Full-length genome analysis of natural isolates of vesicular stomatitis virus (Indiana 1 serotype) from North, Central and South America

Luis L. Rodriguez, Steven J. Pauszek, Thomas A. Bunch† and Kate R. Schumann‡

Plum Island Animal Disease Center, Agricultural Research Service, US Department of Agriculture, Orient Point, Long Island, PO Box 848 Greenport, NY 11944-0848, USA

Most studies on the molecular biology and functional analysis of vesicular stomatitis virus Indiana 1 serotype (VSV-IN1) are based on the only full-length genomic sequence currently deposited in GenBank. This sequence is a composite of several VSV-IN1 laboratory strains passaged extensively in tissue culture over the years and it is not certain that this sequence is representative of strains circulating in nature. We describe here the complete genomic sequence of three natural isolates, each representing a distinct genetic lineage and geographical origin: 98COE (North America), 94GUB (Central America) and 85CLB (South America). Genome structure and organization were conserved, with a 47 nucleotide 3’ leader, five viral genes – N, P, M, G and L – and a 59 nucleotide 5’ trailer. The most conserved gene was N, followed by M, L and G, with the most variable being P. Sequences containing the polyadenylation and transcription stop and start signals were completely conserved among all the viruses studied, but changes were found in the non-transcribed intergenic nucleotides, including the presence of a trinucleotide at the M–G junction of the South American lineage isolate. A 102–189 nucleotide insertion was present in the 5’ non-coding region of the G gene only in the viruses within a genetic lineage from northern Central America. These full-length genomic sequences should be useful in designing diagnostic probes and in the interpretation of functional genomic analyses using reverse genetics.

Introduction

Vesicular stomatitis virus (VSV) is the prototype virus of the family Rhabdoviridae, genus Vesiculovirus. Two serotypes, New Jersey (VSV-NJ) and Indiana (VSV-IN1), cause vesicular disease in livestock in North America, Central America and northern South America. Two subtypes of the Indiana serotype, Cocal (VSV-IN2) and Alagoas (VSV-IN3), cause vesicular disease in livestock in Brazil and Argentina (Rodriguez & Nichol, 1999). VSV-IN1 is widely used as an RNA virus laboratory model because it readily grows to high titres in a variety of tissue culture cells and in laboratory animal models. Due to the high error rate and lack of proofreading activity of its RNA polymerase, VSV has been extensively used as a model for virus evolution (Holland et al., 1982; Domingo et al., 1996).

The viral genome consists of a linear, single-stranded, negative-sense RNA molecule of approximately 11 kb encoding five genes: the nucleocapsid protein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G) and polymerase (L). A 47 nucleotide leader (Le) RNA is transcribed from the 3’ genomic terminus. Only 70 nucleotides – two nucleotides at each of the four gene junctions, three nucleotides at the Le–N junction and a 59 nucleotide trailer (Tr) sequence at the 5’ terminus – are not transcribed (Gallione et al., 1981). Two additional small proteins (C and C’) are encoded in a second open reading frame within the P gene (Spiropoulou & Nichol, 1993). Transcription of the viral genes occurs in sequential order from a single promoter at the 3’ end of the genome resulting in decreasing amounts of each transcript in
the order: 3’ Le–N–P–M–G–L 5’ (Iverson & Rose, 1982). Individual monocistronic mRNAs are transcribed by the viral transcriptase by a mechanism of pause and reinitiation controlled by a 23 nucleotide conserved sequence located at each gene junction, which contains the polyadenylation and transcription initiation signals.

Despite its widespread use as a laboratory model, only one full-length genomic sequence of VSV-IN1 is currently available in GenBank (the same sequence is found under two accession numbers: NC_001560 and J02428). This sequence is based on a composite of several laboratory strains of VSV-IN1 that have been extensively passaged in tissue culture for many years, mainly the San Juan strain (New Mexico, 1956) and the Mudd–Summers strain (Colorado, prior to 1956). Most studies on the molecular biology and functional analyses of the VSV genome are based on these sequences. However, it is not certain that they are representative of strains circulating under natural conditions.

This work describes the first complete genomic sequencing of three natural isolates of VSV-IN1, each representing a distinct genetic lineage of different geographical origin. We show detailed analyses of all genes and gene junctions for each of these viruses and compare them with those of the laboratory strain previously described in the literature. These full-length genomic sequences should be useful in the design and interpretation of future functional genomic analyses using reverse genetics.

Methods

■ Virus strains and RNA extraction. Three viruses, each representing a distinct genetic lineage, were chosen for full-length sequencing. Virus 98COE (North American lineage) was obtained from an equine tongue epithelium sample from outbreaks occurring in the western USA in 1998. Virus 94GUB (Central American lineage) was obtained from a bovine case originating in central Guatemala in 1994. Virus 85CLB (South American lineage) originated from a bovine clinical case in Antioquia, Colombia, in 1985.

Total RNA was extracted directly from 1 g of macerated epithelium by the acid guanidine thiocyanate method using Trizol (Invitrogen), as previously described (Rodriguez et al., 1997). RNA pellets were resuspended in sterile water and kept at −70 °C until tested. In order to obtain enough RNA for sequencing the genomic termini, virion RNA was extracted similarly from sucrose-gradient-purified virions from passage 1 virus in BHK-21 cells, as described previously (Rodriguez et al., 1993).

■ RT–PCR and cloning. Reverse transcription was carried out using random hexamers (Invitrogen) and SuperScript II RNase H Reverse Transcriptase (Invitrogen), following the manufacturer’s instructions. Alternatively, RT–PCR was carried out using the one-tube 17th RNA PCR System (GeneAmp EZ 17th RNA PCR kit, Perkin-Elmer), as previously described (Rodriguez et al., 1997). To determine the sequence of genomic termini, template RNA was extracted from sucrose-gradient-purified virions and ligated using T4 RNA ligase (Mandi et al., 1991). Primers for PCR and sequencing reactions were designed based on the published sequence of VSV-IN1 (GenBank accession no. J02428) or based on newly obtained sequences (primer sequences are available from the corresponding author on request). PCR reactions were performed using Pfu DNA polymerase (Stratagene), according to the manufacturer’s instructions. Nucleic acid amplifications were performed in a 9600 Perkin-Elmer thermocycler (PE Applied Biosystems) using the temperature profiles within the following ranges, depending on expected product size and primer sequence: 94 °C for 3 min, followed by 35 cycles of 94 °C for 30–60 s, 50–60 °C for 30–60 s, 72 °C for 1–3 min, and a final elongation step at 72 °C for 7–15 min. Products were analysed by electrophoresis on agarose gels and visualized by ethidium bromide staining. To characterize the sequence of the insert at the 5’ (genomic sense) non-coding region of the glycoprotein, we cloned an RT–PCR product comprising the region from the G stop codon to the L start codon of virus 94GUB using the Zero Blunt TOPO PCR (Cloning Kit Version D, Invitrogen). Ten different colonies containing an insert were selected and plasmid DNA was purified for sequencing.

■ DNA sequencing, alignments and phylogenetic analysis. Single-band products were purified directly from the RT–PCR reaction using QIAquick PCR purification kits (Qiagen). Multiple-band products were separated in agarose gels and extracted using the QIAquick gel extraction kit (Qiagen). PCR products were sequenced by dideoxy-sequencing using a BigDye Terminator Sequencing kit on 373XL or 3700 automated sequencing instruments (PE Applied Biosystems), as previously described (Rodriguez et al., 2000). The Sequencer software (GeneCodes) was used to analyse nucleotide sequence fragments and assemble contigs. Consensus sequences were derived from at least two independent forward and reverse sequences, but in most cases there was extensive sequence overlap and at least four sequences in each direction were available. Protein alignments were performed using MIGALIGN (DNAStar) or ClustalX (Thompson et al., 1997). Sliding window analysis was performed with similarity scores obtained from ClustalX using the Gonnet PAM250 protein weight matrix (Thompson et al., 1994). Divergence scores (100-similarity score) for each 30-amino acid window were plotted using Excel (Microsoft). Phylogenetic analysis was performed by maximum parsimony using PAUP 4.0 beta version (available from D. L. Swofford) in a Power Macintosh G3. Maximum parsimony settings included a character weighting of 2:1 transition/transversion (ts/tv) ratio, and the branch-swapping algorithm was tree-bisection-reconnection (TBR). Bootstrap analysis was carried out by performing 1000 replicates.

Results and Discussion

Three viral strains originating from North America (98COE), Central America (94GUB) and South America (85CLB), each representing a distinct genetic lineage of VSV-IN1, were chosen for full-length sequencing (Fig. 1). The total genomic lengths were 11336, 11162 and 11155 nucleotides for 94GUB, 98COE and 85CLB, respectively. The only full-length sequence deposited in GenBank (accession no. J02428) is 11161 nucleotides long. Length differences were due to variation in the non-coding region at the 5’ end of the G gene (3’ end of the G mRNA, located upstream of the L cistron) (Fig. 2). Unlike the rule of six observed in Sendai virus, in which the length of DI particles and genomic lengths are always a number divisible by six (Kolakofsky et al., 1998), no particular rules were observed in the length of the VSV-IN1 genomes. Genome structure and organization were conserved among all the viruses studied, with a 47 nucleotide 3’ leader, five viral genes (N, P, M, G and L) and a 59 nucleotide 5’ trailer sequence.
Genome analysis of natural VSV isolates

Genomic termini

We sequenced the genomic termini directly from RT–PCR products using L forward and N reverse primers on purified virion RNA ligated with T4 RNA ligase. The terminal 50 nucleotides at the 3’ terminus were highly conserved among all the viruses with only one nucleotide change (G → A) observed at position 25 in sequence J02428 (Fig. 2). The 5’ terminus was less conserved, with two nucleotide substitutions (A → C) at positions 20 and 26 from the 5’ terminus of 94GUB and 98COE, respectively (Fig. 2). A high degree of sequence conservation was expected for the virus termini in the natural isolates, since these regions contain signals necessary for initiation of transcription and also for replication of the viral RNA (Pattnaik et al., 1995; Whelan & Wertz, 1999).

Non-coding regions and gene junctions

All four non-coding gene junctions in all the viruses studied contained a 23 nucleotide conserved sequence: 3’ AUAC(U)₃NNUUGUUCNNUAAG 5’. These sequences have been the subject of intense investigation since they contain cis-acting signals for transcription stop, polyadenylation and transcription start of each mRNA. As expected, the sequence 3’ AUAC(U)₃ 5’ was conserved in all viruses since it is critical for both termination and polyadenylation of mRNA, and in vitro studies have shown very little tolerance to mutations at these sites (Whelan & Wertz, 1999).

The (U)₃ polyadenylation signal was followed by two non-transcribed nucleotides. These non-transcribed intergenic nucleotides are an important part of the cis-acting signals involved.
Table 1. Intergenic nucleotide diversity among VSV-IN1 lineages

<table>
<thead>
<tr>
<th>Gene junction</th>
<th>North America</th>
<th>Central America</th>
<th>South America</th>
</tr>
</thead>
<tbody>
<tr>
<td>N–P</td>
<td>GA</td>
<td>GA</td>
<td>GA</td>
</tr>
<tr>
<td>P–M</td>
<td>CA</td>
<td>CA</td>
<td>CA</td>
</tr>
<tr>
<td>M–G</td>
<td>GA</td>
<td>GG</td>
<td>GA</td>
</tr>
<tr>
<td>G–L</td>
<td>GA</td>
<td>AA</td>
<td>GA+</td>
</tr>
</tbody>
</table>

* Full-length sequences.
† These intergenic nucleotides were confirmed to be identical in five additional isolates within each genetic lineage.
ND, Not determined.

in termination and reinitiation of transcription in VSV (Barr et al., 1997a; Stillman & Whitt, 1997). The non-transcribed intergenic nucleotides 3′ GA 5′ and 3′ CA 5′ between N and P and between P and M, respectively, were conserved among all viruses sequenced (Table 1). In contrast, the M–G and G–L intergenic nucleotides were more permissive to changes in the natural isolates from Central and South America, with each virus lineage having different intergenic nucleotides at the M–G junction, including the trinucleotide 3′ GGA 5′ in the South American lineage and either 3′ GG 5′ or 3′ GA 5′ in the Central American lineage (Table 1). We determined the sequence of the intergenic nucleotides between M and G and between G and L for five additional viruses from each genetic lineage and found them to be identical to the representative full-length sequence within the North American or South American lineages. However, the G–L junction was variable in the Central American lineage with three different intergenic nucleotides found among the five viruses sequenced (Table 1).

The sequence variability observed at the gene junctions was unexpected, given the importance of these nucleotides in regulating transcription. However, each of the intergenic nucleotide sequences, except for the trinucleotide observed in the South American lineage, had been previously tested in vitro by Barr et al. (1997b) and found to efficiently stop readthrough and attenuate transcription of the downstream mRNA. Our finding of the naturally occurring functional intergenic trinucleotide 3′ GGA 5′ increases the known permissive range of intergenic nucleotides.

**Nucleocapsid**

N is the most abundant viral protein in infected cells. It binds both genomic and replicative intermediate viral RNA and, together with the L and P proteins, is an indispensable component of the transcription–replication complex. We found that N, which is 422 amino acids long, was the most conserved protein among all the viruses (Fig. 3). Only five amino acid substitutions (positions 33, 259 and 360 in 94GUB and positions 11 and 110 in 85CLB) were found among all three viruses sequenced. None of these changes were located within the last 60 amino acids of the C-terminal end of N, where sequences required for interaction with the phosphoprotein and also for encapsidation are believed to be located (Pattnaik et al., 1995). Its high conservation, coupled with the fact that N is the most abundant transcript in infected cells, makes it an ideal target for diagnostic probes.

**Phosphoprotein**

Three major domains have been described in P, which play an essential role in transcription and replication. Domain I (aa 1–137) is responsible for the association of P with L and needs to be phosphorylated in several sites for optimal transcription activity (Takacs et al., 1992; Pattnaik et al., 1997). Domain II (aa 229–242) binds to L and contains two seemingly essential serines, whose phosphorylation by L regulates transcription. Domain III (aa 243–265) is positively charged and is located at the C terminus (Das et al., 1997). Although its length was 265 amino acids in all the viruses studied, P was the most variable of the five structural proteins (Fig. 3). Of a total of 33 amino acid changes, 21 occurred within functional domain I. Six of these changes involved threonines or serines, resulting in a net loss of one phosphorylation site in 94GUB. Domain II was conserved among all viruses, with only one amino acid change at position 237. All basic amino acids in domain III, which are essential for transcription and binding to L, were conserved among all viruses, with only one substitution observed at position 261. Despite the fact that P was the most divergent major viral structural protein (range of divergence 3.4–8.7%), the amino acid changes did not seem to involve any of the critical positions in the functional domains.

**C proteins**

Two small basic proteins, C′ and C (67 and 55 amino acids, respectively), are encoded in a second open reading frame (ORF) within the P gene of VSV-IN1 (Spiropoulou & Nichol, 1993; Kretzschmar et al., 1996). The role of these proteins in
the VSV life cycle remains unknown, since viruses in which C protein expression has been abrogated grow in vitro similarly to the wild-type strain (Kretzschmar et al., 1996). All viruses had the second ORF with conserved start codons for both C' and C. In all cases, the AUG of the smaller C protein was in better context for translation, indicating that it is preferentially translated (Spiropoulou & Nichol, 1993; Kozak, 1986). The C protein was the most divergent among all the VSV proteins (followed by P, which was the most divergent among the structural proteins) (Fig. 3). However, the predicted isoelectric point, dictated by the number of arginines in each protein, remained between 10–86 and 11–28 in all viruses. Interestingly, C was the only VSV gene product with a substantially higher number of non-synonymous versus synonymous substitutions (Fig. 3). This indicates that C is under different selective pressures than P, despite the fact that both proteins are encoded by the same mRNA.

Matrix protein

M is an important component in the virus structure, playing a major role in virus morphology, assembly and budding. It is also involved in inducing host-cell cytopathic effect, inhibition of host gene expression, nuclear transport and apoptosis (Kopecky et al., 2001; Desforges et al., 2001; Petersen et al., 2000; Wagner & Rose, 1996). M, which is 229 amino acids in length, was very conserved at the amino acid level, with only seven differences among all the viruses (positions 54, 133 and 225 in virus J02428; 70 and 221 in 98COE; and 126 and 171 in 94GUB) (Fig. 3). At position 24–27, all viruses had a PPPY motif similar to the late domains in retroviral Gag proteins, which seems to be required for the efficient release of viral particles (Jayakar et al., 2000).

Glycoprotein

G has several functions, including fusion with the cell membrane during virus entry and membrane budding during exit from the cell. It is also the target of neutralizing antibodies and cell-mediated immune responses to VSV (Wagner & Rose, 1996). Trimers of G form the spikes protruding from the viral envelope, which interact with cellular receptors. Functional domains identified in G include glycosylation and palmitoylation sites, a membrane anchor, a cytoplasmic domain and membrane fusion domains (Coll, 1995).

After P, G was the second least conserved of the major viral proteins, with 95–97% amino acid conservation (Fig. 3). There were a total of 41 amino acid differences among the four viruses; 16 of these occurred within 30 amino acids of either the N terminus in the ectodomain, or the C terminus in the cytoplasmic domain. No changes occurred at or near glycosylation or palmitoylation sites. Two amino acid changes (position 258 in 98COE and 259 in 94GUB) were observed at or near epitopes where neutralization-resistant mutations have been selected in vitro (Luo et al., 1988). This could suggest that there is immune selection among these viruses under natural conditions.
conditions. However, no accumulation of changes in these areas of the glycoprotein was observed among viruses from Central and North America previously sequenced (Rodriguez et al., 2000). Two domains involved in fusion, one at amino acid position 118–139 (Shokralla et al., 1998) and the other at position 395–462 (Fredericksen & Whitt, 1998) were completely conserved among the four viruses compared.

**Insertion in the genomic 5' non-coding region of the G gene**

The 5′ (genomic sense) non-coding region of the glycoprotein gene is highly variable both in length and sequence composition, among the genetic lineages of VSV. A reiterative 175 nucleotide insertion of 3′ UUAAAAA 5′ was found between the stop codon of G and the G–L gene junction of 94GUB. We sequenced the G gene 5′ non-coding regions of nine other VSV-IN1 isolates from throughout Central America, and found inserts of variable length only in viruses within one genetic lineage from northern Central America (Guatemala, Honduras, El Salvador) and not in viruses from southern Central America, North or South America (Fig. 1). Insertions had been previously noticed in the VSV-IN1 G mRNA by Bilsel & Nichol (1990). Until now, it was not clear whether this insert was added by the viral polymerase by stuttering during transcription of the mRNA, or if it was present in the viral genome. In order to clarify this, we used gradient-purified virion RNA as template for RT–PCR of this region and confirmed the presence of the insert in the genome. How this insert arose and was maintained in this lineage is not certain. Our data supports the idea that it was created by stuttering of the viral polymerase on the sequence 5′ AAUUUUU 3′ near the U track in the G–L junction of the positive-sense RNA template during virus replication, as first proposed by Bilsel & Nichol (1990). The polymerase stuttering event might have been favoured by the presence of the polyadenylation signal 3′ AUAC 5′, which occurs prior to the 5′ AAUUUUU 3′ only in the northern Central American lineage. Once this insert arose in an ancestral virus, it could have become fixed due to the lack of a recombination mechanism in Mononegavirales (Pringle, 1981).

Since the length of the 5′ non-coding sequence in the G gene was variable among the viruses studied, we wanted to determine its diversity within the quasispecies of a single virus strain. In order to test this, we cloned the resulting RT–PCR product of the G–L region of 94GUB, from the stop codon of G to the start codon of L, into the TOPO vector and selected ten individual colonies for sequencing. Four of the colonies had a 1 nucleotide deletion at position 14, four had a 1 nucleotide deletion at position 55 and one colony had a 30 nucleotide deletion at position 142–171 (Fig. 4). Despite this variability, we found the same consensus 175 nucleotide sequence in at least three independent 94GUB RNA preparations when
RT–PCR products were directly sequenced. Interestingly, the G–L region is also the site where other rhabdoviruses such as snakehead rhabdovirus and rabies virus (Kurath et al., 1997; Johnson et al., 2000; Ravkov et al., 1995) have additional genes. It seems that the G–L junction of rhabdoviruses in general is very permissive to insertions and deletions and perhaps is the site of extinct genes in VSV.

L protein

The VSV polymerase (L) has several functions when associated with P and N including: ATPase, methylase, guanylyl transferase, capping enzyme and poly (A) polymerase among others (Feldhaus & Lesnaw, 1988). This protein was 2109 amino acids in length, with high conservation (97–98%) among the four viruses (Fig. 3). However, the amino acid substitutions observed were not randomly distributed, but rather were concentrated in several discrete regions, particularly at the N- and C-terminal ends.

The exact domains within L associated with each of its functions are not clear but conserved motifs have been described in at least six areas of conservation among the Mononegavirales and these have been proposed as putative active sites (Poch et al., 1990). A sliding window analysis of the L amino acid alignment showed that, for the most part, these areas were of high amino acid conservation among the four virus sequences (Fig. 5). The polymerase motif QGDNQ was found at position 712–716 in all viruses in an area of low divergence within domain III (Fig. 5). The putative template recognition sequences within domain II at amino acids 530–660 were also completely conserved. Since these L sequences are the first from field strains without passage in tissue culture, they could be useful in determining the functional domains of this protein.

The purpose of this work was to determine and analyse the full-length genomic sequences of natural isolates, representing a distinct genetic lineage of VSV-IN1, and to compare these sequences with the only full-length sequence of VSV-IN1 deposited in GenBank. The results showed that overall structure and organization were similar among all the viruses, except for one genetic lineage from northern Central America that had an insertion of variable length at the 5′ non-coding region of the G gene. Protein comparisons showed that N was the most conserved protein, making it a good target for viral detection in diagnostic tests. Based on this finding, we are currently developing detection assays using real-time RT–PCR. Non-coding sequences involved in regulation of transcription and gene expression were completely conserved, but intergenic nucleotides varied in sequence and length in the M–G and G–L gene junctions. These full-length sequences could serve as a baseline for studies on genomic function using reverse genetics.

References


Conserved domains in the New Jersey and Indiana L proteins.


Received 31 January 2002; Accepted 20 May 2002