Nucleotide Sequence of the Genes Coding for the Membrane Glycoproteins E1 and E2 of Rubella Virus

By GABRIELE VIDGREN, KRISTIINA TAKKINEN, NISSE KALKKINEN, LEEVI KÄÄRIÄINEN AND RALF F. PETTERSSON

Recombinant DNA Laboratory, University of Helsinki, Valimotie 7, SF-00380 Helsinki, Finland and 1Ludwig Institute for Cancer Research, Stockholm Branch, Box 60202, S-104 01 Stockholm, Sweden

(Accepted 1 June 1987)

SUMMARY

The complete nucleotide sequence of the genes coding for the two membrane glycoproteins E1 and E2 of rubella virus has been determined from cloned cDNA derived from the 40S genomic RNA. A sequence of 2451 nucleotides extending from the poly(A) tract towards the 5' end is presented. Within one continuous open reading frame E2 is located amino-terminally followed by E1 and a 58 residue long untranslated 3' region preceding the poly(A) tract. The coding regions of E2 and E1 are unusually G/C rich, 71.4% and 66.4% respectively. At the carboxy-terminal end of the coding region of E1, there is an inverted complementary nucleotide sequence, which could form a 13 base pair hairpin structure. Mature E2 and E1 are both preceded by a stretch of uncharged mainly non-polar amino acids, 21 and 20 residues in length, respectively. These could serve as signal peptides that mediate the membrane translocation of the proteins. At the carboxy termini of both proteins there are stretches of hydrophobic amino acids, 19 and 27 residues in length, which probably represent the transmembrane anchors of the proteins. The size of mature E1 is 481 amino acids (mol. wt. 51502), whereas the exact size of mature E2 could not be established as its carboxy-terminal end could not be located in the sequence. A maximum size of 262 amino acids (mol. wt. 28277) is, however, suggested. Between the E2 and E1 genes, there is a stretch of seven amino acids, five of which are arginines, which may serve as cleavage sites for a trypsin-like protease. E1 contains three and E2 four potential sites for asparagine-linked glycosylation. Both proteins are cysteine-rich (5%). Comparison of the rubella virus amino acid sequence to those of several alphaviruses indicated no sequence homology.

INTRODUCTION

The structural characteristics and general strategy of gene expression of rubella virus (RV), the only member of the Rubivirus genus within the family Togaviridae (Porterfield et al., 1978), have recently been clarified (Oker-Blom et al., 1983, 1984; Waxham & Wolinsky, 1983; Toivonen et al., 1983; Oker-Blom, 1984; Kalkkinen et al., 1984; Pettersson et al., 1985). RV contains three structural proteins E1 (Mr 58000), E2 (Mr 42000 to 47000) and C (Mr 33000) (Oker-Blom et al., 1983; Waxham & Wolinsky, 1983; Toivonen et al., 1983). E1 and E2 are membrane glycoproteins forming the viral spikes. E2 can be separated by SDS–PAGE into two (or three) differently migrating species (E2a and E2b) (Oker-Blom et al., 1983) which share the same polypeptide backbone. They probably arise as a result of heterogeneous glycosylation (Kalkkinen et al., 1984). Both proteins contain asparagine-linked glycans, since the Mr of E2 and E1 is reduced to 30000 and 53000 respectively, when they are synthesized in the presence of tunicamycin (Oker-Blom et al., 1983). The C protein is associated with the 40S RNA genome,
which is single-stranded and of positive polarity. The RNA is about 11,000 nucleotides in length and contains a 3' poly(A) tract and a 5' cap structure (Oker-Blom et al., 1984).

The three structural proteins are translated from a subgenomic 24S mRNA (approximate size 3300 nucleotides), which corresponds to the 3' third of the 40S RNA. The proteins are synthesized principally as a 110,000 mol. wt. (p110) precursor as shown by translation in vitro of the 24S mRNA. According to pulse-labelling of RV-infected cells, p110 is cleaved cotranslationally into the three proteins. The gene order in p110 is NH2-C-E2-E1-COOH (Oker-Blom, 1984). Thus, the general strategy of RV gene expression is very similar to that of the alphaviruses (Strauss & Strauss, 1983).

Here we report the complete nucleotide sequence of cDNA encoding the E1 and E2 proteins.

**METHODS**

*Cells and virus.* The origin of RV (Therien strain), cultivation in B-Vero cells in glass roller bottles or plastic dishes, labelling with $^{32}$P and $^{35}$S methionine, and purification of virus have been described previously (Oker-Blom et al., 1983, 1984).

*Synthesis and cloning of cDNA.* Purified RV (about 2 mg per experiment) was solubilized with 2% SDS and the 40S genomic RNA purified by sucrose gradient centrifugation as described (Oker-Blom et al., 1984). Using oligo(dT)$_{12-18}$ (Pharmacia P-L) as primer, both the first and second cDNA strands were synthesized using reverse transcriptase (Life Sciences, St Petersburg, Fla., U.S.A.) under conditions described previously (Söderlund et al., 1981). The cDNA was size-fractionated on a 15 to 30% (w/w) sucrose gradient made in TSE-SDS (50 mM-Tris-HCl pH 7.4, 150 mM-NaCl, 1 mM-EDTA, 0.1% SDS) for 23 h at 25,000 r.p.m. and 23 °C using a Spinco SW28.1 rotor. cDNA larger than 700 bp was tailed with dC residues using terminal transferase (Bethesda Research Laboratories) (Söderlund et al., 1981). The cDNA was ligated to dg-tailed PstI-cleaved pBR322 and transformed into *Escherichia coli* HB101. Tetracycline-resistant, ampicillin-sensitive colonies containing RV-specific inserts were detected by colony hybridization (Grunstein & Hogness, 1975) or by Southern blotting (Southern, 1975) of isolated plasmid DNA using $^{32}$P-labelled RV 40S RNA as a probe. Hybrid plasmids containing inserts larger than about 800 bp were selected for further analyses.

*Sequence determination.* RV-specific inserts were characterized by restriction enzyme mapping. Four partially overlapping clones were selected for sequence determination. Restriction enzyme fragments were isolated from polyacrylamide gels and subcloned in M13mp8 and 9 (Messing, 1983). The nucleotide sequences of subcloned fragments were determined in most cases for both strands by the dideoxy method (Sanger et al., 1977). In some cases, the nucleotide sequence was confirmed by sequencing 5' or 3' end-labelled fragments by the method of Maxam & Gilbert (1980).

*Amino terminus amino acid sequence determination.* Unlabelled RV membrane proteins E1 and E2 were separated by preparative SDS-PAGE together with $^{35}$S methionine labelled virus proteins added as markers (Oker-Blom et al., 1983; Kalkkinen et al., 1984). Following localization by autoradiography, the proteins were electroeluted from the gel slices (Allington et al., 1978) and lyophilized. SDS was removed from the preparations by ion-pair extraction (Henderson et al., 1979) and the proteins were digested with trypsin (3% w/w in 100 μl of 0.1 M-ammonium bicarbonate). Tryptic peptides were separated by HPLC on a 0.46 x 15 cm Vydac 218 TPB5 column (Separations Group, Hesperia, Ca., U.S.A.) by using a gradient of acetonitrile (0 to 60% in 60 min) in 0.1% trifluoroacetic acid and detection at 218 nm. For amino terminus sequence analysis the collected peptides were lyophilized and degraded in an Applied Biosystems 470A gas-phase sequencer using online detection (Applied Biosystems 120A) of the released amino acid derivatives.

*Northern blot analysis.* The viral specificity of the cloned cDNA was analysed by Northern blotting using RNA isolated from RV-infected cells. Poly(A)-containing RNA (about 5 μg per sample) was denatured and fractionated on a 1% agarose gel as described previously (Oker-Blom et al., 1984). The RNA was transferred to nitrocellulose paper and prehybridized as described by Thomas (1980). The filters were then incubated with the various hybrid plasmids which had been labelled by nick translation with $^{32}$P. Filter hybridization was carried out as described previously (Oker-Blom et al., 1984; Thomas, 1980). The filters were washed and analysed by autoradiography.

*Colinearity analysis.* To check that the RV-specific inserts represented continuous stretches of sequences in the viral RNA, the various clones were subjected to colinearity analyses. Hybrid plasmids were linearized with EcoRI (which did not cut the RV-specific inserts). The DNA was denatured in formamide and hybridized to $^{32}$P-labelled RV 40S RNA as described previously (Söderlund et al., 1981). After S1 nuclease digestion, the labelled DNA/RNA duplexes were analysed by electrophoresis on a neutral 1% agarose gel followed by autoradiography.

*Computer analysis.* The nucleotide sequences derived from the various fragments were assembled using the SEQAID program (Peltola et al., 1984). Analyses of the sequences were carried out using the program developed by the University of Wisconsin Genetics Computer Group (UWGCG). The analyses were executed on a VAX 11/750 computer. The hydropathic plot analysis was based on the method developed by Kyte & Doolittle (1982).
Sequence of rubella virus glycoproteins

Sequence homology analyses were done using the DOTPLOT/UWGCG program developed by Maizel & Lenk (1981) or the GAP/UWGCG program, which uses the alignment method of Needleman & Wunsch (1970). For dot plot analyses of protein homology, a window of 30 amino acids and a stringency of 8 were used. The corresponding values for RNA comparisons were 21 (window) and 14 (stringency).

RESULTS

Cloning and sequence analysis of the genes coding for E1 and E2

The genes coding for the RV structural proteins are located at the 3' third of the 40S genomic RNA (Oker-Blom et al., 1984). Using this RNA as template and oligo(dT) as primer, double-stranded cDNA was synthesized, size-fractionated on a sucrose gradient, tailed with dC residues and ligated to dG-tailed pBR322 as described in Methods. Out of 40 hybrid plasmids containing inserts larger than 800 bp, four partly overlapping ones (pKTH345, 348, 349 and 351; Fig. 1) were selected for further analyses; the sizes of the inserts were about 2000, 1500, 800 and 1000 bp, respectively.

We first analysed the RV specificity of the inserts by Northern blot analysis with RNA isolated from RV-infected cells and 32P-labelled nick-translated cDNA inserts as probe. Fig. 2(a; lanes 1, 3, 5 and 7) shows that all clones hybridized to both the 40S and the 24S RV RNA species, whereas no hybridization was detected using RNA from uninfected cells (lanes 2, 4, 6, 8). To ensure that each cDNA insert corresponded to a linear stretch of nucleotides within the RV RNA, we performed a colinearity analysis. The cDNA inserts were denatured and hybridized to 32P-labelled virion 40S RNA. After incubation with S1 nuclease to hydrolyse single strands, the DNA–RNA hybrids formed were analysed on an agarose gel. Fig. 2(b) shows that each insert protected an RNA fragment corresponding in size to the estimated length of the insert. The result confirmed that the inserts are colinear with the RV RNA and are not derived from e.g. defective interfering RNAs. No hybrids were formed when pBR322 without an insert was used in the hybridization reaction (data not shown).
Fig. 2. Analysis of RV-specificity and colinearity of the cDNA inserts. (a) Total RNA from RV-infected (lanes 1, 3, 5, 7) and mock-infected (lanes 2, 4, 6, 8) Vero cells was denatured with DMSO and glyoxal and fractionated on a 1% agarose gel. The RNA was transferred to a nitrocellulose filter. Filter strips were cut out and hybridized with 32P-labelled nick-translated hybrid plasmids pKTH345 (lanes 1 and 2), pKTH351 (lanes 3 and 4), pKTH349 (lanes 5 and 6) and pKTH348 (lanes 7 and 8). After extensive washings the filters were dried and autoradiographed. The positions of the 40S and 24S RNA species are indicated. (b) Purified RV 40S RNA metabolically labelled with 32P was hybridized to linearized and denatured hybrid plasmids in the presence of 70% formamide at 55 °C for 4 h (Söderlund et al., 1981). After dilution single-stranded RNA was hydrolysed with S1 nuclease and the remaining DNA/RNA hybrids were fractionated on a neutral 1% agarose gel. The gels were then dried and autoradiographed. DNA/RNA hybrids formed with pKTH351 (lane 1), pKTH349 (lane 2), pKTH345 (lane 3) and pKTH348 (lane 4) are shown. The positions and sizes in base pairs of various restriction fragments used as markers are indicated to the right.

To sequence the inserts, restriction enzyme fragments derived from the four clones were subcloned in M13mp8 and 9 in both orientations. The sequencing strategy for the subcloned fragments with the dideoxy method is shown in Fig. 1 (middle part). Unusually G/C-rich regions caused problems with compression. Many regions were therefore also sequenced by the Maxam & Gilbert method (Fig. 1, lower part). The insert of pKTH348 had at one end a stretch of 12 A residues and it was therefore assumed to correspond to the extreme 3' end of the 40S RNA.

The assembled nucleotide sequence obtained from the above cDNA inserts covered a continuous sequence starting from the poly(A) tract and extending 2451 nucleotides toward the 5' end. The sequence is presented in Fig. 3. The first nucleotide at the 5' end of the gene coding for mature E2 (i.e. minus its putative signal sequence; see below) has been assigned number 0. The nucleotides upstream have been given negative and those extending towards the 3' poly(A) tract positive numbers.

Primary structure and characteristics of E1 and E2

We have previously shown that the structural proteins of RV are made as a 110000 mol. wt. precursor (p110) (Oker-Blom et al., 1984). Thus, one single continuous open reading frame (ORF) was expected to be found within the determined sequence. One such ORF was identified in the mRNA-sense strand between nucleotides −99 and an UAG codon at positions 2289 to 2291, corresponding to 796 amino acids (Fig. 3 and 4). This ORF presumably extends further upstream into the capsid protein gene. The two other reading frames contained several stop
Fig. 3. Nucleotide sequence and deduced amino acid sequence of RV E1 and E2 genes. The sequence is presented as DNA corresponding to the mRNA-sense in the 5’ to 3’ orientation. The deduced amino acid sequence in one letter code of the large ORF is shown. Numbering of the nucleotide sequence and the amino acid sequence is as described in the text. The amino termini of mature E1 and E2 glycoproteins are indicated. Potential glycosylation sites for N-linked glycans are boxed. The four hydrophobic regions are underlined. Tryptic peptides that were sequenced are indicated by lines above the amino acid sequences. The amino-terminal amino acid sequences of E1 and E2 are from Kalkkinen et al. (1984). The palindromic sequence shown in Fig. 6 at the carboxy terminus of E1 is indicated by the two arrows.
codons (Fig. 4). It is noteworthy, however, that the number of stop codons is surprisingly low. In the complementary (anti-mRNA-sense) strand, there is one large ORF between an AUG at positions 1739 to 1737 (Fig. 3) and an in-phase UAG at positions 215 to 213 (numbers refer to Fig. 3) corresponding to 508 amino acids.

Between the termination codon and the poly(A) tract, there is a non-translated region of 58 nucleotides. This region lacks the AAUAAA sequence known to be a signal for polyadenylation in RNAs transcribed by RNA polymerase II.

Kalkkinen et al. (1984) have determined the amino-terminal sequences of mature E2 and E1 to be X-L-Q-P-R-A-X-M-A-A and E-E-A-F-X-Y-L-X-X-A-P-G, respectively. These sequences were located in the deduced amino acid sequence between positions 1 and 10 and 283 and 294, respectively (Fig. 3, lines above amino acid sequence). We have designated the amino-terminal amino acid (glycine) of mature E2 as number 1. The amino acid sequence of two tryptic peptides derived from E2 and one from E1 were also determined to confirm that the deduced ORF was in the correct phase. These peptides, the sequences of which were determined to be V-G-Q-H-Y-R, G-V-P-A-H-P-G-A-R and L-V-G-A-T-P-E-R, were located between residues 47 to 52 (E2), 198 to 206 (E2) and 527 to 534 (E1) (Fig. 3, indicated by lines above amino acid sequence).

Mature E2 and E1 are preceded by 21 (positions -23 to -3) and 20 (positions 263 to 282) uncharged, mainly hydrophobic amino acids (Fig. 3, underlined). These sequences are typical of signal peptides mediating translocation of the proteins through the endoplasmic reticulum membrane. At the carboxy-terminal end of E1, there is another stretch of hydrophobic amino acids, 27 residues in length (positions 724 to 750). This probably represents the trans-membrane domain anchoring E1 to the lipid bilayer. At the carboxy-terminal end of E2, a few residues upstream from the putative signal sequence of E1, there is a similar 19 residue hydrophobic sequence (positions 237 to 255), which could serve as the membrane anchor domain for E2. The hydrophobic nature of these four regions is also clearly evident from the hydropathy plot determined for the whole ORF (Fig. 5).

The size of mature E1 (481 residues; mol. wt. 51502) is clearly defined by the deduced sequence. The exact size of E2 could not be defined, since the carboxy-terminal sequence of E2 could not be identified (see Discussion). There is a stretch of seven amino acids including five arginines between the putative signal sequence of E1 and the putative trans-membrane domain of E2, which could serve as a cleavage site for a trypsin-like protease. Assuming that cleavage occurs within this region, E2 would have a size of about 260 amino acids (mol. wt. about 28000).

This value is in good agreement with that obtained by SDS–PAGE (Oker-Blom et al., 1983).
Sequence of rubella virus glycoproteins

Fig. 5. Hydropathy plot (positive values represent hydrophobic residues, negative values represent hydrophilic ones), distribution of cysteines (vertical bars) and potential glycosylation sites for N-linked glycans (Y) of the E1 and E2 genes. The amino termini of E1 and E2 are indicated.

Fig. 6. Possible base-paired stem structure, which could form within the palindromic sequence present at the carboxy terminus of E1. Numbers indicate the positions in the nucleotide sequence presented in Fig. 3.

The nucleotide sequence reveals four potential sites for asparagine-linked glycosylation in E2 (N at positions 53, 71, 110, 129 in Fig. 3) and three in E1 (positions 358, 459, 491). These sites have also been marked in the hydropathy plot (Fig. 5). The amino acid composition of E1 and E2 is listed in Table 1. Both proteins are rich in cysteine (5.3% in E2 and 5.0% in E1). Assuming a charge at neutral pH of +1 for lysine and arginine, +0.5 for histidine, −1 for aspartic acid and glutamic acid, E2 (262 amino acids) has a net charge of +18 and E1 +4.5.

Characteristics of the nucleotide sequence

The base composition of the determined nucleotide sequence (2451 residues) was found to be 14.4% A, 38.7% C, 29.7% G and 17.2% U. Thus, the sequence is unusually G/C-rich (68.3%). This is particularly evident in the E2 gene, where the G/C content is 71.4%. In one region in the middle of the E2 gene (residues 330 to 380) there is about 60% C residues followed shortly downstream by a G-rich region (40%).

In the region coding for the carboxy-terminal end of E1, there is an almost perfect inverted complementary nucleotide sequence (Fig. 3, positions 2263 to 2292), which could form a hairpin structure consisting of 13 bp including one mismatch. The ΔG for this hairpin structure was calculated to be −127.6 kJ/mol (Salser, 1977).

DISCUSSION

The amino acid sequence deduced from the nucleotide sequence confirms our previous results showing that (i) E2 and E1 are made as part of a precursor (p110), (ii) E2a and E2b are encoded by one gene, and (iii) E2 is located amino terminally and E1 carboxy terminally in p110 (Oker-
Table 1. Amino acid compositions of E1 and E2

<table>
<thead>
<tr>
<th>Amino acid residue</th>
<th>Number of residues in sequence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala (A)</td>
<td>E1 55 (11.4) E2 31 (11.8)</td>
</tr>
<tr>
<td>Phe (F)</td>
<td>13 (2.7) 6 (2.3)</td>
</tr>
<tr>
<td>Lys (K)</td>
<td>9 (1.9) 1 (0.4)</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>47 (9.8) 28 (10.7)</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>39 (8.1) 20 (7.6)</td>
</tr>
<tr>
<td>Cys (C)</td>
<td>24 (5.0) 14 (5.3)</td>
</tr>
<tr>
<td>Gly (G)</td>
<td>46 (9.6) 27 (10.3)</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>33 (6.9) 25 (9.5)</td>
</tr>
<tr>
<td>Gln (Q)</td>
<td>17 (3.5) 8 (3.1)</td>
</tr>
<tr>
<td>Val (V)</td>
<td>41 (8.5) 12 (4.6)</td>
</tr>
<tr>
<td>Asp (D)</td>
<td>16 (3.3) 9 (3.4)</td>
</tr>
<tr>
<td>His (H)</td>
<td>17 (3.5) 20 (7.6)</td>
</tr>
<tr>
<td>Met (M)</td>
<td>3 (0.6) 6 (2.3)</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>24 (5.0) 20 (7.6)</td>
</tr>
<tr>
<td>Trp (W)</td>
<td>15 (3.1) 9 (3.4)</td>
</tr>
<tr>
<td>Glu (E)</td>
<td>21 (4.4) 4 (1.5)</td>
</tr>
<tr>
<td>Ile (I)</td>
<td>10 (2.1) 2 (0.8)</td>
</tr>
<tr>
<td>Asn (N)</td>
<td>12 (2.5) 5 (1.9)</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>23 (4.8) 11 (4.2)</td>
</tr>
<tr>
<td>Tyr (Y)</td>
<td>16 (3.3) 4 (1.5)</td>
</tr>
<tr>
<td>Total</td>
<td>481</td>
</tr>
</tbody>
</table>

* Residues 283 to 763 (Fig. 3).
† Residues 1 to 262 (Fig. 3) representing E2 without the amino-terminal signal sequence and extending to the putative signal sequence of E1.

Blom et al., 1983, 1984; Oker-Blom, 1984; Kalkkinen et al., 1984). The sizes of E1 and E2 deduced from the sequence (mol. wt. 51 502 and about 28 000) are in good agreement with those obtained by SDS–PAGE for the unglycosylated forms (53 000 and 30 000, respectively) (Oker-Blom et al., 1983).

The nucleotide sequence covering the E1 and E2 genes is unusually G/C-rich (68%). This is also reflected in the low number of stop codons in the two closed reading frames (12 and 11, respectively, Fig. 4). These factors caused some problems in obtaining a correct and reliable sequence. Long G/C-rich regions were therefore sequenced by both sequencing methods and in both directions to resolve compressions. Sequence errors in combination with the scarcity of stop codons in any reading frame may easily result in frameshifts and segments of erroneously deduced amino acid sequences. We believe that our sequence is correct, since the amino-terminal amino acid sequences of five tryptic peptides were localized in the same reading frame (the large ORF). Secondly, our deduced amino acid sequence is, with the exception of a few single base differences, identical with the E1 sequence recently determined by Frey et al. (1986) (see below).

Within the large ORF in the mRNA-sense strand, the previously determined amino-terminal amino acid sequences of mature (virion) E2 and E1 could readily be identified. Both proteins are preceded by a stretch of hydrophobic amino acids typical for signal peptides. This means that E1 is likely to contain an internal signal peptide for translocation through the endoplasmic reticulum membrane, a situation similar to that of the alphaviruses (Strauss & Strauss, 1983). Both proteins also contain a carboxy-terminal hydrophobic region that probably represents the transmembrane domain anchoring the proteins to the lipid bilayer. The potential carboxy-terminal extension (cytoplasmic tail) of E1 is 13 residues in length, whereas the corresponding tail of E2 could not be determined from the sequence. It is likely that the putative signal sequence of E1 is cleaved off the carboxy-terminus of E2. There is a cluster of arginines between the E1 and E2 genes that could be the site for trypsin-like processing.
Sequence of rubella virus glycoproteins

When E2a and E2b were previously digested with a mixture of carboxypeptidases A and B (Kalkkinen et al., 1984) low molar amounts of leucine, isoleucine, tyrosine, alanine, histidine and lysine were liberated. No amino acid sequence corresponding to a cluster of these amino acids could, however, be identified near the N terminus of E1 in the deduced amino acid sequence. Thus, the exact location of the carboxy terminus of E2 remains unidentified. Most probably the small amounts of amino acids obtained originated from the carboxy termini of peptides resulting from contaminating endopeptidase activity in the carboxypeptidase preparations used. It is known that chymotrypsin, generating peptides with hydrophobic carboxy-terminal amino acids, is the most common contaminant of carboxypeptidases A and B.

Between the two membrane proteins of alphaviruses there is a short ‘intergenic peptide’, the 6K peptide (55 to 60 residues), that includes the signal sequence for the downstream membrane protein (Hashimoto et al., 1980; Garoff et al., 1980b; Rice & Strauss, 1981). If such a peptide exists also between E2 and E1 of RV it is much shorter (about 25 residues). To confirm this, the exact cleavage site between the two proteins must be determined.

The E2 protein can be resolved into two bands by SDS–PAGE (Oker-Blom et al., 1983). Based on previous results we have concluded that this may be due to heterogeneous glycosylation. There are four potential sites for asparagine-linked glycosylation in the E2 sequence, possibly allowing attachment of different types of glycans (different number of branches in complex-type glycans, varying degree of sialylation etc.). The large shift in mobility of E2 synthesized in the presence of tunicamycin suggests that all four sites might be used. E1 contains three potential glycosylation sites. Since the difference in mobility between the glycosylated and unglycosylated forms represents only 5000 daltons, it remains to be determined whether all three sites are in fact used.

A large ORF, corresponding to 508 amino acids, was identified between an AUG and a stop codon in the anti-mRNA-sense strand. To date, no positive-strand viruses have been shown to encode proteins in the anti-mRNA sense strands. It will therefore be interesting to study whether this ORF is indeed used.

The 3’ untranslated region is only 58 nucleotides in length. This is substantially shorter than in the alphaviruses Semliki Forest (264 residues; Garoff et al., 1980a), Sindbis (318; Rice & Strauss, 1981) and Ross River (524; Dalgarno et al., 1983) viruses. Sequences important for virus replication are known to be located within a short region close to the poly(A) tract and to be conserved among different members of the alphaviruses (Ou et al., 1981; Strauss & Strauss, 1983; Levis et al., 1986). No homology was found between the alphaviruses and RV within this region. At the 3’ end of the E1 gene, there is a palindromic sequence that could form a stable hairpin structure. Whether this region plays any role in the replication cycle of RV remains to be studied.

Recently, Frey et al. (1986) and Nakhasi et al. (1986) have reported the sequence of the RV E1 gene of the Therien strain and M33 strain, respectively. Our deduced amino acid sequence differs from that of Frey et al. only in three positions (Val659, Ser691, Thr737 in our sequence are replaced by Leu597, Thr630 and Ile619 in their sequence). All changes are due to single base differences. In addition, there is a single base difference in the 3’ untranslated region (T2329 in our sequence is replaced by C2329 in theirs), and next to the poly(A) tract our sequence contains an extra AU dinucleotide. These differences may be due to sequence errors or to mutations that have occurred during cultivation of the Therien strain. The sequence reported by Frey et al. (1986) also contains a short region preceding the signal sequence of E1. Here our sequence runs -C-R-R-A-C-R-R-G-A and theirs -C-R-R-R-L-S-P-P-R, using the one-letter amino acid code.

There are a total of 13 differences between our sequence and that reported by Nakhasi et al. (1986). They include (i) seven single amino acid differences (at positions Ser245 in our sequence (Leu245 in theirs), Ala296 (Pro14), Ala299 (Thr17), Ala350 (Thr68), Gly509 (Ala241), Arg618 (Gly335) Val662 (Phe379), (ii) two amino acids are missing from their sequence (Arg862, Thr819), (iii) a large region, where the amino acid sequence is read from different reading frames (between our nucleotides 1527 and 1835) and which contains numerous nucleotide differences, (iv) starting at nucleotide 1995 (our sequence), their deduced amino acid sequence runs in a different reading frame and terminates prematurely in a TGA codon that does not exist in our
sequence, (v) due to a short frameshift, three amino acids before the cleavage site of E1 have been changed from A-Y-G to P-M-A, (vi) upstream from our nucleotide residue 722, the nucleotide sequences are totally different. Since our sequence and that of Frey et al. (1986) are almost identical in the E1 region (see above), we conclude that the sequence reported by Nakhasi et al. (1986) is likely to contain numerous sequencing errors. Since different strains are being studied, however, some of the single residue changes will be real.

RV is the only member of the Rubivirus genus within the Togaviridae family (Porterfield et al., 1978) and no serological cross-reactivity has been shown to exist between RV and members of the Alphavirus genus. The complete nucleotide sequences of the genes for the structural proteins of several alphaviruses, including Semliki Forest virus (Garoff et al., 1980a, b), Sindbis virus (Rice & Strauss, 1981), Ross River virus (Dal-garno et al., 1983) and Venezuelan equine encephalitis virus (Kinney et al., 1986) have been determined. Comparison of these sequences to that of our sequence displayed no significant homology at either the nucleotide or the amino acid level (see Methods for parameters used). This also included the 3' untranslated region (see above). Thus, although the general strategy of the gene expression of RV and alphaviruses seems to be similar, both the organization of the intergenic region between E1 and E2 and the non-homologous sequence support the view that RV and alphaviruses are not very closely related evolutionarily and justifies the classification of RV in its own genus within the family Togaviridae.

We wish to thank Ms. Annikki Kallio for excellent technical assistance in the cultivation of RV and the sequence determination, and Kai Korpela for performing some of the initial Maxam & Gilbert sequencing. We also thank Ulf Hellman and Christer Wernstedt, Ludwig Institute for Cancer Research, Uppsala Branch, for determining the nucleotide sequences are totally different. Since our sequence and that of Frey almost identical in the E1 region (see above), we conclude that the sequence reported by Nakhasi et al. (1986) is likely to contain numerous sequencing errors. Since different strains are being studied, however, some of the single residue changes will be real.

REFERENCES


Sequence of rubella virus glycoproteins


(Received 21 May 1987)