Genomic analysis of diverse rubella virus genotypes

Yumei Zhou,1,2 Hiroshi Ushijima2 and Teryl K. Frey1

1Department of Biology, Georgia State University, Atlanta, GA, USA
2Department of Developmental Medical Sciences, Institute of International Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Based on the sequence of the E1 glycoprotein gene, two clades and ten genotypes of Rubella virus have been distinguished; however, genomic sequences have been determined for viruses in only two of these genotypes. In this report, genomic sequences for viruses in an additional six genotypes were determined. The genome was found to be well conserved. The viruses in all eight of these genotypes had the same number of nucleotides in each of the two open reading frames (ORFs) and the untranslated regions (UTRs) at the 5′ and 3′ ends of the genome. Only the UTR between the ORFs (the junction region) exhibited differences in length. Of the nucleotides in the genome, 78% were invariant. The greatest observed distance between viruses in different genotypes was 8.74% and the maximum calculated genetic distance was 14.78 substitutions in 100 sites. This degree of variability was similar among regions of the genome with two exceptions, both within the P150 non-structural protein gene: the N-terminal region that encodes the methyl/guanylyltransferase domain was less variable, whereas the hypervariable domain in the middle of the gene was more divergent. Comparative phylogenetic analysis of different regions of the genome was done, using sequences from 43 viruses of the non-structural protease (near the 5′ end of the genome), the junction region (the middle) and the E1 gene (the 3′ end). Phylogenetic segregation of sequences from these three genomic regions was similar with the exception of genotype 1B viruses, among which a recombinational event near the junction region was identified.

INTRODUCTION

Rubella virus is an important human pathogen that causes an acute, contagious disease known as rubella, 3-day measles or German measles, and severe birth defects (known as congenital rubella syndrome) when infection occurs during the first trimester of pregnancy (Chantler et al., 2001). Rubella virus is the single member of the genus Rubivirus in the family Togaviridae and is an enveloped, single-stranded, positive-polarity RNA virus with a genome of approximately 10 kb. The genome contains two long open reading frames (ORFs): the 5′-proximal ORF (NSP-ORF) encodes two non-structural proteins, P150 and P90, that function in RNA replication, and the 3′-proximal ORF (SP-ORF) encodes three structural proteins: the capsid protein, C, and two envelope glycoproteins, E1 and E2. The SP-ORF is translated from a subgenomic RNA synthesized in infected cells (Frey, 1994). The genome also contains untranslated regions (UTRs) at its 5′ and 3′ ends and between the ORFs (known as the junction region).

Although rubella occurs worldwide, vaccination efforts with live-attenuated vaccines have been concentrated in developed countries. Currently, approximately 50% of countries have national vaccination efforts against rubella (Robertson et al., 2003). Isolation and genetic sequencing of rubella viruses has been most thorough in countries pursuing elimination (Bosma et al., 1996; Frey et al., 1998; Icenogle et al., 2006; Katow, 2004; Katow et al., 1997a, b; Reef et al., 2002; Saitoh et al., 2006); however, collections have recently been assembled from other regions of the world (Donadio et al., 2003; Katow, 2004; Zheng et al., 2003c, c). Recently, a standard taxonomy for rubella viruses was adopted based on sequences of a standard window within the E1 gene and supported by sequencing of the SP-ORF of selected viruses (WHO, 2005). The taxonomy consists of two clades (corresponding to the previous genotypes I and II [Frey et al., 1998; Zheng et al., 2003a]) containing a total of ten genotypes, seven in clade 1 (1a, 1B, 1C, 1D, 1E, 1F and 1g) and three in clade 2 (2A, 2B and 2c); the genotypes designated in lower case are provisional. Within the E1 gene, maximal variation among clade 1 viruses is 5.8%, that among clade 2 viruses is 8.0%, and it is 8.2% between the two clades (Zheng et al., 2003a). Geographically, clade 1 viruses circulate worldwide, whilst clade 2 viruses thus far have been restricted to Eurasia.

Thus far, ten complete genomic sequences of Rubella virus have been reported, which represent only two of the ten

Supplementary tables are available in JGV Online.
genotypes (eight sequences of genotype 1a viruses and two sequences of genotype 2A viruses). Among these sequences, genetic genomic variability is similar to that in the E1 gene, with the exception of a 'hypervariable region' (HVR) of greater variability in the middle of the P150 non-structural protein gene (Hofmann et al., 2003; Zheng et al., 2003b). Given the lack of representation of the majority of the genotypes in the current genomic database, the first goal of this study was to expand the number of genomic sequences, using viruses in our collection from six additional genotypes (1B, 1C, 1D, 1E, 2B and 2C). The second goal of this study was to extend phylogenetic analysis to 5′ regions of the genome, which had not previously been done. To this end, the sequence of the non-structural protease-encoding region within the P150 gene was determined and compared phylogenetically with the sequences of the junction region and the E1 gene from 43 viruses representing eight genotypes.

METHODS

Viruses, cells, RNA extraction, cDNA amplification and DNA sequencing. The viruses analysed in this study are listed in Table 1 (genomic sequences) and Supplementary Table S1, available in JGV Online (genomic regions). The Cba strain was provided by Dr Marta Zapata, University of Cordoba, Cordoba, Argentina. A monolayer of Vero cells (25 cm² T-flask or 60 mm plate) was infected with each virus. Three to five days post-infection, the culture medium was removed and total cellular RNA was extracted by using Tri-Reagent (Molecular Research Center) using the manufacturer’s protocol. The RNA extracted from one 25 cm² T-flask or one 60 mm plate was resuspended in 50 μl double-distilled H₂O and stored at −80 °C until use. cDNA was synthesized in a 20 μl reaction volume reaction containing 5 μl denatured (95 °C, 5 min) RNA template, 4 μl 5× Reverse Superscript buffer (Invitrogen), 4 μl 2.5 mM dNTPs, 1 μl 0.1 M dithiothreitol, 1 μl 4 μM 3′E1/808 reverse primer (5′-TTTTTTTTTCTATACACGCAAC-3′; Tₙ followed by the complement of nt 9751–9762 of the rubella virus genome), 1 μl (40 units) RNasin (Promega) and 1 μl (200 units) Superscript reverse transcriptase III (Invitrogen). The reaction was incubated at 55 °C for 60 min and then stored at −20 °C prior to use in PCR. Each 50 μl PCR contained 25 μl 2× GC buffer (TaKaRa), 8 μl 2.5 mM dNTPs, 3 μl cDNA template, 1 μl 40 μM appropriate forward and reverse primers and 0.5 μl (2.5 units) LA Taq polymerase (TaKaRa). Cycling parameters were determined according to the manufacturer’s (TaKaRa) recommendations. For genomic sequencing, 10–11 overlapping fragments encompassing the entire genome were amplified by using appropriate primer pairs. The primers used to amplify the genomic regions are listed in Supplementary Table S2, available in JGV Online. Amplified fragments were purified following agarose-gel electrophoresis by using a QiAquick gel extraction kit (Qiagen). Sequencing reactions were performed bidirectionally by using appropriate primers and cycle-sequencing kits (ABI PRISM BigDye Terminator v. 3.1; PE Applied Biosystems) and resolved by using a 3100 Genetic Analyzer (Applied Biosystems). The 5′ and 3′ ends of the genome were determined by using a 5′/3′ FirstChoice RLM-RACE kit (Ambion Inc.).

Sequence analysis. For cataloguing and storage, sequences were input into free online sequence-alignment software (ALIGN Query, GENESTREAM search network server IGH, Montpellier, France; http://xylian.igh.cnrs.fr/bin/align-guess.cgi). The assembled nucleotide sequences were aligned by using the CLUSTAL_W multiple sequence-alignment program version 1.8 (Henikoff & Henikoff, 1994) and the PileUp program in the GCG software package (Genetics Computer Group, version 11.0; Accelrys Inc.). The TN93 substitution model (Tamura & Nei, 1993) with discrete gamma-distributed rate heterogeneity with eight gamma rate categories (Yang, 1994) (TN93+I+G model) was used as a substitution model for phylogenetic reconstruction, as it was found statistically to be the best fit for our datasets. Maximum-likelihood (ML) phylogenetic analysis

<table>
<thead>
<tr>
<th>Virus</th>
<th>Isolation site and year</th>
<th>Genotype</th>
<th>GenBank accession no.</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fth_USA64</td>
<td>Connecticut, USA, 1964</td>
<td>1a</td>
<td>M15240</td>
<td>Dominguez et al. (1990)</td>
</tr>
<tr>
<td>RA27_USA64</td>
<td>Pennsylvania, USA, 1964</td>
<td>1a</td>
<td>L78917</td>
<td>Pugachev et al. (1997)</td>
</tr>
<tr>
<td>M33_USA61</td>
<td>New Jersey, USA, 1961</td>
<td>1a</td>
<td>X05259, X72393</td>
<td>Clarke et al. (1987)</td>
</tr>
<tr>
<td>TO-W_JAP67</td>
<td>Toyama, Japan, 1967</td>
<td>1a</td>
<td>AB047330</td>
<td>Kakizawa et al. (2001)</td>
</tr>
<tr>
<td>TO-V_JAP67</td>
<td>Toyama, Japan, 1967</td>
<td>1a</td>
<td>AB047329</td>
<td>Kakizawa et al. (2001)</td>
</tr>
<tr>
<td>ULR_GER84</td>
<td>Bratislava, Slovakia, 1974</td>
<td>1a</td>
<td>AF435865</td>
<td>Hofmann et al. (2003)</td>
</tr>
<tr>
<td>SUR_SVK74</td>
<td>Bratislava, Slovakia, 1974</td>
<td>1a</td>
<td>AF435866</td>
<td>Hofmann et al. (2003)</td>
</tr>
<tr>
<td>Cba_ARG88</td>
<td>Cordoba, Argentina, 1988</td>
<td>1B</td>
<td>DQ085339</td>
<td>This report</td>
</tr>
<tr>
<td>GUZ_GER92</td>
<td>Stuttgart, Germany, 1992</td>
<td>1B</td>
<td>DQ388280</td>
<td>This report</td>
</tr>
<tr>
<td>Anim_MEX97</td>
<td>Baja California, Mexico, 1997</td>
<td>1C</td>
<td>DQ085341</td>
<td>This report</td>
</tr>
<tr>
<td>JCC_NZL91</td>
<td>Auckland, New Zealand, 1991</td>
<td>1D</td>
<td>DQ388281</td>
<td>This report</td>
</tr>
<tr>
<td>6423_ITA97</td>
<td>Pavia, Italy, 1997</td>
<td>1E</td>
<td>DQ085343</td>
<td>This report</td>
</tr>
<tr>
<td>BRDH_CN80</td>
<td>Beijing, China, 1980</td>
<td>2A</td>
<td>AY258323</td>
<td>Zheng et al. (2003b)</td>
</tr>
<tr>
<td>BRI_CN79</td>
<td>Beijing, China, 1979</td>
<td>2A</td>
<td>AY258322</td>
<td>Zheng et al. (2003b)</td>
</tr>
<tr>
<td>AN5_KOR96</td>
<td>Seoul, Korea, 1996</td>
<td>2B</td>
<td>DQ085342</td>
<td>This report</td>
</tr>
<tr>
<td>1-11_ISR68</td>
<td>Tel-Aviv, Israel, 1968</td>
<td>2B</td>
<td>DQ085338</td>
<td>This report</td>
</tr>
<tr>
<td>C74_RUS97</td>
<td>Moscow, Russia, 1997</td>
<td>2c</td>
<td>DQ085340</td>
<td>This report</td>
</tr>
<tr>
<td>C4_RUS67</td>
<td>Moscow, Russia, 1967</td>
<td>2c</td>
<td>DQ388279</td>
<td>This report</td>
</tr>
</tbody>
</table>
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), fastDNAm1 (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 each of these 19 viruses, with three exceptions, the genomes were invariant. Not surprisingly, the parameter of rate heterogeneity, ω, 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 ω = 1.35. These small ω 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

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 surprising 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, ω, 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 ω = 1.35. These small ω 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.
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 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 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
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, hyper-variable 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

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

*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–G)/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.
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
Conservation of the MT domain sequence has also been higher than to the encoded amino acid (Pugachev et al., 1998). The HVR encodes a proline- and arginine-rich internal HVR exhibiting greater variability. Although the two genomic regions were equally divergent. Both observed and genetic distances 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 \(z\). Sequence collections with low \(z\) 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 \(z\) is 1 (or >1). The low \(z\) 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, Suguanthi Suppiah and Hui Zhao for preliminary sequence determinations and Ping Jiang for processing sequencing reactions and gels.

REFERENCES


