba ARG88
0754 GER92
TOM UNK86
LOP USA98
ESP USA99
AGU USA98
FLO USA97
Anim MEX97
VEL USA98
Bou USA99
ULR GER84
SUR SLK74
RN UNK86
HPV USA61
M33 USA61
CEN BEL63
Fth USA64
RA27 USA64
TO-V JAP67
TO-W JAP67
C68 RUS73
5298 ITA94
AN5 KOR96
MTS IND00
4861 ITA93
BRDII CN80
BRI CN79
I-11 ISR68
1E
1D
1B
1C
1a
0.1
(ii)
AGU USA98
FLO USA97
Anim MEX97
ESP USA99
LOP USA98
Bou USA99
VEL USA98
0754 GER92
TOM UNK86
RAY USA98
ROCA USA98
FIB USA98
6431 ITA97
6423 ITA97
LOWER CA97
AN3 KOR96
JC2 NZL91
SAL USA97
GUZ GER92
4008 ITA91
I76 ISR92
FAN UNK78
Cba ARG88
SUR SLK74
ULR GER84
RN UNK86
Fth USA64
RA27 USA64
C68 RUS73
TO-V JAP67
TO-W JAP67
HPV USA61
M33 USA61
CEN BEL63
5298 ITA94
AN5 KOR96
MTS IND00
4861 ITA93
I-11 ISR68
BRI CN79
BRDII CN80
1C
1E
1D
1B
1a
0.01

Fig. 4. Phylogenetic trees based on sequence of genomic regions. (a) Trees were constructed using the sequence of the (i) non-structural protease (NP, nt 3035–3973), (ii) junction region (JR, nt 6351–6829, including the 3′ end of the NS-ORF, the UTR between the ORFs and the 5′ end of the C gene) and (iii) WHO standard E1 window (nt 8731–9469). On all three trees, genotype 1B viruses are indicated by dots (filled or empty for the clusters of five or two viruses, respectively, on the JR-based tree), and in (ii), genotype 1B viruses with a 2 nt deletion at nt 6480–6481 are indicated by stars. (b) Trees were constructed by using nt 5720–6554 (i) or 6555–6814 (ii), on either side of the putative recombination break point. Differential segregation of two of the genotype 1B viruses is indicated by arrows. All trees were constructed with TREE-PUZZLE (version 5.2; 100 000 puzzle steps). Reliability values are indicated on each node and genetic distance (substitutions in 100 nt) calculated by using the TN93+Γ substitution model is indicated by the bar below each tree.

divergent genotypes. The most striking finding was the genomic uniformity of rubella viruses, as it was discovered that 78 % of the nucleotides in the genomes of the viruses from the eight genotypes were invariant and these viruses preserved identical genomic dimensions across the two ORFs and two of the three UTRs. Only in the junction region (the UTR between the ORFs) of genotype 2B viruses, which had a 1 nt deletion, and a subset of genotype 1B viruses, which had a 2 nt deletion, was any plasticity observed. Such strict uniformity of genomic topology is highly unusual among RNA viruses (Huang *et al.*, 2004; Kang *et al.*, 2004; Kinney *et al.*, 1998; Saleh *et al.*, 2003; Takahashi *et al.*, 2003; Tarbatt *et al.*, 1997; van Cuyck *et al.*, 2003; Yang *et al.*, 2004). Sequence diversity was also low; across the eight genotypes, the maximum observed distance was <9 % and the maximum calculated genetic distance was 14.8 substitutions in 100 sites. Regardless of this difference, substitution saturation had not occurred, indicating that, despite the limited sequence diversity among rubella viruses, sufficient phylogenetic signal was retained to support the groupings observed (Salemi & Vandamme, 2003; Xia, 2000; Xia & Xie, 2001).

A sequence-similarity profile revealed, with two exceptions, a comparable pattern across the genome with local windows of similarity and dissimilarity varying about a relatively uniform mean, indicating that most genomic regions, including both virion protein and replicase protein genes, were equally divergent. Both observed and genetic distances between these genomic regions were comparable. The two exceptions were both within the P150 gene, with the N-terminal MT domain exhibiting less variability and the internal HVR exhibiting greater variability. Although the MT domain was predicted to encode both methyl- and guanylyltransferase activities (Rozanov *et al.*, 1992; neither activity has been demonstrated experimentally), the fact that 90 % of the nucleotide residues within this region are conserved raises the possibility that this region serves as a CAE in addition to encoding protein sequence. Consistent with this possibility, the phenotype of a cell culture-potentiating mutation discovered at nt 164 of the RA27/3 genome was found to be due to the nucleotide itself rather than to the encoded amino acid (Pugachev *et al.*, 2000). Conservation of the MT domain sequence has also been observed in other alpha-like family viruses (Gouvea *et al.*, 1998). The HVR encodes a proline- and arginine-rich domain of P150 termed the 'proline hinge' (Koonin *et al.*, 1992), although this domain contains several adaptor motifs that could serve to facilitate the association of P150 with other proteins. If this domain serves as a structural hinge between functional domains within the P150 protein, this could explain the lower constraint on sequence conservation within the HVR in comparison with the rest of the genome. On the other hand, Hofmann *et al.* (2003) reported data suggesting that the HVR among clade 1 viruses was under positive selection at the amino acid level. It should be pointed out that 'hypervariable region' is a relative term, in that HVRs in the genomes of other viruses are often more variable than the rubella virus HVR. For example, in the hepatitis E virus HVR, variability is >50 % (Arankalle *et al.*, 1999; Gouvea *et al.*, 1998; Nishizawa *et al.*, 2003; van Cuyck *et al.*, 2003).

With the exception of the HVR, nucleotide residues or sites across the genome showed a strong heterogeneity in rate of divergence, as indicated by the low value of the rate-heterogeneity parameter α. Sequence collections with low α values exhibit an L-shaped distribution on a graph of number of sites versus rate of divergence, rather than the bell-shaped curve generated when α is 1 (or >1). The low α value reflects the fact that roughly 80 % of the residues in the rubella virus genome were invariant in this collection of genomic sequences. The percentage of invariant residues at first and second codon positions was 93 %, compared with 48 % at third codon positions (Y. Zhou, unpublished data), and thus maintenance of amino acid sequence is a substantial component of the conservation of nucleotide sequence. Among third codon positions, the G+C content was 81 mol%, compared with 63 mol% among first and second codon positions (Y. Zhou, unpublished data), and thus there was selection for G and C residues. This selection was also evident in the HVR, within which the G+C content was 81 mol%, compared with 70 mol% for the genome.

Among the nucleotide substitutions at the 20 % of genomic sites that exhibited variability, transitions were strikingly more abundant than transversions; across the genome, the transition to transversion ratio, K, was 7.0 and varied among genomic regions from 4.5 to 13.4. Thus, the rubella virus genome exhibited the transition over transversion preference that has been well documented in DNA genomes (Meyer *et al.*, 1999; Salemi & Vandamme, 2003). This preference has been attributed to the facts that transitions are more likely to lead to silent mutations in amino acid

sequence and that, during replication, it is more likely that a mutation to a nucleotide of equal size (transition) will occur than to a nucleotide of different size (transversion). In RNA genomes, the possibility of both G–C and G–U pairing would also favour transitions in the replication process. Interestingly, pyrimidine transitions were favoured over purine transitions by a ratio (Y/R) of 2.7 across the entire genome; Y/R varied from 0.9 to 3.7 in genomic regions. Within the HVR, the most variable region of the genome, both K and Y/R were lower than for the entire genome and most of the other genomic regions, indicating that the variability in this region was generated in part by relaxing of the genomic preference for pyrimidine transitions over transversions and purine transitions.

Phylogenetic analysis of rubella viruses has traditionally been done on the basis of E1 gene or subE1 gene sequences and a standard taxonomy was proposed recently, based on a window within the E1 gene, that was substantiated by using complete SP-ORF sequences (WHO, 2005). The second goal of this study was to extend phylogenetic analysis to the 5′ region of the genome and we found that generally comparable trees, in terms of both overall variability and phylogenetic clustering, were generated with sequence windows in the NSP-ORF. The exception was a group of seven genotype 1B viruses that formed a branch in a tree based on NP sequence, but formed two branches on the basis of JR sequence. Intriguingly, a deletion in the junction region of five of these seven viruses did not segregate with the two phylogenetic branches on the JR tree. Analysis revealed a recombinational event, putatively near the 5′ end of the C gene, that led to the generation of the two branches on the JR tree. There was one previous report of a natural recombination event in *Rubella virus* (in the E1 gene; Zheng *et al.*, 2003a), but the origin of the recombinant strain was in doubt because one of the parents was related closely to a commonly used laboratory strain. Thus, this was the first conclusive evidence of rubella virus recombination in nature.

Interestingly, the E1 sequences of the seven genotype 1B viruses did not cluster on the E1-based tree and, in comparison with the NP- and JR-based trees, this could be due to divergence or additional recombinational events. As can be seen in the tree in Fig. 1, genotype 1B consists of two sub-branches that would not necessarily appear to be related if fewer sequences were employed (e.g. the E1-based tree in Fig. 4). It is also to be noted that all of the WHO reference strains are on one of these sub-branches. Thus, for this genotype, phylogenetic analysis using sequences from the NSP-ORF region of the genome could be useful in assessing relatedness.

ACKNOWLEDGEMENTS

This research was supported by a grant from the National Institutes of Health (AI21389). We thank Xianfeng Chen for software assistance, Duping Zheng, Suganthi Suppiah and Hui Zhao for preliminary sequence determinations and Ping Jiang for processing sequencing reactions and gels.

REFERENCES

Arankalle, V. A., Paranjape, S., Emerson, S. U., Purcell, R. H. & Walimbe, A. M. (1999). Phylogenetic analysis of hepatitis E virus isolates from India (1976–1993). *J Gen Virol* **80**, 1691–1700.

Bosma, T. J., Best, J. M., Corbett, K. M., Banatvala, J. E. & Starkey, W. G. (1996). Nucleotide sequence analysis of a major antigenic domain of the E1 glycoprotein of 22 rubella virus isolates. *J Gen Virol* **77**, 2523–2530.

Chantler, J. K., Wolinsky, J. S. & Tingle, A. (2001). Rubella virus. In *Fields Virology*, 4th edn, pp. 963–990. Edited by D. M. Knipe & P. M. Howley. Philadelphia, PA: Lippincott Williams & Wilkins.

Clarke, D. M., Loo, T. W., Hui, I., Chong, P. & Gillam, S. (1987). Nucleotide sequence and in vitro expression of rubella virus 24S subgenomic messenger RNA encoding the structural proteins E1, E2 and C. *Nucleic Acids Res* **15**, 3041–3057.

Dominguez, G., Wang, C. Y. & Frey, T. K. (1990). Sequence of the genome RNA of rubella virus: evidence for genetic rearrangement during togavirus evolution. *Virology* **177**, 225–238.

Donadio, F. F., Siqueira, M. M., Vyse, A., Jin, L. & Oliveira, S. A. (2003). The genomic analysis of rubella virus detected from outbreak and sporadic cases in Rio de Janeiro state, Brazil. *J Clin Virol* **27**, 205–209.

Frey, T. K. (1994). Molecular biology of rubella virus. *Adv Virus Res* **44**, 69–160.

Frey, T. K., Abernathy, E. S., Bosma, T. J., Starkey, W. G., Corbett, K. M., Best, J. M., Katow, S. & Weaver, S. C. (1998). Molecular analysis of rubella virus epidemiology across three continents, North America, Europe, and Asia, 1961-1997. *J Infect Dis* **178**, 642–650.

Gouvea, V., Snellings, N., Popek, M. J., Longer, C. F. & Innis, B. L. (1998). Hepatitis E virus: complete genome sequence and phylogenetic analysis of a Nepali isolate. *Virus Res* **57**, 21–26.

Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J Mol Biol* **243**, 574–578.

Hofmann, J., Renz, M., Meyer, S., von Haeseler, A. & Liebert, U. G. (2003). Phylogenetic analysis of rubella virus including new genotype I isolates. *Virus Res* **96**, 123–128.

Huang, F. F., Sun, Z. F., Emerson, S. U., Purcell, R. H., Shivaprasad, H. L., Pierson, F. W., Toth, T. E. & Meng, X. J. (2004). Determination and analysis of the complete genomic sequence of avian hepatitis E virus (avian HEV) and attempts to infect rhesus monkeys with avian HEV. *J Gen Virol* **85**, 1609–1618.

Icenogle, J. P., Frey, T. K., Abernathy, E., Reef, S. E., Schnurr, D. & Stewart, J. A. (2006). Genetic analysis of rubella viruses found in the United States between 1966 and 2004: evidence that indigenous rubella viruses have been eliminated. *Clin Infect Dis* **43** (Suppl. 3), S133–S140.

Kakizawa, J., Nitta, Y., Yamashita, T., Ushijima, H. & Katow, S. (2001). Mutations of rubella virus vaccine TO-336 strain occurred in the attenuation process of wild progenitor virus. *Vaccine* **19**, 2793–2802.

Kang, S. Y., Yun, S. I., Park, H. S., Park, C. K., Choi, H. S. & Lee, Y. M. (2004). Molecular characterization of PL97-1, the first Korean isolate of the porcine reproductive and respiratory syndrome virus. *Virus Res* **104**, 165–179.

Katow, S. (2004). Molecular epidemiology of rubella virus in Asia: utility for reduction in the burden of diseases due to congenital rubella syndrome. *Pediatr Int* **46**, 207–213.

Katow, S., Minahara, H., Fukushima, M. & Yamaguchi, Y. (1997a). Molecular epidemiology of rubella by nucleotide sequences of the rubella virus E1 gene in three East Asian countries. *J Infect Dis* **176**, 602–616.

Katow, S., Minahara, H., Ota, T. & Fukushima, M. (1997b). Identification of strain-specific nucleotide sequences in E1 and NS4 genes of rubella virus vaccine strains in Japan. *Vaccine* **15**, 1579–1585.

Kinney, R. M., Pfeffer, M., Tsuchiya, K. R., Chang, G. J. & Roehrig, J. T. (1998). Nucleotide sequences of the 26S mRNAs of the viruses defining the Venezuelan equine encephalitis antigenic complex. *Am J Trop Med Hyg* **59**, 952–964.

Koonin, E. V., Gorbalenya, A. E., Purdy, M. A., Rozanov, M. N., Reyes, G. R. & Bradley, D. W. (1992). Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. *Proc Natl Acad Sci U S A* **89**, 8259–8263.

Lund, K. D. & Chantler, J. K. (2000). Mapping of genetic determinants of rubella virus associated with growth in joint tissue. *J Virol* **74**, 796–804.

Meyer, S., Weiss, G. & von Haeseler, A. (1999). Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* **152**, 1103–1110.

Milne, I., Wright, F., Rowe, G., Marshall, D. F., Husmeier, D. & McGuire, G. (2004). TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics* **20**, 1806–1807.

Nishizawa, T., Takahashi, M., Mizuo, H., Miyajima, H., Gotanda, Y. & Okamoto, H. (2003). Characterization of Japanese swine and human hepatitis E virus isolates of genotype IV with 99% identity over the entire genome. *J Gen Virol* **84**, 1245–1251.

Pugachev, K. V., Abernathy, E. S. & Frey, T. K. (1997). Genomic sequence of the RA27/3 vaccine strain of rubella virus. *Arch Virol* **142**, 1165–1180.

Pugachev, K. V., Galinski, M. S. & Frey, T. K. (2000). Infectious cDNA clone of the RA27/3 vaccine strain of Rubella virus. *Virology* **273**, 189–197.

Reef, S. E., Frey, T. K., Theall, K., Abernathy, E., Burnett, C. L., Icenogle, J., McCauley, M. M. & Wharton, M. (2002). The changing epidemiology of rubella in the 1990s: on the verge of elimination and new challenges for control and prevention. *JAMA* **287**, 464–472.

Robertson, S. E., Featherstone, D. A., Gacic-Dobo, M. & Hersh, B. S. (2003). Rubella and congenital rubella syndrome: global update. *Rev Panam Salud Publica* **14**, 306–315.

Rozanov, M. N., Koonin, E. V. & Gorbalenya, A. E. (1992). Conservation of the putative methyltransferase domain: a hallmark of the 'Sindbis-like' supergroup of positive-strand RNA viruses. *J Gen Virol* **73**, 2129–2134.

Saitoh, M., Shinkawa, N., Shimada, S., Segawa, Y., Sadamasu, K., Hasegawa, M., Kato, M., Kozawa, K., Kuramoto, T. & other authors (2006). Phylogenetic analysis of envelope glycoprotein (E1) gene of rubella viruses prevalent in Japán in 2004. *Microbiol Immunol* **50**, 179–185.

Saleh, S. M., Poidinger, M., Mackenzie, J. S., Broom, A. K., Lindsay, M. D. & Hall, R. A. (2003). Complete genomic sequence of the Australian south-west genotype of Sindbis virus: comparisons with other Sindbis strains and identification of a unique deletion in the 3′-untranslated region. *Virus Genes* **26**, 317–327.

Salemi, M. & Vandamme, A.-M. (2003). *The Phylogenetic Handbook: a Practical Approach to DNA and Protein Phylogeny.* Cambridge: Cambridge University Press.

Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol* **13**, 964–969.

Takahashi, K., Kang, J. H., Ohnishi, S., Hino, K., Miyakawa, H., Miyakawa, Y., Maekubo, H. & Mishiro, S. (2003). Full-length sequences of six hepatitis E virus isolates of genotypes III and IV from patients with sporadic acute or fulminant hepatitis in Japan. *Intervirology* **46**, 308–318.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512–526.

Tarbatt, C. J., Glasgow, G. M., Mooney, D. A., Sheahan, B. J. & Atkins, G. J. (1997). Sequence analysis of the avirulent, demyelinating A7 strain of Semliki Forest virus. *J Gen Virol* **78**, 1551–1557.

van Cuyck, H., Juge, F. & Roques, P. (2003). Phylogenetic analysis of the first complete hepatitis E virus (HEV) genome from Africa. *FEMS Immunol Med Microbiol* **39**, 133–139.

WHO (2005). Standardization of the nomenclature for genetic characteristics of wild-type rubella viruses. *Wkly Epidemiol Rec* **80**, 126–132.

Xia, X. (2000). *Data Analysis in Molecular Biology and Evolution.* Boston: Kluwer Academic Publishers.

Xia, X. & Xie, Z. (2001). DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* **92**, 371–373.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**, 306–314.

Yang, D. K., Kim, B. H., Kweon, C. H., Kwon, J. H., Lim, S. I. & Han, H. R. (2004). Molecular characterization of full-length genome of Japanese encephalitis virus (KV1899) isolated from pigs in Korea. *J Vet Sci* **5**, 197–205.

Zheng, D. P., Frey, T. K., Icenogle, J., Katow, S., Abernathy, E. S., Song, K. J., Xu, W. B., Yarulin, V., Desjatskova, R. G. & other authors (2003a). Global distribution of rubella virus genotypes. *Emerg Infect Dis* **9**, 1523–1530.

Zheng, D. P., Zhou, Y. M., Zhao, K., Han, Y. R. & Frey, T. K. (2003b). Characterization of genotype II rubella virus strains. *Arch Virol* **148**, 1835–1850.

Zheng, D. P., Zhu, H., Revello, M. G., Gerna, G. & Frey, T. K. (2003c). Phylogenetic analysis of rubella virus isolated during a period of epidemic transmission in Italy, 1991-1997. *J Infect Dis* **187**, 1587–1597.

Microbiol. Immunol., 51(1), 117–125, 2007

Editor-Communicated Paper

Tandem Repeats of Lactoferrin-Derived Anti-Hepatitis C Virus Peptide Enhance Antiviral Activity in Cultured Human Hepatocytes

Ken-ichi Abe[1,*], Akito Nozaki[1,*], Kazushi Tamura[2], Masanori Ikeda[1], Kazuhito Naka[1], Hiromichi Dansako[1], Hiro-o Hoshino[2], Katsuaki Tanaka[3], and Nobuyuki Kato*[,1]

[1]*Department of Molecular Biology, Okayama University Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, Okayama, Okayama 700–8558, Japan,* [2]*Department of Virology and Preventive Medicine, Gunma University Graduate School of Medicine, Maebashi, Gunma 371–8511, Japan, and* [3]*Gastroenterological Center, Yokohama City University of Medical Center, Yokohama, Kanagawa 236–0004, Japan*

Communicated by Dr. Masanobu Ohuchi: Received November 8, 2006. Accepted November 13, 2006

Abstract: **Previously, we found that bovine and human lactoferrin (LF) specifically inhibited hepatitis C virus (HCV) infection in cultured non-neoplastic human hepatocyte-derived PH5CH8 cells, and we identified 33 amino acid residues (termed C-s3-33; amino acid 600–632) from human LF that were primarily responsible for the binding activity to the HCV E2 envelope protein and for the inhibiting activity against HCV infection. Since the anti-HCV activity of C-s3-33 was weaker than that of human LF, we speculated that an increase of E2 protein-binding activity might contribute to the enhancement of anti-HCV activity. To test this possibility, we made two repeats [$(C\text{-}s3\text{-}33)_2$] and three repeats [$(C\text{-}s3\text{-}33)_3$] of C-s3-33 and characterized them. Far-Western blot analysis revealed that the E2 protein-binding activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ became stronger than that of the C-s3-33, and that the binding activity of $(C\text{-}s3\text{-}33)_3$ was stronger than that of $(C\text{-}s3\text{-}33)_2$. Using an HCV infection system in PH5CH8 cells, we demonstrated that the anti-HCV activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ became stronger than that of the C-s3-33. Furthermore, using a recently developed infection system with a VSV pseudotype harboring the green fluorescent protein gene and the native E1 and E2 genes, we demonstrated that the antiviral activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ were stronger than that of C-s3-33. These results suggest that tandem repeats of LF-derived anti-HCV peptide are useful as anti-HCV reagents.**

Key words: **Hepatitis C virus, Lactoferrin, Anti-HCV peptide, E2 protein-binding activity**

Hepatitis C virus (HCV) infection frequently causes chronic hepatitis, which progresses to liver cirrhosis and hepatocellular carcinoma (28). HCV is an enveloped positive single-stranded RNA (9.6 kb) virus belonging to the *Flaviviridae*. The HCV genome encodes a large polyprotein precursor of about 3,000 amino acids (aa), which is cleaved by the host and viral proteases into at least 10 proteins: the core, envelope 1 (E1), E2, p7, and non-structural 2 (NS2), NS3, NS4A, NS4B, NS5A, and NS5B (7, 8, 18). These HCV proteins function not only in virus replication but may also affect a variety of cellular functions, including gene expression, signal transduction, and apoptosis (1, 17).

Approximately 170 million people worldwide are infected with HCV (32). The combination of a pegylated interferon with ribavirin is the current standard therapy for chronic hepatitis C and yields a sustained virological response rate of about 55% (6). This means that about 45% of patients with chronic hepatitis C are still threatened by the progress of the disease to cirrhosis and hepatocellular carcinoma.

Although the entry mechanism of HCV remains unclear, to date, several candidates for HCV receptors

*Address correspondence to Dr. Nobuyuki Kato, Department of Molecular Biology, Okayama University Graduate School of Medicine, Dentistry, and Pharmaceutical Sciences, 2–5–1 Shikata-cho, Okayama, Okayama 700–8558, Japan. Fax: +81–86–235–7392. E-mail: nkato@md.okayama-u.ac.jp

*Both authors contributed equally to this work.

Abbreviations: aa, amino acids; DMEM, Dulbecco's modified Eagle's medium; E2, envelope 2; GFP, green fluorescent protein; HCV, hepatitis C virus; LF, lactoferrin; MBP, maltose-binding protein; NS2, non-structural 2; TF, transferrin; VSV, vesicular stomatitis virus.

have been reported: CD81, the scavenger receptor class B type I, the mannose-binding lectins DC-SIGN and L-SIGN, low-density lipoprotein receptors, etc. (4). Most of them have been identified as interacting materials with a soluble and truncated form of the HCV E2 protein, because of the lack of efficient HCV proliferation in cell cultures, although several culture systems using PCR for detection of HCV infection have been reported (20). However, a major advance in investigating HCV entry has been achieved by the development of pseudotype viruses bearing HCV E1 and E2 proteins assembled onto retrovirus particles (2, 9) or vesicular stomatitis virus (VSV) particles (3, 23, 30). Extensive characterization of the pseudotype viruses has shown that these mimic the early steps of the HCV life cycle. This system has allowed the study of the role of candidate receptors in the early steps of HCV infection (4).

We previously found that bovine and human lactoferrins (LFs) specifically prevented HCV infection in cultured human non-neoplastic hepatocyte PH5CH8 cells using the PCR method for detection of HCV infection (10, 12). Regarding these findings, some clinical studies have demonstrated that monotherapy with bovine LF improves the serum HCV RNA and/or alanine aminotransferase levels in patients with chronic hepatitis C (15, 16, 27, 31).

LF is an 80-kDa mammalian iron-binding glycoprotein and consists of two homologous globular lobes (an N-lobe and a C-lobe), each with a single iron (Fe^{3+}) binding site. It is structurally related to the plasma iron-transport protein transferrin (TF). LF's biological roles include activities in the host defense mechanism as well as in iron metabolism (21, 22). Unlike TF, LF is a primary defense protein against microbial infection. LF possesses strong bacteriostatic and bactericidal activities against pathogenic bacteria, as well as inhibitory activity against pathogenic viruses (5, 21, 22, 33).

LF's preventive mechanism against HCV infection has been thought to be the direct interaction between LF and HCV; indeed, by Far-Western blot analysis using thioredoxin-fused LF fragments expressed in *Escherichia coli* (*E. coli*) and the soluble E2 protein expressed in Chinese hamster ovary cells, we demonstrated that the 93 carboxyl aa of LF (human, bovine, and horse), termed C-s3, specifically bound to the E2 protein (25). On the other hand, Yi et al. (34) independently reported that the E1 and E2 proteins could bind to human and bovine LFs, although the binding region of LF was not identified. Furthermore, we identified the 33 aa of human LF (termed C-s3-33; aa 600–632), which was primarily responsible for the E2 protein-binding activity, and demonstrated that maltose-binding protein (MBP)-fused C-s3-33 prevented HCV infection in PH5CH8 hepatocyte cells (25). However, the E2 protein-binding activity and the anti-HCV activity of C-s3-33 were obviously weaker than those of human LF. Therefore, we presumed that the increase of the E2 protein-binding activity would lead to the enhancement of anti-HCV activity.

To evaluate this idea, we made tandem repeats of C-s3-33, and compared their E2 protein-binding activities and anti-HCV activities with those of the C-s3-33. Here, we report our findings that the anti-HCV activity of the tandem repeats were stronger than that of the monomer when accompanied by the enhancement of the E2 protein-binding activity, by analyses using not only the HCV infection system but also the infection system of a VSV pseudotype bearing the native forms of HCV E1 and E2 proteins.

Materials and Methods

Cell cultures. Simian virus 40 large T antigen-immortalized non-neoplastic human PH5CH8 hepatocytes were maintained as described previously (11, 24). Human hepatoblastoma HepG2 cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum.

Construction of expression plasmids for E. coli. The pMAL-c2X (hLF600–632) (25) expression plasmid for the MBP-fused C-s3-33 LF fragment, was used as a template for the PCR using a primer set of hLFB6 5'-TGATAGGATCCGTGGTGTCTCGGATGATAAGG-3' containing the *Bam*HI recognition site (underlined) (25) and 632R6A 5'-ATCCATCCGAGACACCACAAACTTGTCCGGGCAGTCAGATCC-3' containing an extra 18 nts (underlined) encoding the amino-terminal 6 aa of the C-s3-33 LF fragment. After PCR (20 cycles) using KOD-plus DNA polymerase (Toyobo, Osaka, Japan), the amplified PCR product was used as a template for a second PCR using the primer set of hLFB6 and 632R 5'-TAATAAAGCTTTTAAAACTTGTCCGGGCAGTCAGATCC-3' containing the *Hin*dIII recognition site (underlined) (25). After PCR (35 cycles) using KOD-plus DNA polymerase, amplified PCR products (approximately 200 bp for the two-repeat form and approximately 300 bp for the three-repeat form) were subcloned into the *Bam*HI and *Hin*dIII sites of pMAL-c2X, and were used as expression plasmids for the production of the MBP-fused (C-s3-33)$_2$ and (C-s3-33)$_3$.

To prepare an expression plasmid for the production of the MBP-fused C-s3-33-relevant fragment (aa 587–619) of human TF, pCXbsr/huTF (29) encoding full-length human TF was used as a template for the PCR using a primer set of hTF587F 5'-TGATAG-

GATCCGTGGTCACACGG-3' containing the *Bam*HI recognition site (underlined) and hTF619R 5'-TAATAAAGCTTTTAAAAGTTGCCCG-3' containing the *Hind*III recognition site (underlined). After PCR (35 cycles) using KOD-plus DNA polymerase, the amplified PCR product was subcloned into the *Bam*HI and *Hind*III sites of pMAL-c2X, and was used as the expression plasmid.

Expression and purification of the MBP-fused protein. Expression and purification of the MBP-fused LF fragment [C-s3-33, $(C\text{-}s3\text{-}33)_2$, or $(C\text{-}s3\text{-}33)_3$] or the MBP-fused C-s3-33-relevant fragment of human TF were carried out as described previously (25). Briefly, the expression plasmid for MBP-fused protein was transformed into the *E. coli* strain JM109. The transformants were cultured at 37 C for several hours, and the harvested cells were sonicated. After removal of insoluble cellular debris by centrifugation, the supernatant obtained as the soluble fraction was applied onto an amylose resin affinity column (New England Biolabs) to obtain the MBP-fused protein. The purity of the obtained MBP-fused protein was evaluated to be more than 95% by electrophoresis on 10% SDS-PAGE gels. The concentration of the purified MBP-fused protein was determined by using Coomassie protein assay reagent (Pierce). The MBP2 (43 kDa) produced from the pMAL-c2X with a stop codon inserted into the *Xmn*I site was used as a control protein.

Far-Western blot analysis. Far-Western blot analysis was carried out as described previously (25). Briefly, 0.5 μg of human LF, MBP2, and MBP-fused LF fragments were resolved by 10% SDS-PAGE and transferred to polyvinylidene difluoride membranes. After blocking with N-buffer (25), a binding reaction was carried out using the secreted form of the E2 protein (E2-681) consisting of aa 384–681 expressed in Chinese hamster ovary cells as a probe (14), and then rat monoclonal antibody, MO-12 (13), against E2 protein was used for the detection of E2 protein-bound MBP-fused LF fragments.

Assay for anti-HCV activity of MBP-fused protein. An assay for anti-HCV activity of the MBP-fused LF fragment was carried out by the method described previously (25). Briefly, 2 μl (2×10^4 HCV) of the HCV-positive serum HCV-O (previously described as 1B-2 (19)) (genotype 1b) and the MBP-fused LF fragment (final concentration, 0.5, 1.0, and 2.0 mg/ml) were pre-incubated for 60 min at 4 C and then inoculated onto to the PH5CH8 cells (1.5×10^4 cells were cultured for 2 days before viral inoculation on a 96-well plate). After incubation of the cells for 90 min at 37 C, the cells were washed three times with PBS and further cultured for 1 day at 32 C. Cellular RNA (0.5 μg) prepared by ISOGEN extraction kit (Nippon Gene Co., Toyama, Japan) was used for the quantitative analysis of HCV RNA using LightCycler PCR as described previously (26). As the positive and negative controls for anti-HCV activity, human LF and MBP2, respectively, were used.

Assay for anti-VSV pseudotype activity of MBP-fused protein. For this assay, the VSV pseudotype VSVΔG*(HCV), bearing the native forms of HCV E1 and E2 proteins from the O strain (19), was used. VSVΔG*(HCV) was prepared by introducing the native form of E1 and E2 proteins into recombinant VSV, VSVΔG*, which harbors the green fluorescent protein (GFP) gene instead of the VSV G envelope protein gene (30). An assay for the anti-VSV pseudotype activity of the MBP-fused LF fragment was carried out by a method described previously (30). Briefly, VSVΔG*(HCV) (Approximately 100 IU/assay) was pre-incubated with the MBP-fused LF fragment (final concentration, 0.1–1.0 mg/ml) at 37 C for 60 min and inoculated onto PH5CH8 or HepG2 cells (1.5×10^4 cells were cultured for 2 days before viral inoculation on a 96-well plate). After incubation of the cells for 90 min at 37 C, the cells were washed with DMEM three times and incubated with fresh culture medium. VSVΔG*G was used as a control in this assay. After 24 hr of incubation, each infectious titer was determined by counting the number of GFP-expressing cells under a fluorescence microscope. As the positive and negative controls for the assay, human LF and MBP2 were used, respectively. Human TF and an MBP-fused C-s3-33-relevant fragment of human TF were also used for the assay.

Results

Two and Three Repeats of the Human LF Fragment (C-s3-33) Strengthened the E2 Protein-Binding Activity

Previously we found that bovine and human LFs prevented HCV infection in PH5CH8 cells via direct interaction between LF and HCV (10, 12), and we further identified 33 aa residues (C-s3-33; aa 600–632 of human LF) as an essential and minimum domain possessing binding activity for the HCV E2 protein (secreted form consisting of aa 384–681) and inhibiting activity against HCV infection (25). This result suggested that the E2 protein-binding activity contributes to the anti-HCV activity. However, the E2 protein-binding activity of C-s3-33 was somewhat weaker than that of human LF (25), and the anti-HCV activity of C-s3-33 (IC_{50}=20 μM) in the infection system using PH5CH8 cells was also weaker than that of human LF (IC_{50}=5 μM) (25). To improve these points, we first tried to

enhance the E2 protein-binding activity of C-s3-33 by the multiplication of C-s3-33. Initially, we made pMAL-c2X-based expression vectors encoding two, three, and four repeats of C-s3-33 as MBP-fused proteins, and then expressed them in *E. coli*. We successfully purified two repeats $(C\text{-}s3\text{-}33)_2$ and three repeats $(C\text{-}s3\text{-}33)_3$ of C-s3-33 as soluble forms of the MBP-fused protein; the purification of the four repeats of C-s3-33 failed due to problems with solubility. Using the MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$, we performed Far-Western blot analysis to compare their E2 protein-binding activities. The result revealed that the E2 protein-binding activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ became stronger than that of the C-s3-33, and the binding activity of $(C\text{-}s3\text{-}33)_3$ was stronger than that of $(C\text{-}s3\text{-}33)_2$ (Fig. 1). Although the E2 protein-binding activity of C-s3-33 was weaker than that of human LF, the binding activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$

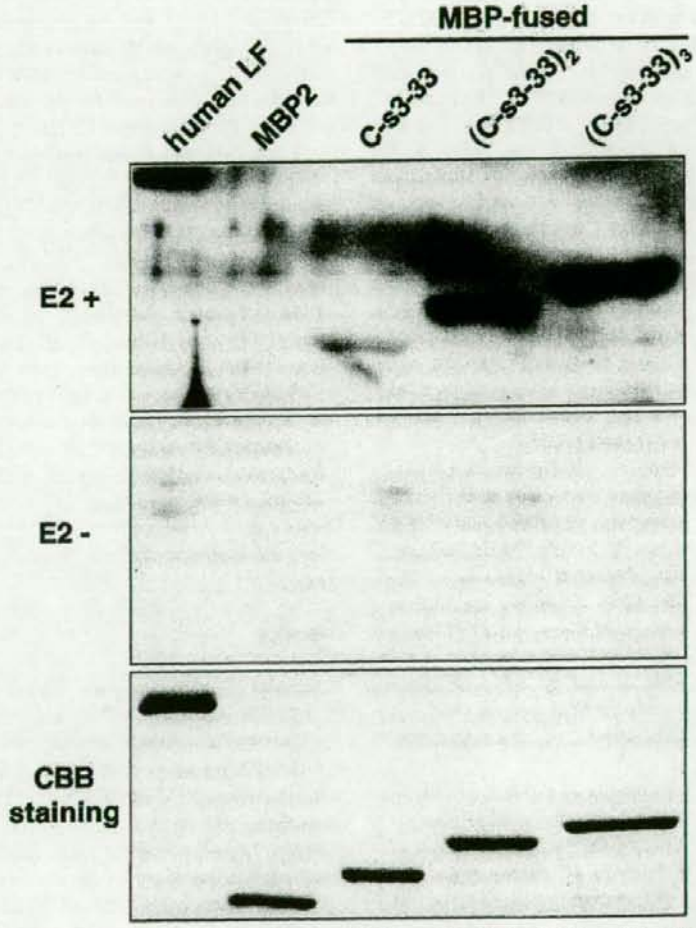

Fig. 1. Comparison of the E2 protein-binding activities of MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$, and $(C\text{-}s3\text{-}33)_3$. MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$, and $(C\text{-}s3\text{-}33)_3$ (0.5 μg each) were resolved by 10% SDS-PAGE. Human LF and MBP2 (0.5 μg each) were also used for the assay as control materials. Far-Western blot analysis using the E2 protein expressed in Chinese hamster ovary cells (14) as a probe was performed as described under "Materials and Methods." Rat monoclonal antibody MO-12 (13) against the E2 protein was used for the detection of the E2 protein bound to MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$, and $(C\text{-}s3\text{-}33)_3$ as well as human LF. Far-Western blot analysis in the absence of the E2 protein was also performed. The bottom panel shows the results for human LF, MBP2, and MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$, and $(C\text{-}s3\text{-}33)_3$ detected by staining with Coomassie Brilliant Blue.

became comparable with that of human LF (Fig. 1). To exclude the possibility of cross-reactions between C-s3-33 and the anti-E2 antibody, we performed a Far-Western blot analysis in the absence of the E2 protein. No significant bands were obtained in this control experiment (Fig. 1). The Far-Western blot analysis using normal rat serum instead of anti-E2 antibody also detected no significant bands (data not shown). These results suggest that the specific E2 protein-binding activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ increase with the degree of multiplication of C-s3-33.

$(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ Efficiently Prevented HCV Infection in PH5CH8 Cells

Since we obtained the expected results that the E2-binding activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ were stronger than that of C-s3-33, we next compared their anti-HCV activities in our HCV infection system using PH5CH8 cells (10, 25). The obtained result (Fig. 2) revealed that the anti-HCV activities of $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$ (IC_{50}=10 μM in both) became stronger than that of the C-s3-33 (IC_{50}=23 μM), although their activities were somewhat weaker than that of human LF (IC_{50}=5 μM). These results support the previous suggestion that the E2 protein-binding activity of C-s3-33 contributes to the inhibition of HCV infection (inoculum HCV-O) in human hepatocyte cells (25). However, in our HCV infection system, we failed to clearly show a difference in inhibiting activities between $(C\text{-}s3\text{-}33)_2$ and $(C\text{-}s3\text{-}33)_3$, because each standard deviation became somewhat large value due to the low level of cell culture-based HCV infection (20, 25, 31). In order to improve this point, we developed an infection system with VSVΔG*(HCV), a VSV pseudotype bearing the native E1 and E2 proteins derived from HCV-O (30), and this VSV pseudotype was used for further analysis as described below.

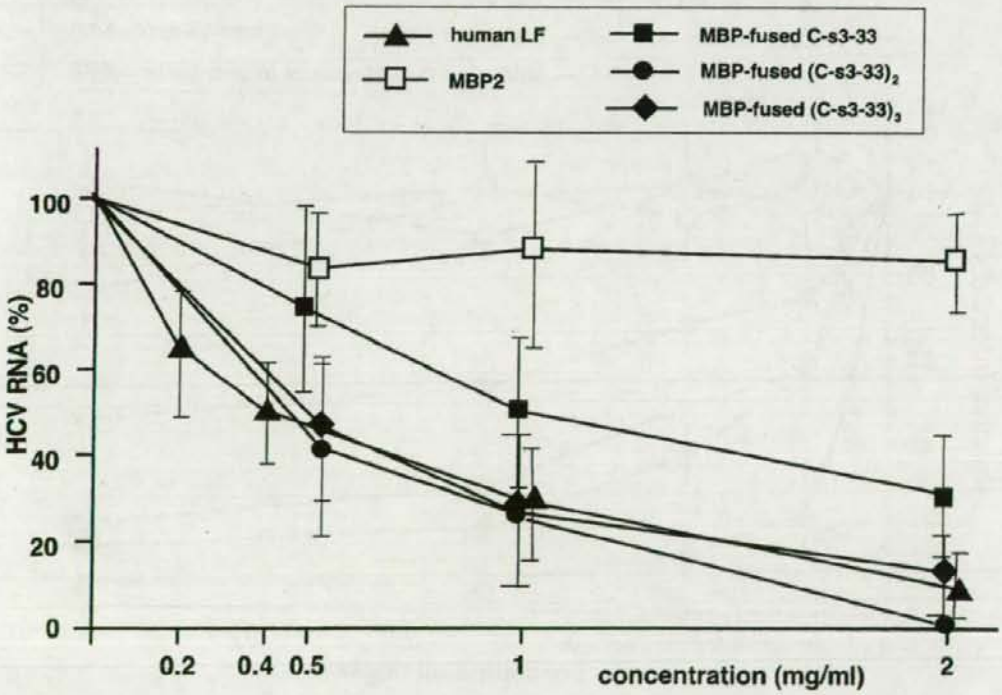

Fig. 2. Anti-HCV activities of MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$, and $(C\text{-}s3\text{-}33)_3$ in an HCV infection system using PH5CH8 cells. PH5CH8 cells and the inoculum HCV-O were used for the HCV-inhibiting assay, as described under "Materials and Methods." The number in the ordinate axis indicates the percent of HCV RNA determined by real-time LightCycler PCR (26). Approximately 2,000 copies of HCV RNA per μg of cellular RNA were reproducibly obtained using this HCV infection system (10, 26). In addition to the MBP-fused C-s3-33, $(C\text{-}s3\text{-}33)_2$, and $(C\text{-}s3\text{-}33)_3$, human LF and MBP2 were also used for the assay as control materials. The data are means ± SD of triplicates from three independent experiments.

Antiviral Effects of (C-s3-33)$_2$ and (C-s3-33)$_3$ against VSVΔG(HCV) Infection in PH5CH8 Cells*

Since PH5CH8 cells showed good susceptibility to our developed VSV pseudotype, VSVΔG*(HCV) (30), we examined the antiviral effects of (C-s3-33)$_2$ and (C-s3-33)$_3$ against VSVΔG*(HCV) infection in PH5CH8 cells, and compared them with those of the C-s3-33 and human LF. In this experiment, the antiviral effects of human TF and a C-s3-33-relevant fragment of human TF were also examined. The results (Fig. 3) clearly showed that human LF (IC_{50}=0.6 μM) strongly inhibited VSVΔG*(HCV) infection, but that human TF and the C-s3-33-relevant fragment of human TF did not, nor did MBP2, suggesting that inhibition against VSVΔG*(HCV) infection also occurred in an LF-specific manner as observed previously in the HCV infection system (25, 31). These results support previous findings (23, 30) using the VSV pseudotype infection system. Furthermore, we obtained clear results that C-s3-33 showed inhibiting activity against VSVΔG*(HCV) infection, and that its inhibiting activity was increased with multiplication of C-s3-33. The IC_{50} doses of C-s3-33, (C-s3-33)$_2$, and (C-s3-33)$_3$ were 17 μM, 5.0 μM, and 3.0 μM, respectively. This result indicates that antiviral activity of C-s3-33 is improved by the duplication and triplication of C-s3-33, although the antiviral activity of (C-s3-33)$_3$ is still weaker than that of human LF. We confirmed that these LF fragments did not inhibit VSVΔG*G (control virus without E1 and E2 proteins) infection in PH5CH8 cells (data not shown). In summary, our results suggest that direct interaction of the C-s3-33 fragment with the E2 protein in VSVΔG*(HCV) prevents the virus infection in PH5CH8 cells.

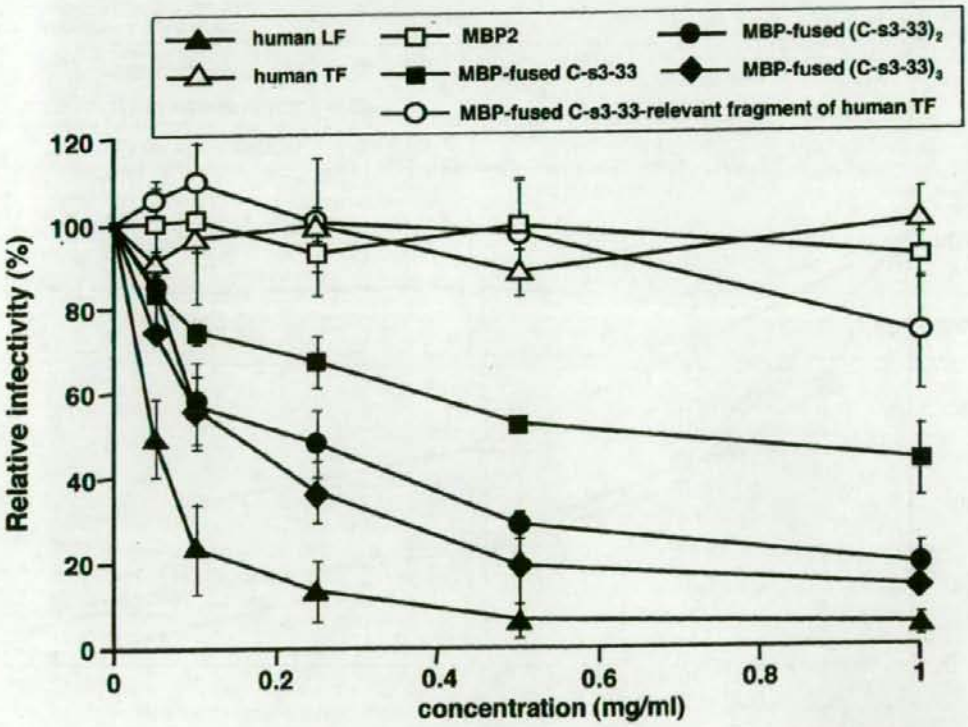

Fig. 3. Antiviral activity of the MBP-fused C-s3-33, (C-s3-33)$_2$, and (C-s3-33)$_3$ in the infection system of pseudotype virus using PH5CH8 cells. PH5CH8 cells and the VSV pseudotype, VSVΔG*(HCV), were used for the HCV-inhibiting assay, as described under "Materials and Methods." The number in the ordinate axis indicates the relative infectivity (%) calculated by counting GFP-positive cells. Approximately 100 GFP-positive cells per one assay were reproducibly obtained using this pseudotype infection system (30). In addition to the MBP-fused C-s3-33, (C-s3-33)$_2$, and (C-s3-33)$_3$, human LF, human TF, MBP2, and an MBP-fused C-s3-33-relevant fragment of human TF were also used for the assay as controls. The data are means ± SD of three independent experiments.

Antiviral Effects of (C-s3-33)$_2$ and (C-s3-33)$_3$ against VSVΔG(HCV) Infection in HepG2 Cells*

We have shown the inhibiting activities of LF fragments against HCV infection or VSV pseudotype infection in PH5CH8 cells; however, it is not clear whether or not the LF fragments used in this study show inhibiting activities against virus infection in cells other than PH5CH8 cells. To clarify this point, HepG2 cells were used for the analysis, because HepG2 cells showed the highest susceptibility to VSVΔG*(HCV) among 25 cell lines examined (30). As a consequence, we obtained similar results (Fig. 4) with those obtained in the infection system using PH5CH8 cells. The IC_{50} doses of C-s3-33, (C-s3-33)$_2$, and (C-s3-33)$_3$ were >12 μM, 7.6 μM, and 3.9 μM, respectively, indicating that, again, the inhibiting activity was increased with multiplication of C-s3-33, although antiviral activity of (C-s3-33)$_3$ was still weaker than that of human LF (IC_{50}=1.2 μM). In conclusion, our results indicated that tandem repeats of C-s3-33 enhanced the inhibiting activity in cell culture-based HCV infection.

Discussion

In our previous (30) and present studies, we showed that pretreatment of VSV pseudotypes with bovine and human LFs reduced the infectivity of VSVΔG*(HCV) and VSVΔG*(E2) bearing only the E2 protein in a dose-dependent manner, whereas pretreatment with TF did not. In contrast, LFs partially inhibited the infectivity of VSVΔG*(E1) bearing only the E1 protein (30). These results suggested that the interaction of LF and the E2 protein is the main contributing factor to the prevention of HCV infection. This idea has been strongly supported by the results obtained in this study. We demonstrated that tandem repeats of C-s3-33, an anti-HCV peptide derived from human LF, enhanced the E2 protein-binding activity and the inhibiting activity

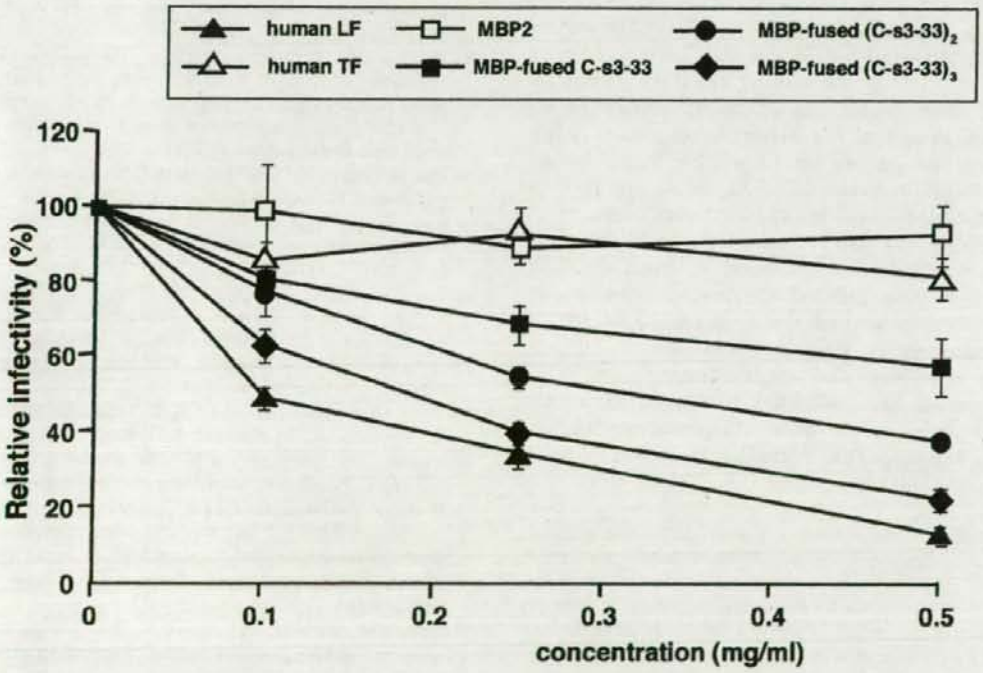

Fig. 4. Antiviral activity of the MBP-fused C-s3-33, (C-s3-33)$_2$, and (C-s3-33)$_3$ in the infection system of the pseudotype virus using HepG2 cells. HepG2 cells and the VSV pseudotype, VSVΔG*(HCV), were used for the HCV-inhibiting assay, as described in "Materials and Methods." The number in the ordinate axis indicates the relative infectivity (%) calculated by counting GFP-positive cells. Approximately 100 GFP-positive cells per one assay were reproducibly obtained using this pseudotype infection system (30). In addition to the MBP-fused C-s3-33, (C-s3-33)$_2$, and (C-s3-33)$_3$, human LF, human TF, and MBP2 were used for the assay as controls. The data are means ± SD of three independent experiments.

against infection by HCV or the VSV pseudotype, VSVΔG*(HCV), in human hepatic cell lines. These results strongly suggest that the direct interaction between C-s3-33 and the E2 protein plays a central role in the inhibition of HCV infection by LF.

Since C-s3-33 or repeated forms of C-s3-33 could prevent HCV and VSVΔG*(HCV) infection, C-s3-33 must bind to a region other than the region (aa 441–500 of E2 protein) required for heteromeric complex formation between E1 and E2 proteins. Our preliminary results suggested that the C-s3-33 bound to aa 411–500 and aa 600–661 of the E2 protein, indicating that the target sites of C-s3-33 may be plural. This result suggests a rather complex interaction between C-s3-33 and the E2 protein. To clarify this point, further comprehensive analysis will be needed.

Although tandem repeats of C-s3-33 enhanced the anti-HCV activity compared with that of the C-s3-33, the fact that their antiviral activities were still several-fold weaker than that of original human LF remains a subject to be resolved. As one approach to increase anti-HCV activity, tandem repeats of C-s3-33-relevant fragment of bovine LF may be useful, because we previously observed that the anti-HCV activity of bovine LF (IC_{50}=1.5 μM) was stronger than that of human LF (IC_{50}=5.0 μM) (26), and that the E2 protein-binding activity of the C-s3 (93 aa)-relevant fragment of bovine LF was stronger than that of C-s3 (25). Since 10 aa out of 33 aa differ between C-s3-33 and its relevant fragment of bovine LF, some aa substitutions between both fragments may help to further increase the anti-HCV activity of LF-derived peptides. Alternatively, some spacer between the C-s3-33 repeats may be needed. Therefore, further trials will be needed to achieve the maximum anti-HCV activity of C-s3-33.

In conclusion, the results of the present study demonstrated that tandem repeats of human LF-derived 33 aa prevented HCV infection more strongly than the 33 aa, and suggest that this repeated form will be useful as a novel anti-HCV reagent.

We thank T. Nakamura and A. Morishita for their technical assistance. This work was supported by a Grant-in-Aid for the Third-term Comprehensive 10-year Strategy for Cancer Control and by a Grant-in-Aid for research on hepatitis, both from the Ministry of Health, Labour, and Welfare, Japan; and by the Program for the Promotion of Fundamental Studies in Health Science of the Pharmaceutical and Medical Device Agency (PMAD), Japan. K.A. was supported by a Research Fellowship from the Japan Society for the Promotion of Science (JSPS) for Young Scientists.

References

1) Bartenschlager, R., and Lohmann, V. 2000. Replication of hepatitis C virus. J. Gen. Virol. **81**: 1631–1648.
2) Bartosch, B., Dubuisson, J., and Cosset, F.L. 2003. Infectious hepatitis C virus pseudo-particles containing functional E1-E2 envelope protein complexes. J. Exp. Med. **197**: 633–642.
3) Buonocore, L., Blight, K.J., Rice, C.M., and Rose, J.K. 2002. Characterization of vesicular stomatitis virus recombinants that express and incorporate high levels of hepatitis C virus glycoproteins. J. Virol. **76**: 6865–6872.
4) Cocquerel, L., Voisset, C., and Dubuisson, J. 2006. Hepatitis C virus entry: potential receptors and their biological functions. J. Gen. Virol. **87**: 1075–1084.
5) Farnaud, S., and Evans, R.W. 2003. Lactoferrin—a multifunctional protein with antimicrobial properties. Mol. Immunol. **40**: 395–405.
6) Feld, J.J., and Hoofnagle, J.H. 2005. Mechanism of action of interferon and ribavirin in treatment of hepatitis C. Nature **436**: 967–972.
7) Hijikata, M., Kato, N., Ootsuyama, Y., Nakagawa, M., and Shimotohno, K. 1991. Gene mapping of the putative structural region of the hepatitis C virus genome by *in vitro* processing analysis. Proc. Natl. Acad. Sci. U.S.A. **88**: 5547–5551.
8) Hijikata, M., Mizushima, H., Tanji, Y., Komoda, Y., Hirowatari, Y., Akagi, T., Kato, N., Kimura, K., and Shimotohno, K. 1993. Proteolytic processing and membrane association of putative nonstructural proteins of hepatitis C virus. Proc. Natl. Acad. Sci. U.S.A. **90**: 10773–10777.
9) Hsu, M., Zhang, J., Flint, M., Logvinoff, C., Cheng-Mayer, C., Rice, C.M., and McKeating, J.A. 2003. Hepatitis C virus glycoproteins mediate pH-dependent cell entry of pseudotyped retroviral particles. Proc. Natl. Acad. Sci. U.S.A. **100**: 7271–7276.
10) Ikeda, M., Nozaki, A., Sugiyama, K., Tanaka, T., Naganuma, A., Tanaka, K., Sekihara, H., Shimotohno, K., Saito, M., and Kato, N. 2000. Characterization of antiviral activity of lactoferrin against hepatitis C virus infection in human cultured cells. Virus Res. **66**: 51–63.
11) Ikeda, M., Sugiyama, K., Mizutani, T., Tanaka, T., Tanaka, K., Sekihara, H., Shimotohno, K., and Kato, N. 1998. Human hepatocyte clonal cell lines that support persistent replication of hepatitis C virus. Virus Res. **56**: 157–167.
12) Ikeda, M., Sugiyama, K., Tanaka, T., Tanaka, K., Sekihara, H., Shimotohno, K., and Kato, N. 1998. Lactoferrin markedly inhibits hepatitis C virus infection in cultured human hepatocytes. Biochem. Biophys. Res. Commun. **245**: 549–553.
13) Inudoh, M., Kato, N., and Tanaka, Y. 1998. New monoclonal antibodies against a recombinant second envelope protein of hepatitis C virus. Microbiol. Immunol. **42**: 875–877.
14) Inudoh, M., Nyunoya, H., Tanaka, T., Hijikata, M., Kato, N., and Shimotohno, K. 1996. Antigenicity of hepatitis C virus envelope proteins expressed in Chinese hamster ovary cells. Vaccine **14**: 1590–1596.
15) Ishii, K., Takamura, N., Shinohara, M., Wakui, N., Shin, H.,

Sumino, Y., Ohmoto, Y., Teraguchi, S., and Yamauchi, K. 2003. Long-term follow-up of chronic hepatitis C patients treated with oral lactoferrin for 12 months. Hepatol. Res. **25**: 226–233.

16) Iwasa, M., Kaito, M., Ikoma, J., Takeo, M., Imoto, I., Adachi, Y., Yamauchi, K., Koizumi, R., and Teraguchi, S. 2002. Lactoferrin inhibits hepatitis C virus viremia in chronic hepatitis C patients with high viral loads and HCV genotype 1b. Am. J. Gastroenterol. **97**: 766–767.
17) Kato, N. 2001. Molecular virology of hepatitis C virus. Acta Med. Okayama **55**: 133–159.
18) Kato, N., Hijikata, M., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S., Sugimura, T., and Shimotohno, K. 1990. Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. Proc. Natl. Acad. Sci. U.S.A. **87**: 9524–9528.
19) Kato, N., Ikeda, M., Sugiyama, K., Mizutani, T., Tanaka, T., and Shimotohno, K. 1998. Hepatitis C virus population dynamics in human lymphocytes and hepatocytes infected *in vitro*. J. Gen. Virol. **79**: 1859–1869.
20) Kato, N., and Shimotohno, K. 2000. Systems to culture hepatitis C virus. Curr. Top. Microbiol. Immunol. **242**: 261–278.
21) Levay, P.F., and Viljoen, M. 1995. Lactoferrin: a general review. Haematologica **80**: 252–267.
22) Lonnerdal, B., and Iyer, S. 1995. Lactoferrin: molecular structure and biological function. Annu. Rev. Nutr. **15**: 93–110.
23) Matsuura, Y., Tani, H., Suzuki, K., Kimura-Someya, T., Suzuki, R., Aizaki, H., Ishii, K., Moriishi, K., Robison, C.S., Whitt, M.A., and Miyamura, T. 2001. Characterization of pseudotype VSV possessing HCV envelope proteins. Virology **286**: 263–275.
24) Noguchi, M., and Hirohashi, S. 1996. Cell lines from non-neoplastic liver and hepatocellular carcinoma tissue from a single patient. In Vitro Cell. Dev. Biol. Anim. **32**: 135–137.
25) Nozaki, A., Ikeda, M., Naganuma, A., Nakamura, T., Inudoh, M., Tanaka, K., and Kato, N. 2003. Identification of a lactoferrin-derived peptide possessing binding activity to hepatitis C virus E2 envelope protein. J. Biol. Chem. **278**: 10162–10173.
26) Nozaki, A., and Kato, N. 2002. Quantitative method of intracellular hepatitis C virus RNA using LightCycler PCR. Acta Med. Okayama **56**: 107–110.
27) Okada, S., Tanaka, K., Sato, T., Ueno, H., Saito, S., Okusaka, T., Sato, K., Yamamoto, S., and Kakizoe, T. 2002. Dose-response trial of lactoferrin in patients with chronic hepatitis C. Jpn. J. Cancer Res. **93**: 1063–1069.
28) Saito, I., Miyamura, T., Ohbayashi, A., Harada, H., Katayama, T., Kikuchi, S., Watanabe, Y., Koi, S., Onji, M., Ohta, Y., Choo, Q.L., Houghton, M., and Kuo, G. 1990. Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. Proc. Natl. Acad. Sci. U.S.A. **87**: 6547–6549.
29) Tamura, T., Nozaki, A., Abe, K., Dansako, H., Naka, K., Ikeda, M., Tanaka, K., and Kato, N. 2005. cDNA microarray analysis of lactoferrin expression in non-neoplastic human hepatocyte PH5CH8 cells. Biochim. Biophys. Acta **1721**: 73–80.
30) Tamura, K., Oue, A., Tanaka, A., Shimizu, N., Takagi, H., Kato, N., Morikawa, A., and Hoshino, H. 2005. Efficient formation of vesicular stomatitis virus pseudotypes bearing the native forms of hepatitis C virus envelope proteins detected after sonication. Microbes Infect. **7**: 29–40.
31) Tanaka, K., Ikeda, M., Nozaki, A., Kato, N., Tsuda, H., Saito, S., and Sekihara, H. 1999. Lactoferrin inhibits hepatitis C virus viremia in patients with chronic hepatitis C: a pilot study. Jpn. J. Cancer Res. **90**: 367–371.
32) Thomas, D.L. 2000. Hepatitis C epidemiology. Curr. Top. Microbiol. Immunol. **242**: 25–41.
33) van der Strate, B.W., Beljaars, L., Molema, G., Harmsen, M.C., and Meijer, D.K. 2001. Antiviral activities of lactoferrin. Antiviral Res. **52**: 225–239.
34) Yi, M., Kaneko, S., Yu, D.Y., and Murakami, S. 1997. Hepatitis C virus envelope proteins bind lactoferrin. J. Virol. **71**: 5997–6002.

JOURNAL OF EXPERIMENTAL ZOOLOGY (MOL DEV EVOL) 306B:59–69 (2006)

Isolation of the Feline α1,3-Galactosyltransferase Gene, Expression in Transfected Human Cells and its Phylogenetic Analysis

BIBHUTI BHUSAN ROY, ATSUSHI JINNO-OUE, MASAHIKO SHINAGAWA, AKIRA SHIMIZU, KAZUSHI TAMURA, NOBUAKI SHIMIZU, ATSUSHI TANAKA, AND HIROO HOSHINO*

Department of Virology and Preventive Medicine, Gunma University Graduate School of Medicine, Maebashi, Gunma 371-8511, Japan

ABSTRACT The enzyme alpha 1,3-galactosyltransferase (α1,3-GT), which catalyzes synthesis of terminal α-galactosyl epitopes (Gal α1,3Gal β1-4GlcNAc-R), is produced in non-primate mammals, prosimians and new-world monkeys, but not in old-world monkeys, apes and humans. We cloned and sequenced a cDNA that contains the coding sequence of the feline α1,3-GT gene. Flow cytometric analysis demonstrated that the α-galactosyl epitope was expressed on the surface of a human cell line transduced with an expression vector containing this cDNA, and this α-galactosyl epitope expression subsided by α-galactosidase treatment. The open reading frame of the feline α1,3-GT cDNA is 1,113 base pairs in length and encodes 371 amino acids. The nucleotide sequence and its deduced amino acid sequence of the feline α1,3-GT gene are 88–90% and 85–87%, respectively, similar to the reported sequences of the bovine, porcine, marmoset and cebus monkey α1,3-GT genes, while they are 88% and 82–83%, respectively, similar to those of the orangutan and human α1,3-GT pseudogenes, and 81% and 77%, respectively, similar to the murine α1,3-GT gene. Thus, the α1, 3-GT genes and pseudogenes of mammals are highly similar. Ratios of non-synonymous nucleotide changes among the primate pseudogenes as well as the primate genes are still higher than the ratios of non-primates, suggesting that the primate α1,3-GT genes tend to be divergent. *J. Exp. Zool. (Mol. Dev. Evol.) 306B:59–69, 2006.*

Alpha 1,3-galactosyltransferase (α1,3-GT) (EC 2.4.1.151) is responsible for the synthesis of terminal α-galactosyl epitopes (Gal α1,3Gal β1-4GlcNAc-R) in sugar chains. These epitopes are produced by non-primate mammals, prosimians and new-world monkeys and are found on the cell surface ($>10^6$ epitopes/cell) as well as in secreted glycoproteins. The full-length α1,3-GT sequences of human and orangutan (Koike et al., 2002) and some partial sequences, similar to α1,3-GT genes of non-primate mammals, have been detected in humans and higher primates (Larsen et al., '90; Joziasse et al., '91). These sequences have been judged to be pseudogenes because of the generation of premature stop codons due to multiple base deletions (Larsen et al., '90; Joziasse et al., '91; Koike et al., 2002). The expression of α1,3-GT protein is not detected in humans and other catarrhines (Spiro and Bhoyroo, '84; Galili et al., '87, '88; Thall and Galili, '90), thus resulting in the production of large amounts of a natural antibody against the α-galactosyl epitope (Galili et al., '84, '87).

The α1,3-GT is a Golgi membrane-bound enzyme that catalyzes the addition of α-galactosyl epitopes to existing β-galactose terminals accord-

Grant sponsor: Grant-in-Aids from the Japanese Society for the Promotion of Sciences; Grant sponsor: CREST; Grant sponsor: Japan Health Sciences Foundation; Grant sponsor: 21st Century COE Program.

The nucleotide sequence reported in this paper has been submitted to the GenBank™ with the accession number AY167024 (feline α1, 3-GT).

Present Address: Bibhuti Bhusan Roy, McGill AIDS Centre, Lady Davis Institute, Jewish General Hospital, Montreal, Que., Canada H3T 1E2.

*Correspondence to: Hiroo Hoshino, Department of Virology and Preventive Medicine, Gunma University Graduate School of Medicine, Showa-Machi 3-39-22, Maebashi, Gunma 371-8511, Japan.
E-mail: hoshino@med.gunma-u.ac.jp

Received 26 May 2005; Accepted 27 July 2005
Published online 10 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jez.b.21072.

ing to the following reaction (Basu and Basu, '73; Blanken and Van den Eijnden, '85; Elices et al., '86):

UDP galactose+β-D-galactosyl-1,4-*N*-acetyl-D-glucosaminyl-R→ UDP+α-D-galactosyl-1,3-β-D-galactosyl-1,4-*N*-acetyl-D-glucosaminyl-R, in which R may be a glycoprotein or a glycolipid. α1,3-Galactose and less abundant α1,6-linked galactose can be cleaved by the enzyme α-galactosidase (EC 3.2.1.22).

The whole mRNA for α1,3-GT has been found in cells of non-primate mammals and marmoset (Joziasse et al., '89; Larsen et al., '89; Henion et al., '94; Strahan et al., '95; Koike et al., 2002), but only partial mRNAs have been detected in those of old-world monkeys (OWM) (rhesus monkey, green monkey and patas monkey) or human cells (Joziasse et al., '89; Joziasse, '92), indicating that the regulatory sequences can still be functional. Several homologs of the α1,3-GT gene have been described in the human genome, but all of them contain several frame-shift mutations that lead to the generation of premature internal stop codons (Larsen et al., '90; Joziasse et al., '91). One homolog is thought to correspond to the original α1,3-GT gene because it contains intronic sequences (Joziasse et al., '91) as well as one exonic sequence corresponding to the largest part of the catalytic domain of the enzyme (Larsen et al., '90), and this gene has been localized on chromosome 9. Another one, which does not contain any intronic sequences and is located on chromosome 12, corresponds to a copy of the α1,3-GT gene (Larsen et al., '90; Joziasse et al., '91). This pseudogene has also been found in apes and OWM (Galili and Swanson, '91).

By Northern blot analysis, 3.6–3.9 kb α1,3-GT transcripts are detected in bovine and marmoset cells, but not detected in human and OWN cells (Joziasse et al., '89). For the first time, Koike et al. (2002) have detected the full-length sequences of α1,3-GT of orangutans and humans using sensitive PCR-based methods. Comparison of the deduced amino acid sequences with the marmoset gene sequence has revealed that in the human α1,3-GT pseudogene sequence, three single-nucleotide deletions are present at the site corresponding to the amino acid positions 81, 256 and 284 of marmoset α1,3-GT protein, and result in the appearance of the stop codons at positions 268 and 362. Thus, only non-functional proteins are made even when the full-length mRNA is overexpressed and translated in human cells.

A study on the murine α1,3-GT gene has revealed that it is distributed over nine exons that span at least 35 kb of the genomic sequence (Joziasse, '92). Transcription of this gene results in the production of four distinct mRNAs that are generated by the alternative splicing. Translation of these individual mRNAs produces four related isoforms of α1,3-GT consisting of 337, 349, 359 and 371 amino acids.

In this study, we have described the isolation and characterization of a cDNA clone comprising the complete coding sequence of feline (cat) α1,3-GT that is capable of catalyzing the synthesis of α-galactosyl epitopes on human cells. Phylogenetic analysis was performed to evaluate the evolutionary relationship among the six different full coding sequences of α1,3-GT genes and two pseudogenes.

MATERIALS AND METHODS

Cell culture

8C feline kidney cells (Fischinger et al., '73) were cultured in Eagle's minimum essential medium (EMEM) supplemented with 10% fetal calf serum (FCS) at 37°C under 5% CO_2 in humidified air. A human osteosarcoma cell line, HOS, infected with a Melanesian strain of human T-cell leukemia virus type I (HTLV-I), termed HOS/HTLV-I_{MEL5}, was cultured in Dulbecco's modified EMEM (Hoshino et al., '93).

Reverse transcriptase-polymerase chain reaction

When 8C cells became confluent, they were collected in an Eppendorf tube using a scraper. Total RNA was isolated from 8C cells using an RNA extraction kit, SepaGene (Sanko-Junyaku Co. Ltd., Tokyo, Japan), according to the manufacturer's protocol. cDNA was synthesized using total cellular RNA as the template, using the SuperScript™ Preamplification System for First-Strand cDNA Synthesis kit (GibcoBRL, Life Technologies™, Carlsbad, CA) in accordance with the manufacturer's protocol. PCR was used to amplify the α1,3galactosyltransferase gene in cDNA preparation. Two sets of primers were used for PCR. One set was made according to the porcine α1,3-GT sequence as follows: 5′-ATGAATGTCAAAGGAAGAGTGGTTCTGTCA-3′ as a forward primer and 5′-TCAGATGTTATTTCTAACCAAATTATACTC-3′ as a reverse primer. The other set was designed using DNA sequences of α1,3-GT of pig (Strahan et al., '95), mouse (Larsen et al., '89), cow (Joziasse et al., '89) and

marmoset (Galili et al., '88). Namely, the regions where the nucleotide sequences of α1,3-GT are well conserved were selected and used to make PCR primers as follows: 5′-GGAGAAAATAAT GAATGTCAA-3′ as a forward primer and 5′-TCAGATGTTATTTCTAACCAAATT-3′ as a reverse primer. The initiation and stop codons in the primers are underlined.

Cloning of the feline α1,3-GT and sequencing

The PCR product was cloned into a TA cloning vector, pCR2.1 (Invitrogen, Carlsbad, CA), and the ligated product was used to transform DH5α *E. coli* strain. *E. coli* colonies were tested for mini-scale plasmid preparation, and colonies containing a 1.1 kb insert after multiple enzyme digestion were selected. Clones containing inserts were sequenced using Texas Red-labeled M13 forward and reverse primers for sequencing DNA. A Hitachi SQ-5500 automated DNA sequencer and software (Hitachi, Tokyo, Japan) were used to determine DNA sequences as described previously (Jinno et al., '98). The plasmid vector pCR2.1 containing the insert and the expression vector pcDNA3 (Clontech Co., Palo Alto, CA) were digested with *BamHI* and *NotI* and purified. The insert in the pCR2.1 vectors was then re-ligated to the digested pcDNA3 vector DNA. After transformation of competent cells, DH5α, by this ligation product, plasmids purified from colonies contained an insert of about 1.1 kb. It was confirmed by enzyme digestion of the plasmid DNA with *StuI*, *SpeI*, *BamHI* and *StuI* that the inserts were in a correct direction (data not shown).

Transfection of the feline α1,3-GT gene and detection of its expression

The mammalian expression vector pcDNA3 harboring the gene of the feline α1,3-GT was transfected into a clonal line of HTLV-I-infected HOS cells using LipofectAMINE (GibcoBRL) to examine the effects of α1,3-GT expression on HTLV-I infection, and neomycin-resistant cells were selected as described elsewhere (Jinno et al., '98). In short, the cells were seeded at 4×10^5 cells/ml/well into 12-well plates and incubated overnight. The cells in nine wells were independently transfected with pcDNA3 DNA harboring the gene of the feline α1,3-GT or pcDNA3 DNA alone using LipofectAMINE (GibcoBRL). The cells were selected with neomycin after 24 hr incubation and maintained for 3 weeks. The expression of α-galactosyl epitope on the surface of HOS and HOS/HTLV-I cells and feline cells was detected by flow cytometry (FCM) (Cyto ACE-100, Auto cell screener, Japan Spectroscopic Co., Ltd.) after treatment with fluorescent isothiocyanate labeled *Bandeiraea simplicifolia* Isolectin B_4 (FITC/BS-IB_4) (Sigma Chemical Co., St. Louis, MO) (Wood et al., '79; Azimzadeh et al., '97; Bracy et al., '98). Briefly, cells were collected into Eppendorf tubes from cultured plates, washed with washing solution (cold PBS with 1% FCS and 0.01% NaN3), pelleted by centrifugation at 5,000 rpm for 5 min and treated for 1 hr with diluted FITC/BS-IB_4 (3.3 pg/ml) on ice. The cells were washed and fixed with 1% paraformaldehyde and analyzed by FCM. In seven (#1–#7) out of nine wells seeded with the feline α1,3-GT-transduced cells, 20–45% of cells were positively stained with FITC/BS-IB_4 (Table 1). These cells were cloned using 96-well plates by seeding 1 cell/well and thus obtained clones were screened again by FCM.

α-Galactosidase treatment

Cells were treated with α-galactosidase as described by Bracy et al. ('98). Briefly, cells were seeded at 4×10^5 cells/well into 6-well plates in 2 ml and incubated overnight. The culture medium was replaced with fresh medium and the cells were incubated for another day. The cells were collected into two Eppendorf tubes from each cultured plate, washed with the washing solution

TABLE 1. Flow cytometry of the cells transfected with the feline α1,3-GT gene

Cell	BS-IB4-positive cells (%)
HOS/HTLV-I	4
HOS/HTLV-I/pcDNA3	5
HOS/HTLV-I/fGT	
#1	45
#2	44
#3	38
#4	32
#5	27
#6	27
#7	22
#8	4
#9	4

Human HOS/HTLV-I cells in nine culture wells were transfected with the feline α1,3-GT gene independently. Neomycin-resistant cells were selected. These HOS cells were processed for FCM using FITC/BS-IB4 lectin and examined for the expression of α-galactosyl epitope on the cell surface. HOS/HTLV-I/pcDNA3 cells were used as a control negative for terminal α1,3-galactose. The cut-off value for FCM was set at the fluorescence intensity that gave 5% positive for the control cells as shown in Figs. 1 and 2.