Chapter 3

Determination of the whole genome consensus sequence of the prototype rotavirus DS-1 strain

3.0 Introduction

The human rotavirus type A DS-1 strain (RVA/Human-tc/USA/DS-1/1976/G2P[4]) is the prototype of rotavirus strains in the DS-1-like genogroup (subgroup I, short electropherotype) (Wyatt *et al*., 1982). The rotavirus DS-1 strain was selected for use in the attempt to recover rotavirus by reverse genetics because it is a well characterised laboratory strain (Flores *et al.*, 1982). The rotavirus DS-1 strain has a G2P[4] serotype which is among the G1 to G4 types that are most prevalent, worldwide, in combination with P[4], P[6] and P[8] (Gentsch *et al*., 2005). This indicates that the rotavirus DS-1 strain is important for vaccine development. The wild-type rotavirus DS-1 strain was isolated in 1976 in Washington D.C. (USA) from a gastroenteritis patient (Kalica *et al.*, 1981). Primary adaptation from stool, of the relatively slow-growing strain was performed in MA104 cells without infection of primary cells (Wyatt *et al*., 1983). The cell culture-adapted DS-1 strain now propagates very well.

The first whole genome sequence for the rotavirus DS-1 strain was reported by Heiman et al. (2008). However, the extreme 5'- and 3'-terminal sequences could not be determined directly as genome segment-specific terminal primers were used for RT-PCR amplification (Heiman et al., 2008). While recent rotavirus DS-1 nucleotide sequences were determined from amplicons (Matthijnssens *et al*., 2008a, Heiman *et al*., 2008, Zao *et al*., 1999), the sequences were determined using Sanger sequencing. The other nucleotide sequences were obtained from cloned cDNA (Gorziglia *et al.*, 1988). Cloning may introduce bias i.e., certain regions (usually AT-rich) are less likely to be sequenced (McMurray *et al*., 1998). A deeper sequencing coverage, which lacks in Sanger sequencing technology (Sanger *et al*., 1977), would be required to efficiently sequence these regions. Furthermore, mutations can be introduced if low-fidelity DNA polymerases which lack proofreading ability are used during PCR amplification (Vanhercke *et al.*, 2005, Cline *et al*., 1996, Malet *et al*.,

2003). Additional mutations can be introduced when clones are amplified in bacterial cells. For instance, a severely biased selection of defective hepatitis C virus cDNA clones which could not be expressed was reported (Forns *et al*., 1997). Therefore, in order to develop a selection-free and successful reverse genetics system where viable rotavirus can be recovered, it was envisaged that the consensus sequence of the entire rotavirus DS-1 genome was needed. To date, no consensus sequence of the rotavirus DS-1 genome has been determined. The consensus sequence represents the most viable and predominant sequence of the viral population (Domingo, 2006). Viral population variants may confer useful traits, in *trans* by complementation, or may interfere with the replication of other population variants (González-López *et al*., 2004). Therefore, for rotavirus reverse genetics, it was thought that using a consensus sequence would eliminate the potential of unfit viral RNA from interfering with rotavirus recovery.

The lack of a consensus sequence is attributed to the absence of next-generation sequencing technology in the past. Several next-generation sequencing platforms were recently developed and are now available on the market. These platforms use similar or different chemistries to determine the nucleotide sequence. For instance, 454® pyrosequencing (Roche) determines the nucleotide sequence in a sequencing by synthesis approach, while the newer Ion Torrent technology (Life Technologies™) uses a semi-conductor device for non-optical nucleotide sequencing (Margulies *et al*., 2005, Rothberg *et al*., 2011, Lin *et al*., 2008). Other next-generation sequencing technologies use di-base sequencing by ligation such as SOLID™ system or reversible terminator sequencing as in the Illumina platform (Metzker, 2010, Anderson and Schrijver, 2010, Lin *et al*., 2008). An in-depth technical comparison of next-generation sequencing platforms is available in Metzker, 2010. For this study, 454® pyrosequencing was selected for determining the consensus whole-genome sequence of the rotavirus DS-1 strain. The selection was based on local availability of the technology, the relatively shorter turn-around-time as well as the superior read length which enables efficient read assembly (Metzker, 2010).

Towards the objective of obtaining the rotavirus DS-1 consensus sequence, with exact 5'- and 3'-terminal sequences, it was decided to use a combination of sequence-independent genome amplification (Potgieter *et al*., 2009) and 454®

pyrosequencing (Margulies *et al*., 2005). The whole genome consensus sequence of the rotavirus DS-1 strain would subsequently be compared to all the rotavirus DS-1 genome segment sequences that are available in the GenBank sequence database.

3.1 Materials and Methods

3.1.1 Cells and virus

MA104 cells (kindly provided by Dr. A. C. Potgieter, Onderstepoort Veterinary Institute) were maintained in Dulbecco's minimal essential medium (DMEM) (Hyclone). The DMEM contained 1% Penicillin/streptomycin/amphotericin (PSA; Lonza), supplemented with 1% non-essential amino acids (NEAA; Gibco) and 5% heat-inactivated foetal bovine serum (FBS; Hyclone). The culture-adapted rotavirus DS-1 strain used in this study was kindly provided, at 5 x 10^6 focus forming units/ml, by Dr Carl Kirkwood (Murdoch Children's Research Institute, Melbourne, Australia) following 10 passages from initial adaptation in MA104 cells. The lyophilised DS-1 rotavirus was reconstituted in a total of 1 ml DMEM containing 1% PSA, non-essential amino acids and 10 µg/ml porcine trypsin IX (Sigma), but no FBS. Activation was performed in a water bath (Grant) at 37 ^oC for 30 minutes. MA104 cells in a 25 cm^2 flask were washed twice with PBS (Sigma) and the activated DS-1 virus was adsorbed onto the cells for 45 minutes, at room temperature, with rocking. Following adsorption, the virus was allowed to replicate in a total of 5 ml DMEM supplemented with 1 µg/ml trypsin at 37 ^oC in 5% CO_2. The 25 cm^2 flask was examined daily for cytopathic effect (CPE) for up to 5 days. The virus was passaged again in 25 cm^2 flasks, with preparation of stocks, at each passage, that were stored at -80 ^oC.

3.1.2 Extraction of dsRNA

A third passage for dsRNA extraction was performed, at a multiplicity of infection of 0.9, in two 75 cm^2 flasks and the virus harvested after 5 days of incubation when the CPE had advanced to more than 75%. Any cells at the bottom of a rotavirus DS-1-infected flask were scraped off and the total contents of a 75 cm^2 flask were centrifuged for 10 minutes at 110 x *g* at room temperature. The pellet was used for dsRNA extraction and the supernatant stored at 4 ^oC. The pellet was resuspended in

DMEM to a final volume of 200 µl followed by the addition of 100 µl of an extraction buffer (20 mM Tris, 10 mM $CaCl_2$, 0.8% NaCl, pH 7.4). Concentrated Vertrel XF (100 µl; Du Pont) was added to enhance homogenisation of the sample. Five volumes of TRIzol® reagent (Invitrogen) were added, mixed and allowed to stand for 2 minutes followed by three volumes of chloroform, mixed and allowed to stand for 5 minutes. Centrifugation followed at 4 ^{o}C for 15 minutes at 15 000 x *g*. The aqueous phase was carefully removed and added to a microcentrifuge tube. An equal volume of isopropanol was added, mixed by inversion and allowed to stand for 10 minutes at room temperature to precipitate the dsRNA. Centrifugation at 15 000 x *g* for 15 minutes followed and the dsRNA pellet was dried by centrifugation again at room temperature for a further 15 minutes at 15 000 x *g*. Extracted dsRNA was dissolved in 95 µl elution buffer (EB; Qiagen) for 10 minutes at room temperature with intermittent mixing. The extraction was analysed on a 1% agarose gel (Hispanagar) that contained 1 µg/ml ethidium bromide and visualised using a ChemiGenius image analyser (Syngene) and Gene Snap software (Syngene). Single-stranded RNA was removed by precipitation in 2 M LiCl at 4^{o}C for 16 hours, followed by centrifugation at 15 000 x *g* rpm at 4 ^{o}C for 30 minutes. This was followed by purification of dsRNA from the supernatant using a MinElute gel extraction kit (Qiagen) according to the manufacturer's instructions. Elution of dsRNA was performed by adding 32 µl elution buffer (EB: 10mM Tris/HCl, pH 8.5) to give a final volume of 30 µl of purified dsRNA. The quality of dsRNA was analysed by 1% agarose (Hispanagar) gel electrophoresis and visualised with a ChemiGenius image analyser (Syngene) and Gene Snap software (Syngene).

3.1.3 Oligo-ligation

To facilitate sequence-independent genome amplification, oligo-ligation was performed. The oligo-ligation reaction was performed as described by Potgieter *et al*. (2009). The PC3-T7 loop oligonucleotide (200 ng/µl) was ligated to dsRNA in a volume of 12 µl. The sequence of the loop oligonucleotide is 5'-p–GGATCCCGGGAATTCGGTAATACGACTCACTATATTTTTATAGTGAGTCGTATTA–OH-3' (TIB MOLBIOL; Potgieter *et al*., 2009). The ligation reaction contained 50 mM HEPES/NaOH (pH 8.8; Sigma), 18 mM $MgCl_2$, (Sigma), 0.01% bovine serum albumin (BSA; TaKaRa), 1 mM ATP (Roche), 3 mM DTT (Roche), 10% DMSO

(Sigma), 20% PEG_{6000} (Calbiochem), and 30 U T4 RNA ligase (TaKaRa). The ligation reaction was incubated at 37 ^{o}C for 16 hours, followed by purification using the MinElute gel extraction kit following the manufacturer's instructions. Elution of ligated dsRNA was performed with 10 µl of EB (Qiagen) and centrifugation at 15 000 x *g* for 1 minute.

3.1.4 Sequence-independent cDNA synthesis

Purified and ligated dsRNA (8 µl) was denatured by adding 300 mM methyl mercury hydroxide (MMOH; Alfa Aesar), to a final concentration of 30 mM, with incubation at room temperature for 30 minutes in a fume hood. The cDNA synthesis reaction contained 1 mM dNTPs (TaKaRa), 1X Transcriptor High Fidelity buffer (Roche), 30 mM 2-mercaptoethanol (Sigma), 0.5 U RNase inhibitor (Roche) and 10 U Transcriptor High Fidelity Reverse Transcriptase (Roche). The reaction was incubated at 42 ^{o}C for 45 minutes, followed by 55 ^{o}C for 15 minutes. Removal of excess dsRNA and cDNA annealing was performed by adding 1 M NaOH (to a final concentration of 0.1 M) and incubation at 65 ^{o}C for 30 min., followed by the addition of 1M HCl (final concentration of 0.1 M) and 1 M Tris/HCl (pH 8.3) to a final concentration of 0.1 M and incubation at 65 ^{o}C for 1 hour.

3.1.5 PCR amplification of the DS-1 cDNA genome

Amplification of genomic cDNA using the polymerase chain reaction (PCR) was performed with a PC2 primer (5'-p–CCGAATTCCCGGGATCC-3'; TIB MOLBIOL) described by Potgieter *et al*. (2009). The PC2 primer is complementary to the PC3-T7 loop. The reaction mix contained 1X Phusion^{TM} High Fidelity buffer (Finnzymes), 2.5 mM dNTPs (TaKaRa) and 1 U Phusion^{TM} High Fidelity DNA polymerase (Finnzymes). The cycling conditions were the same as described by Potgieter *et al*., 2009, i.e. initial 72 ^{o}C for 1 minute to fill in cDNA ends to ensure amplification of intact DNA, denaturation at 94 ^{o}C for 2 minutes followed by 25 cycles of denaturation at 94 ^{o}C for 30 seconds, annealing at 67 ^{o}C for 30 seconds and extension at 72 ^{o}C for 4 minutes. The PCR reaction was replicated 8 times to ensure sufficient cDNA for 454® pyrosequencing (at least 5 µg was required). PCR amplicons were analysed using 1% agarose gel (Hispanagar) electrophoresis in 1X TAE (0.04 M Tris-acetate;

0.001 M EDTA; pH 8.0). The agarose gel contained ethidium bromide and was analysed using a ChemiGenius image analyser (Syngene) and Gene Snap software (Syngene). PCR amplicons were purified by pooling the amplicons from the 8 PCR replicates followed by mixing with 5 volumes of binding buffer (PB buffer; Qiagen). The mix was transferred to a QIAquick® spin column (Qiagen) and centrifuged at 15 700 x *g* for 1 minute to bind cDNA. A wash step was performed using 750 µl of wash buffer (PE buffer; Qiagen) and centrifugation at 15 700 x *g* for 1 minute. An additional centrifugation was performed to remove residual PE buffer. Elution of purified cDNA was performed by adding 39 µl EB buffer (QIAquick® gel extraction kit, Qiagen). The concentration of cDNA was measured using a NanoDrop® 1000 spectrophotomer (Thermo Scientific).

3.1.6 Whole genome 454® pyrosequencing and analyses

Whole genome sequencing was performed at Inqaba Biotec™ (Pretoria, South Africa) using GS FLX Titanium (Roche) pyrosequencing technology. A 33 µl volume containing 13.9 µg of PCR amplified, purified, cDNA was submitted for sequencing. Sequence data was received in a standard flowgram format file (sff) and assembly was performed using the SeqMan Pro module of Lasergene™ v8.1 software (DNASTAR®). The consensus sequences were exported to the MEGAlign module for manual inspection and editing if required. Similar rotavirus DS-1 sequences in GenBank were identified using the Basic Local Alignment Search Tool (BLAST). Sequences retrieved from GenBank for comparison with the consensus sequence obtained by pyrosequencing are listed in Table 3.1. Sequence alignment was performed using the Sequence Viewer v6.4 (CLC Bio).

3.1.7 Modelling of protein and RNA structures

Rotavirus protein structures similar to the rotavirus DS-1 proteins were identified in the protein data bank (PDB) using the Phyre protein fold recognition sever at http://www.sbg.bio.ic.ac.uk/~phyre/ (Kelley and Sternberg, 2009). Modelling of protein structures and localisation of specific amino acid residues was achieved using UCSF Chimera (Pettersen *et al*., 2004). RNA folding was performed by minimum free energy prediction of secondary structures with RNA-fold at

http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi (Zuker and Stiegler, 1981, Schuster *et al.*, 1994).

Table 3.1. Rotavirus DS-1 nucleotide Sequences retrieved from GenBank for comparison with the consensus DS-1 nucleotide sequence obtained by pyrosequencing

GS1 (VP1)	GS2 (VP2)	GS3 (VP3)	GS4 (VP4)	GS5 (NSP1)	GS6 (VP6)	GS7 (NSP3)	GS8 (NSP2)	GS9 (VP7)	GS10 (NSP4)	GS11 (NSP5/6)
DQ870505	DQ870506	AY277914	EF672577	L18945	DQ870507	EF136660	EF672580	AB118025	AF174305	EF672583
AF044360	DQ490202	EF583027	DQ141310	EF672578	EF583027	EF672579	L04529	L20813	EF672582	M33608
AF106302	EF583026	-	AJ540227	*GQ414546	EF619345	-	-	EF672581	-	-
EF583025	-	-	AB118025	*DQ146688	-	-	-	-	-	-
DQ870505	-	-	-	*EF554088	-	-	-	-	-	-

GS: genome segment
*: DS-1-like strains

3.2 Results

3.2.1 Sequence-independent genome amplification and determination of the whole genome consensus sequence

To determine the consensus sequence of the whole genome of the rotavirus DS-1 virus obtained from the MCRI, rotavirus DS-1 dsRNA was extracted (Figure 3.1A), reverse-transcribed into cDNA and amplified using sequence independent whole-genome amplification (Figure 3.1B). Approximately 14 µg of cDNA was submitted for 454® pyrosequencing at Inqaba Biotec™. To achieve this quantity of cDNA, eight separate amplifications were performed followed by pooling and purification of the cDNA.

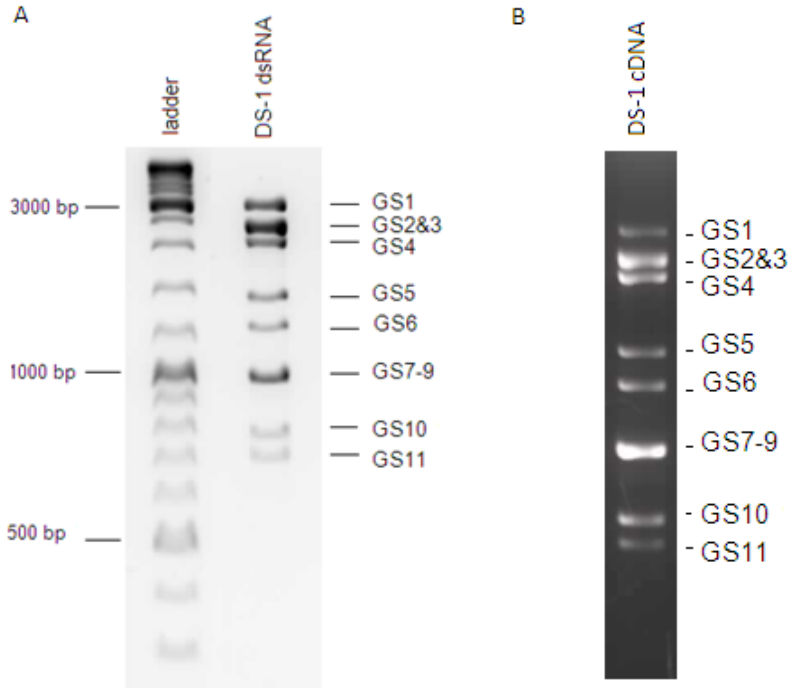

Figure 3.1. Gel electrophoresis of the rotavirus DS-1 dsRNA and cDNA thereof. Genome segment is abbreviated GS. **A**, 1% agarose gel electrophoresis of rotavirus DS-1 dsRNA. The gel contained 0.1 µg/ml ethidium bromide and electrophoresis was was performed in a 1X Tris/borate/EDTA (TBE) buffer. The molecular size marker is a GeneRuler™ (#SM1173) DNA ladder (Fermentas). **B**, 1% agarose gel electrophoresis of sequence-independent amplified rotavirus DS-1 cDNA. Electrophoresis was performed in 1X Tris/acetate/EDTA (TAE) buffer.

Pyrosequence data comprised of 10180 nucleotide sequence reads of approximately 400 bp each. The Lasergene™ SeqMan Pro (DNASTAR®) was used for sequence assembly. A total of 36 contigs were obtained and used to determine the complete consensus sequence for each genome segment as illustrated in Figure 3.2.

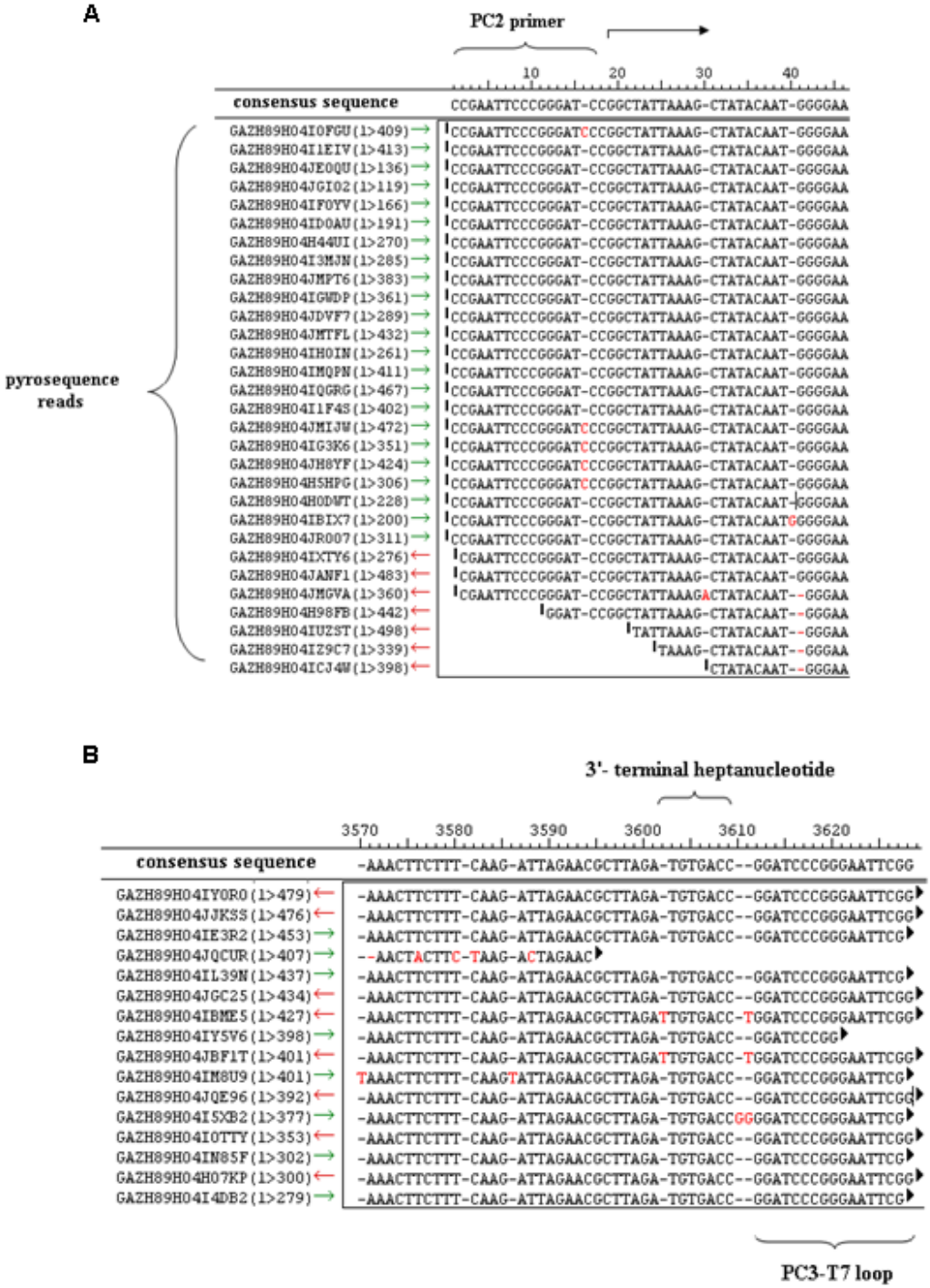

Figure 3.2. A representation of a contig aligment used in the determination of the consensus sequence. The red and green arrows indictae the direction of sequencing of the pyrosequence reads. Identification of nucleotide sequences of the PC2 primer and 5'-terminal end sequences **(A)** and PC3-T7-loop and 3'-terminal end sequences **(B)** marking the begining and end of a complete consensus sequence of each genome segment. The start of the conserved 5'-terminal nucleotide sequences is indicated by the arrow in **A**.

The total size of the consensus genome obtained was 18 612 bp. The whole-genome consensus sequence was submitted to GenBank and the following accession numbers were assigned: HQ650116–HQ650126 (Table 3.2). The average depth of sequencing coverage was 204-fold. Genome segment 1 had the lowest depth of coverage while genome segment 8 the highest depth of coverage. The range of coverage for all 11 genome segments was 22–458-fold. The depth of coverage for the other genome segments is indicated in Table 3.2.

Table 3.2. Summary of the rotavirus DS-1 whole-genome consensus sequence data, obtained with 454® pyrosequencing, in comparison to DS-1 sequences in GenBank.

Genome segment	GenBank Accession number	Size (bp)	Depth of coverage (-fold)	Amino acids in coding region	Nucleotide differences			Amino acid differences	
					Insertions	Deletions	Point variations	Insertions	Point variations
1 (VP1)	HQ650116	3302	22	1088	None	None	None	None	None
2 (VP2)	HQ650117	2684	89	879	None	None	3	None	None
3 (VP3)	HQ650118	2591	49	835	None	None	11	None	None
4 (VP4)	HQ650119	2359	155	775	None	None	28	None	[a] 52S→H, 53W→G, 106S→I, 107S →A, 142T→M, 144T→K, 150N→D, 230R→S, 245K→R, 280I→V, 380T→I, and 397N/D→I
5 (NSP1)	HQ650120	1563	80	493	1	[b]2	5	1-7 MKSLVEA	[c]137 S→L, 338 I→T, 381E→D, 488R→G
6 (VP6)	HQ650121	1356	263	397	None	None	1	None	None
7 (NSP3)	HQ650122	1064	291	313	None	None	1	[d]1-3 MLK	[e]51D→G
8 (NSP2)	HQ650123	1059	458	317	None	None	1	None	194G→D
9 (VP7)	HQ650124	1062	384	326	None	None	5	None	None
10 (NSP4)	HQ650125	751	195	175	None	None	None	None	None
11 (NSP5/6)	HQ650126	821	262	200/92	6	None	6	139-140QE	[f]9G→S and 49N→R

[a] Differences in the translated VP4 amino acid sequence: residues 52-150 occur in VP8* region; residue 150 is a glycosylation site; residue 230 is a trypsin cleavage site; residue 380 occurs in an antigenic region and residue 397 in a hydrophobic region

[b] Deletion of T(U)A at nucleotide 1553–1554 in the 3'-UTR

[c] Residue 137 occurs in a region that is important for cytoskeletal localisation while the other changes occur in a region that interacts with interferon regulatory factor 3

[d] A GenBank sequence EF672579 described as a complete ORF sequence lacks the first three residues MLK

[e] Amino acid difference in translated NSP3 occurs in an RNA-binding region

[f] The 9G→S difference in the translated NSP5 amino acid sequence occurs in a region that interacts with VP1

3.2.2 General analyses of the rotavirus DS-1 genome segment consensus sequences

The nucleotide sequence 5'-GGC(U/A)$_7$- was obtained in the 5'-untranslated regions (UTRs) of all genome segments. The 3'-tetranucleotide terminal sequence -GACC-3' was obtained for all genome segments. The 3'-terminal heptanucleotide sequence, -U**G**UGACC-3', was present in 9 genome segments. In genome segments 2 and 10, the heptanucleotide sequence obtained was -U**A**UGACC-3'. Using the RotaC classification tool (Maes *et al*., 2009), the full genotype of the rotavirus DS-1 strain was confirmed as G2-P[4]-I2-R2-C2-M2-A2-N2-T2-E2-H2.

The comparison of the rotavirus DS-1 whole-genome consensus sequence to all rotavirus DS-1 genome sequences available in GenBank revealed nucleotide and amino acid differences (Table 3.2). Nucleotide sequence differences were observed in 10 of the 11 genome segments. Only the consensus sequences of genome segment 1 (HQ650116; VP1) and genome segment 10 (NSP4; HQ650125) were 100% identical to the genome segment 1 sequences available in GenBank. While nucleotide differences were observed in genome segments 2 (VP2), 3 (VP3), 6 (VP6), and 9 (VP7), there were no associated amino acid changes (Table 3.2). Amino acid differences were observed in the deduced amino acid sequences of VP4, NSP1, NSP2, NSP3 and NSP5 (Table 3.2).

3.2.3 Analyses of DS-1 genome segments encoding structural proteins VP1–VP4, VP6 and VP7

The genome segment 1 (HQ650116; VP1) consensus nucleotide sequence was identical to the four rotavirus DS-1 genome segment 1 sequences in GenBank i.e., DQ870505 (full length genome segment), EF583025 (full length open reading frame), AF106302 (427 bp corresponding to nucleotides 44–470 of the consensus sequence) and AF044360 (561 bp corresponding to nucleotides 38–553 of the consensus sequence).

The consensus nucleotide sequence for genome segment 2 (VP2; HQ650117) was compared to three nucleotide sequences in GenBank. DQ870506 is a full length genome segment sequence, EF583026 a full length open reading frame (ORF) and

DQ490202 a partial sequence of 508 bp corresponding to a region of the consensus sequence spanning nucleotides 457–964. Three nucleotide differences were observed between the consensus sequence and the GenBank sequences. At nucleotide position 577 the consensus sequence contained T(U) while the GenBank sequences contain C at this position. The consensus sequence as well as EF583026 indicated that C was present at nucleotide position 727, while DQ870506 and DQ490202 contained T(U) at this position. At nucleotide position 901 the consensus sequence displayed T(U), while C is present in the GenBank sequences. However, no amino acid differences were observed between the deduced amino acid sequences of VP2.

For genome segment 3 (VP3; HQ650118), the consensus nucleotide sequence was 99% identical to a full length nucleotide sequence, AY277914, and a full length ORF, EF583027. The nucleotide differences observed were 13G→A, 17C→T(U), 18T(U)→C, 559C→T(U) (C in AY277914), 1238C→T(U) (C in AY277914), 2401A→G, 2536A→G, (A in AY277914), 2577A→G and 2582T(U)→(C). Despite these nucleotide differences, there were no amino acid changes.

The consensus nucleotide sequence for genome segment 4 (VP4; HQ650119) was compared to four GenBank sequences i.e., AB118025, AJ540227, EF672577 and DQ141310. A 2% variation resulted from a total of 28 nucleotide differences (Table 3.3) between all the genome segment 4 sequences except DQ141310. The GenBank sequence DQ141310 is a fragment of 492 bp and it was 100% identical the corresponding region of the consensus sequence nucleotide sequence (nucleotide position 186–678).

Table 3.3. Nucleotide differences observed between the DS-1 genome segment 4 consensus sequence and the DS-1 genome segment 4 sequences in GenBank.

Nucleotide position	Accession number of DS-1 genome segment 4 (VP4)				
	HQ650119[§]	EF672577	AB118025	AJ540227	DQ141310
18	G	G	A	A	-
162	A*	A	A	U	-
163	C	C	C	U	-
164	A*	A	A	C	-
192	G	G	G	U	G
318	A	A	A	U	A
324	U	U	U	A	U
325	A	A	A	U	A
326	U*	U	U	C	U
327	C	C	C	A	C
328	G*	G	G	U	G
378	U	U	U	G	U
393	U	U	U	C	U
434	U*	U	U	C	U
440	A*	A	A	C	A
457	G*	G	A	A	G
699	U*	U	A	A	-
714	A	A	A	G	-
717	U	U	U	A	-
732	U	U	U	G	-
743	G*	G	G	A	-
847	G*	G	G	A	-
861	A	A	A	U	-
954	C	A	C	C	-
999	U[†]	C	C	C	-
1148	U*	U	C	C	-
1198	A	A	A	G	-
1199	U*[†]	A	A	A	-

[§]Consensus sequence of genome segment 4 (VP4)
* Associated with amino acid changes
[†] Nucleotide observed only in the consensus sequence and not in any GenBank sequences
- No sequence data available

Twelve of the 28 nucleotide differences resulted in amino acid changes (Table 3.2; Table 3.3). Changes observed in the VP8* region of VP4 were 52S→H, 53W→G, 106S→I, 107S→A, 142T→M, 144T→K, 150N→D and 230R→S. Close scrutiny of the VP8* structure, using a DS-1 model PDB ID: 2AEN (Monnier *et al*., 2006), indicated that I106 and A107 are located internally in the VP8* structure (Figure 3.3A) while M142, K144 and D150 are located on the surface of the VP8* structure (Figure 3.3B). A 245K→R change was observed in the region between VP8* and VP5*. Amino acid differences 280I→V and 380T→I were observed in the antigenic domain of VP5*. Furthermore, a novel isoleucine at position 397, in a hydrophobic region, was observed. Based on a rhesus rotavirus model PBD ID: 1SLQ (Dormitzer et al., 2004), the residues V280, I380 and I397 were located on the surface of the predicted VP5* structure (Figure 3.3C). Pyrosequence data showed that at the corresponding codon position encoding amino acid residue 397, 30% (approximately 150 reads) of sequence reads indicated an AAT(U) codon encoding an asparagine residue (Figure 3.4).

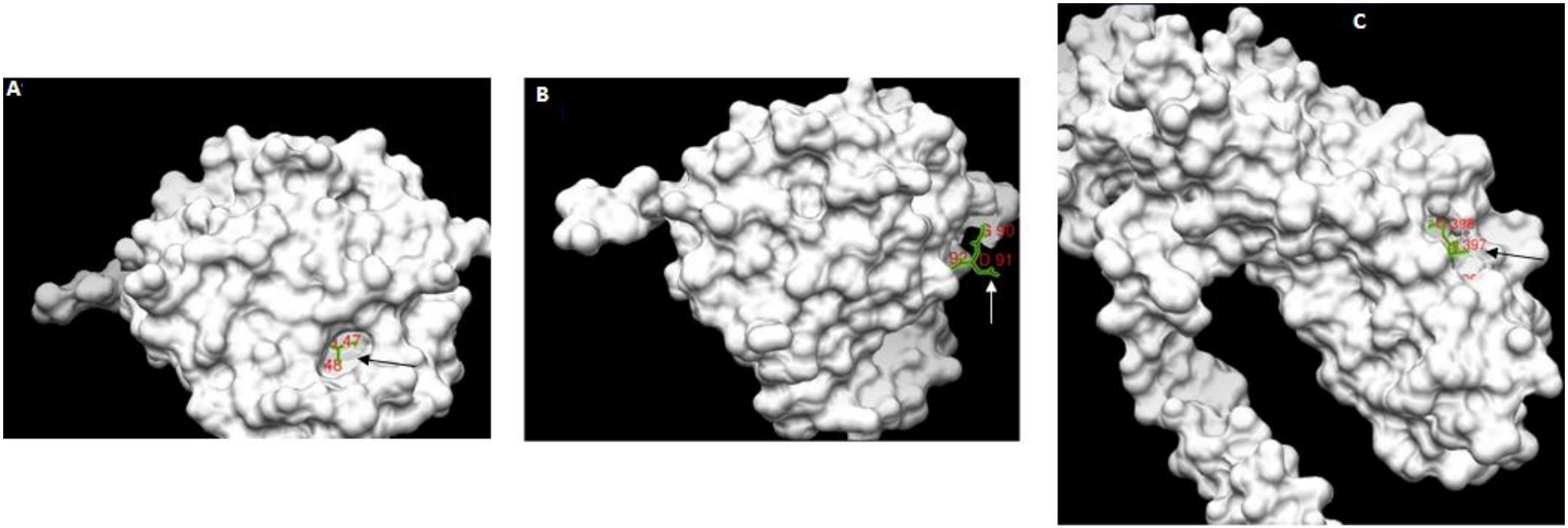

Figure 3.3. Models of rotavirus VP4 protein structures predicted using Chimera UCSF software. The models depict the determination of the location of specific amino acids. **A**, Model of VP8* structure based on a DS-1 structure PDB ID: 2AEN (Monnier *et al.*, 2006) showing the internal location of amino acid residues I106 and A107 (residues 47 and 48 in the model). **B**, VP8* structural model (based on PDB ID: 2AEN) indicating the surface location of D91 (D150 equivalent in the deduced consensus VP4 amino acid sequence). **C**, VP5* model based on PBD ID: 1SLQ (Dormitzer *et al*., 2004) showing the surface location of the novel I397.

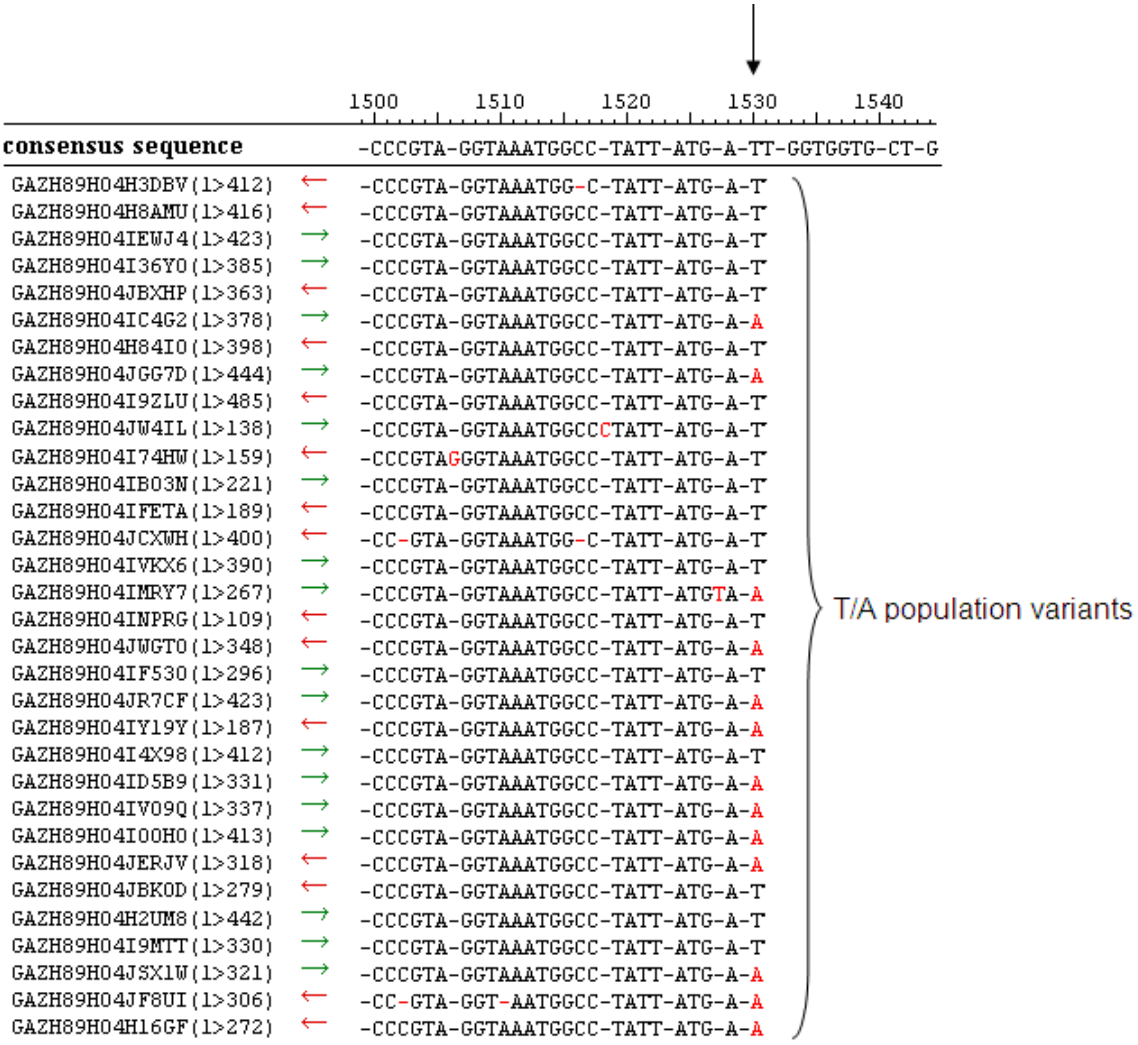

Figure 3.4. A representation of sequence reads indicating nucleotide population variants in genome segment 4 (VP4). The nucleotide position 1530 (nucleotide position 1191 in the edited complete consensus sequence), indicated by the vertical arrow, shows that a T(U) was called as the consensus base and the resulting consensus codon was ATT (encoding a novel isoleucine at residue 397). The nucleotide T(U) was called as a consensus due to the higher number of pyrosequence reads containing the nucleotide when compared to pyrosequence reads containing the nucleotide A. Where an A was present at the same position in the other reads, the resulting codon was AAT (encoding asparagine at residue 397).

For genome segment 6 (VP6), the consensus sequence (HQ650121) was 100% identical to DQ670507 (a full length genome segment) and EF583028 (ORF sequence). The GenBank sequence EF619345 is a 328 bp fragment and it was 98% identical to the genome segment 6 consensus nucleotide sequence. The nucleotide differences observed were 749C→T(U), 751G→A, 1115C→T(U) and 1127A→T(U). However, no amino acid differences were observed.

No nucleotide differences were observed between the genome segment 9 (VP7) consensus sequence HQ650122 and GenBank sequences L20813 and EF672581. However, a GenBank sequence AB118023 was 99% identical to the VP7-encoding consensus sequence HQ650122. Five nucleotide differences observed were 660C→G, 1011A→T(U), 1039A→T(U), 1041A→G and 1043T(U)→(C). The deduced amino acid sequences were, however, 100% identical.

3.2.4 Analysis of genome segments encoding the DS-1 non-structural proteins NSP1–NSP6

The consensus DS-1 genome segment 5 obtained in this study (HQ650120) was 1563 bp in length. A full length genome segment 5 sequence L18945 containing 1564 bp and a partial 1461 bp sequence (EF672578) are available in GenBank for the DS-1 strain. L18945 was 99% identical to the consensus sequence (HQ650120) while EF672578 was identical to the corresponding region of the consensus sequence spanning nucleotides 32–1492. At the 5'-terminal end, the consensus genome segment 5 (HQ650120) contained an A insertion at nucleotide position 11 resulting in a start codon at nucleotide positions 11–13 (Table 3.1). The ORF of L18945 and EF672578 starts at nucleotide position 32. Therefore the ORF of the consensus sequence was 21 nucleotides longer at the 5'-end, than the ORFs of GenBank sequences L18945 and EF672578. The NSP1 amino acid sequence deduced from the consensus nucleotide sequence contained 493 amino acid residues. The rotavirus DS-1 consensus NSP1 amino acid sequence therefore contained 7 additional amino acids (MKSLVEA) at the N-terminal end when compared to L18945 and EF672578 (Figure 3.6). Alignment of the rotavirus DS-1 consensus sequence to rotavirus DS-1-like strains such as GER1H-09, N26-02 and B1711 showed that their first 7 N-terminal residues were 100% identical (Figure 3.5). However, published sequences for other DS-1-like strains such as the Tb-Chen strain (G2P[4]) do not have the MKSLVEA sequence at the N-terminal end of NSP1. In the 3'-UTR end, the consensus genome segment 5 (HQ650120) contained deletions of the nucleotides T(U)A at positions 1553 and 1554 (Table 3.1). Therefore, GenBank sequence L18945 was a single nucleotide longer than the consensus sequence obtained in this study (HQ650120).

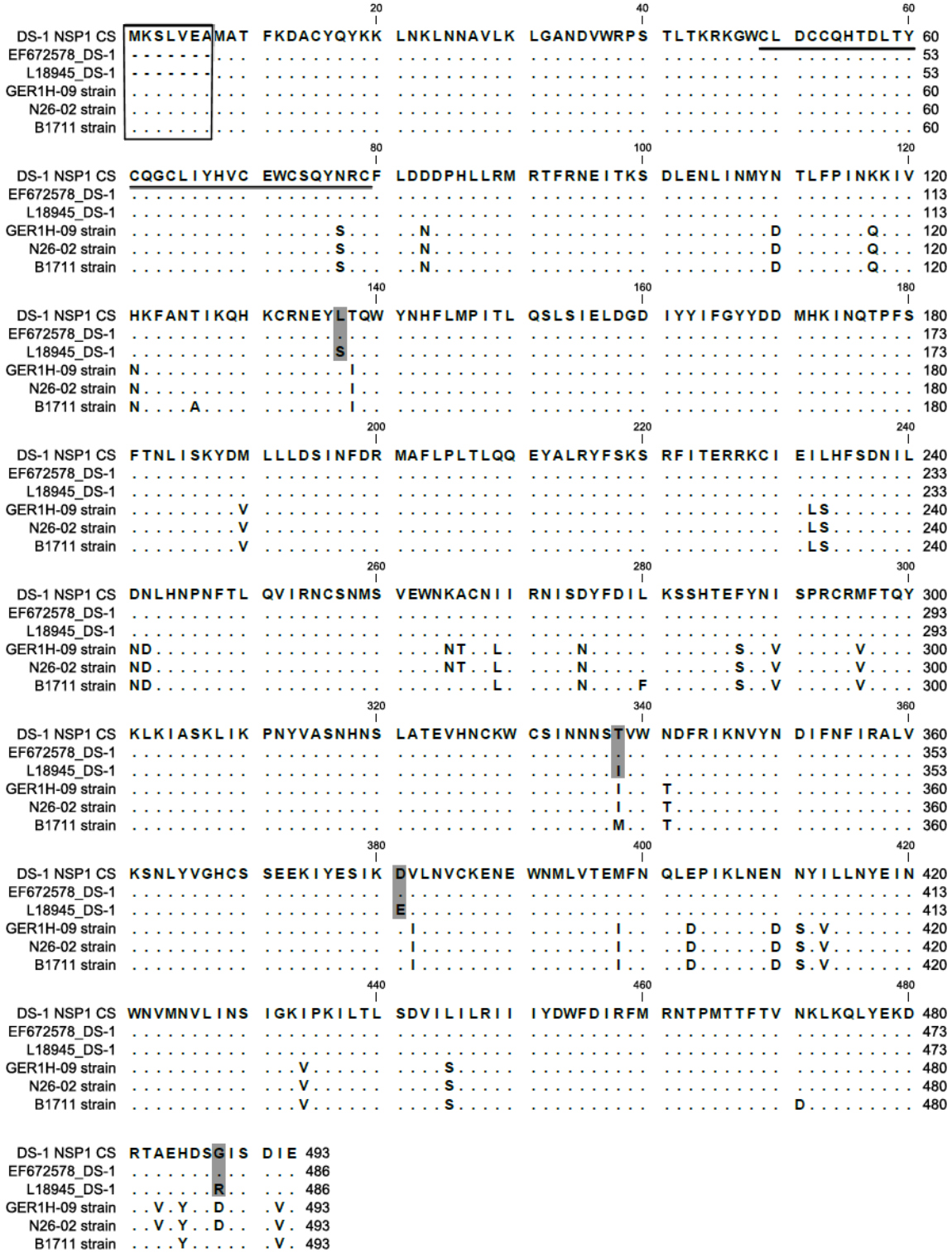

Figure 3.5. Alignment of the deduced DS-1 NSP1 consensus sequence amino acid (DS-1 NSP1 CS; HQ650120) with rotavirus DS-1 sequences (EF672578 and L18945) and selected rotavirus DS-1-like sequences (GER1H-09, N26-02 and B1711) in GenBank. The first 7 residues, MKSLVEA, observed at the N-terminal end of the translated consensus sequence and not in GenBank rotavirus DS-1 sequences but present in the DS-1-like strains B1711, G9P[6]; GER1H-09, G8P[6]; and N26-02, G12P[6] are boxed and differences between the deduced consensus sequence and the rotavirus DS-1 sequences are shaded. The conserved rotavirus DS-1 cysteine-rich zinc finger motif is underlined.

The genome segment 7 (NSP3) consensus nucleotide sequence (HQ650122) was 1064 bp and the ORF spanned nucleotides 26–967. The ORF translated into a deduced NSP3 which contained 313 amino acid residues. A full length genome segment sequence (EF136660) as well as a 933 bp sequence (EF672579) encoding DS-1 NSP3 are available from GenBank. EF672579 corresponds to a region from nucleotide 35–967 of the consensus sequence. EF672579 is nine amino acids shorter than the ORF of the consensus sequence and therefore missing MLK, the first three residues. The GenBank sequence EF136660 contains the nucleotide A at position 177 while the consensus sequence and EF672579 contain the nucleotide G at the same position. This nucleotide difference results in a 51D→G change in the RNA-binding region. Based on the rotavirus SA11 NSP3 structure PDB ID: 1KNZ (Deo *et al*., 2002), the amino acid position 51 is located on the surface of the predicted NSP3 structure (Figure 3.6A).

The consensus genome segment 8 (HQ650123) was 1059 bp and the ORF spanned nucleotides 47–1000. The deduced amino acid sequence contained 317 amino acid residues. Two DS-1 NSP2-encoding sequences are available in GenBank i.e., a full length ORF (EF672580) and a full length genome segment (L04529). EF672580 was identical to the consensus sequence. A single nucleotide difference was observed at nucleotide position 627 between the consensus sequence which contained the nucleotide A, and L04529 which contained the nucleotide G. The nucleotide difference resulted in the amino acid difference 194G→D. Using a rotavirus SA11 structure PDB ID: 1L9V (Jayaram et al., 2002), the amino acid position 51 was mapped to the surface of NSP2 (Figure 3.6B).

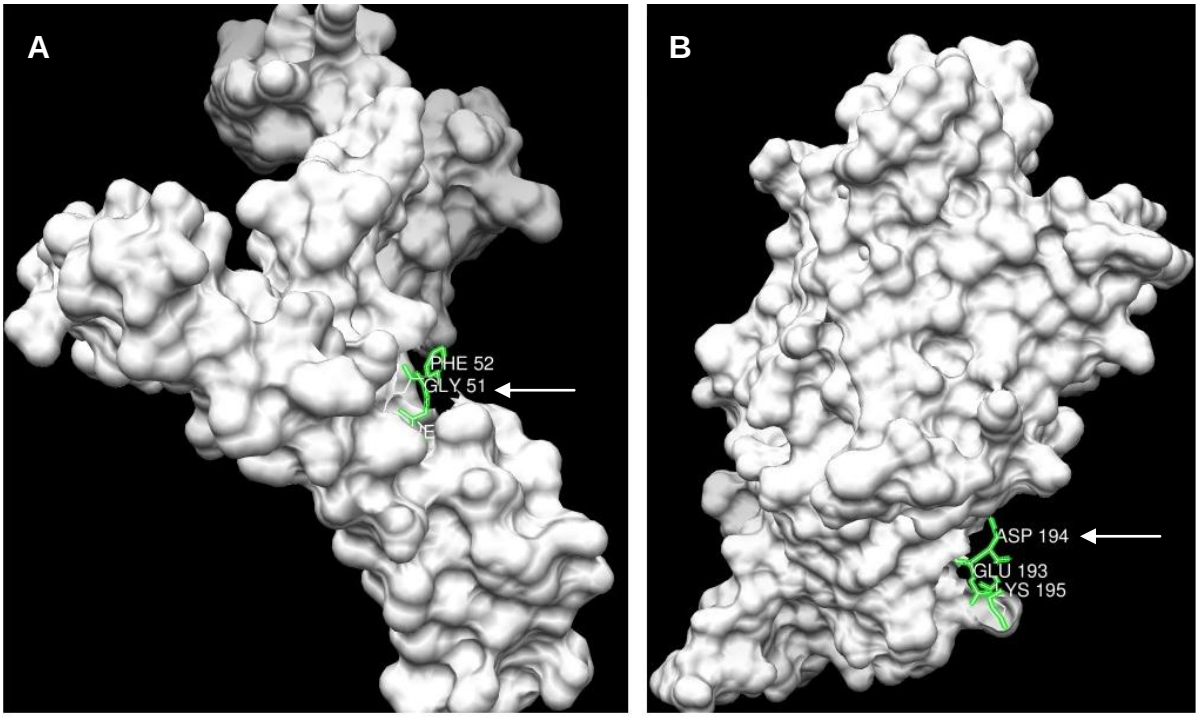

Figure 3.6. Models of rotavirus NSP3 (genome segment 7; A) and NSP2 (genome segment 8; B) structures generated using Chimera UCSF software. A, Depiction of the surface location of amino acid residue 51 (indicated by arrow) in which a 51D→G variation was observed in NSP3. **B**, A structure of NSP2 showing the surface location of amino acid 194 (indicated by arrow). A 194G→D change was observed at this position.

Genome segment 10 (NSP4) sequences available in GenBank are 528 bp (EF672582) and 687 bp (AF174305) long. The consensus genome segment 10 sequence determined with 454® pyrosequencing (HQ650125) was 751 bp and the ORF spanned nucleotide position 42–569. AF174304 was 100% identical, and corresponds, to the ORF of the consensus sequence (nucleotide position 35–721). EF672582 was 100% identical to a region in the consensus sequence spanning nucleotides 42–569. The 34 5'-terminal and 30 3'-terminal nucleotide sequences of genome segment 10 (NSP4) have not been reported previously. These sequences were determined in this study as 5'-GGCTTTTAAAAGTTCTGTTCCGAGAGAGCGCGTG-3' at the 5'-terminus, and 5'-GTTAATGGAAGGAACGGTCTTAATATGACC-3' at the 3'-terminus (Figure 3.7).

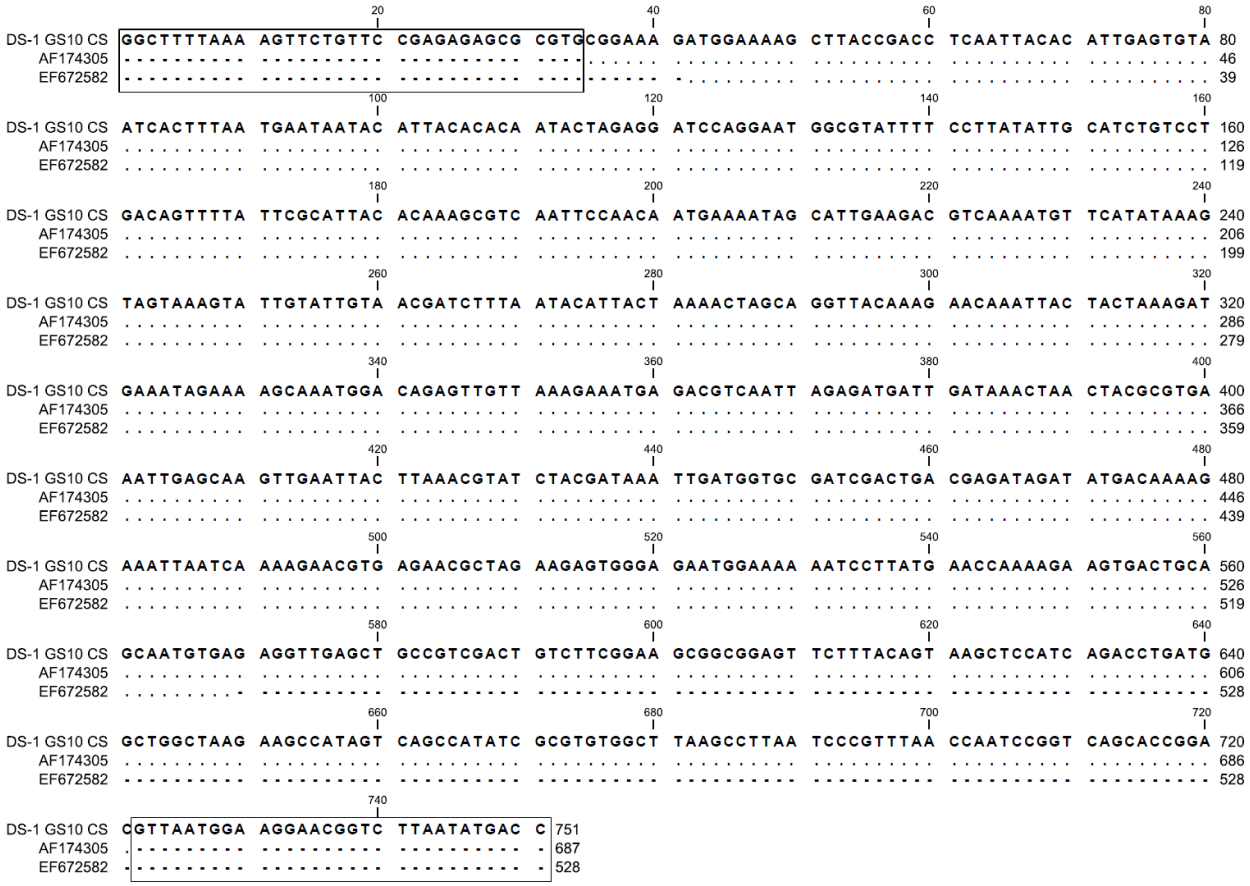

Figure 3.7. Alignment of the consenus rotavirus DS-1 genome segment 10 (NSP4; HQ650125) nucleotide sequence with genome segment 10 sequences from GenBank. The alignment indicates the newly determined terminal sequences (boxed).

The consensus sequence for genome segment 11 (NSP5/6) obtained in this study, HQ650126) was 821 bp (Figure 3.8). The consensus sequence (HQ650126) was compared to a full length rotavirus DS-1 genome segment 11 sequence (GenBank ID: M33608) of 667 bp. A 148 bp nucleotide sequence (nucleotides 625–772) was not present in the 3'-UTR of M33608. A partial GenBank sequence EF672583 lacks 21 nucleotides at the 5'-terminus and 60 nucleotides at the 3'-terminus (Figure 3.8). This partial sequence was, however, 100% identical to the corresponding region of the consensus genome segment 11 sequence (Figure 3.8). No internal duplications were observed in the consensus genome segment 11 sequence. Six potentially *cis*-acting 5'-CUUC- motifs (Li et al., 2010) were present in M33608 and the consensus sequence at nucleotides 49–52, 53–56, 62–65, 89–92, 206–209, 392–395 (Figure 3.10). A seventh 5'-CUUC- motif is present at nucleotide 568–571 in M33608 and at nucleotide position 574–577 in the consensus sequence. A potentially *cis*-acting 5'-

GGGAGCUCCC- palindrome (Li *et al*., 2010) spanned nucleotide position 794–803 in the consensus genome segment 11 sequence and is present at nucleotide position 640-649 in M33608 (Figure 3.9).

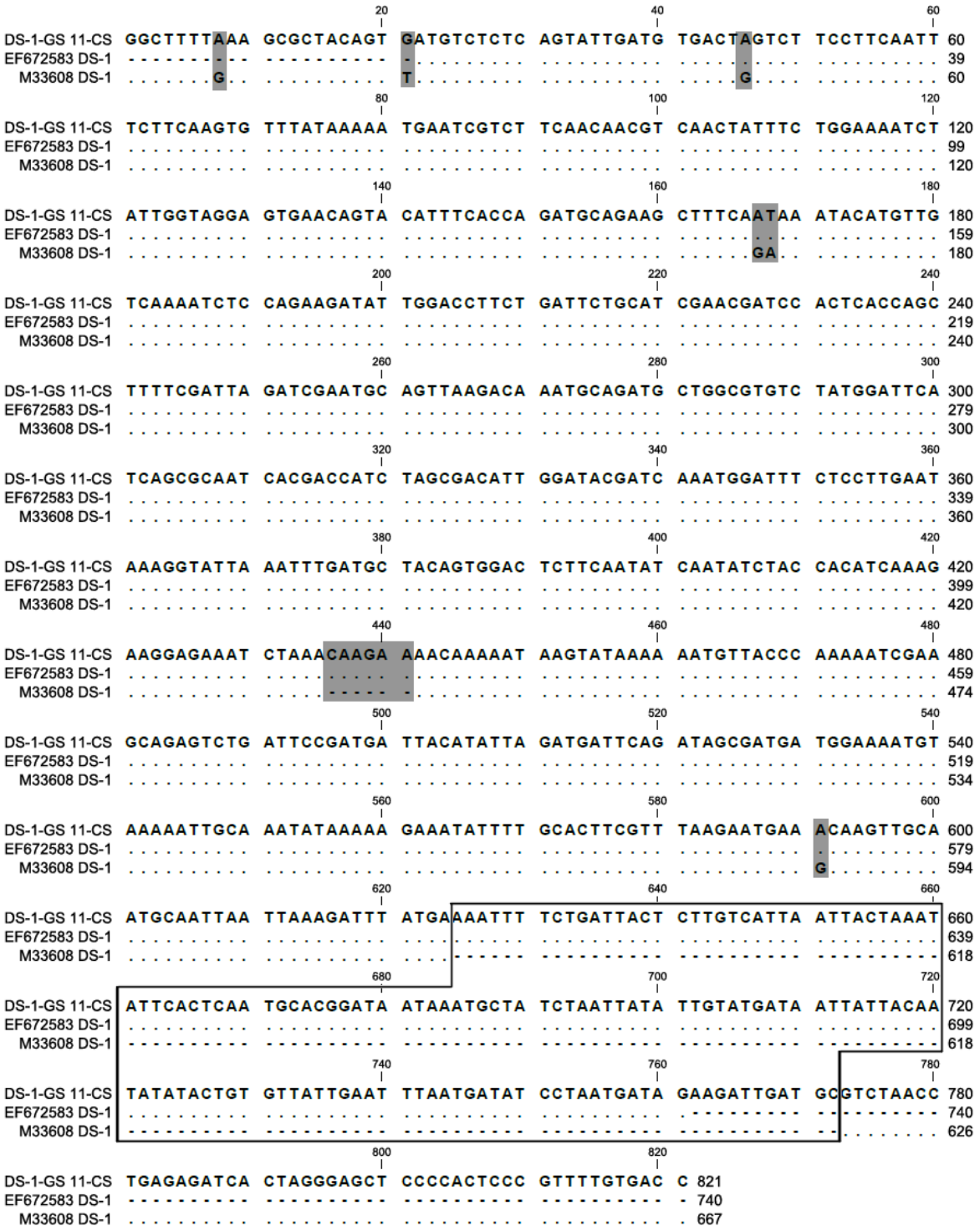

Figure 3.8. Alignment of genome segment 11 consensus nucleotide sequence (DS-1-GS 11-CS) with rotavirus DS-1 genome segment 11 sequences from GenBank (M33608 and EF672583). Boxed sequences (nucleotide 625–772) indicate the 148-nucleotide deletion in M33608. Nucleotide changes and a 6-base deletion in M33608 at position 436–441 are shaded.

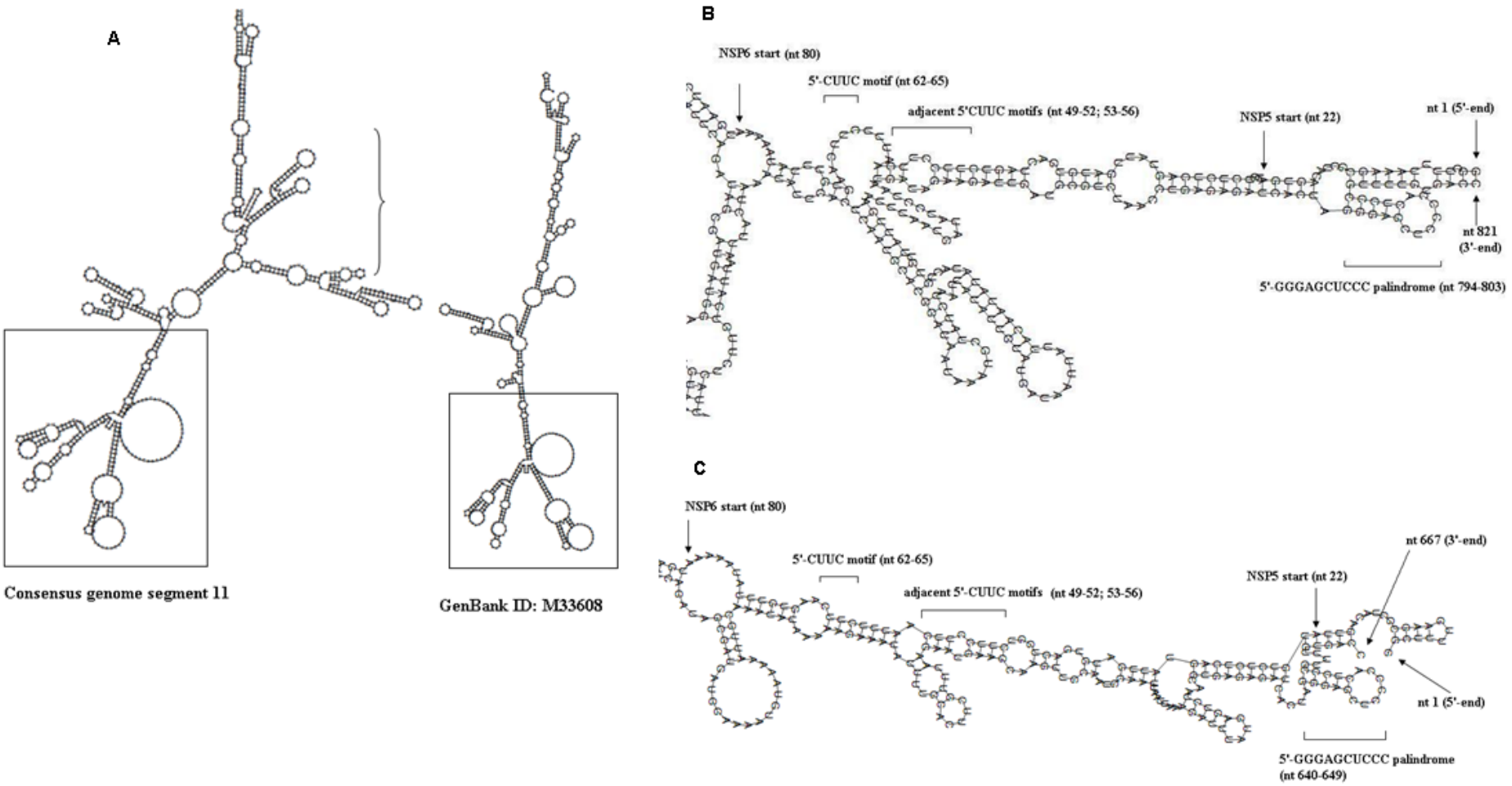
A
Consensus genome segment 11
GenBank ID: M33608
B
NSP6 start (nt 80)
5'-CUUC motif (nt 62-65)
adjacent 5'CUUC motifs (nt 49-52; 53-56)
NSP5 start (nt 22)
nt 1 (5'-end)
nt 821
(3'-end)
5'-GGGAGCUCCC palindrome (nt 794-803)
C
NSP6 start (nt 80)
5'-CUUC motif (nt 62-65)
adjacent 5'-CUUC motifs (nt 49-52; 53-56)
nt 667 (3'-end)
NSP5 start (nt 22)
nt 1 (5'-end)
5'-GGGAGCUCCC palindrome
(nt 640-649)

Figure 3.9. Comparison of the secondary RNA structures of the DS-1 consensus genome segment 11 (HQ650126) and M33608 in GenBank predicted with RNAfold. A, An overview of genome segment 11 secondary structures showing identical folding patterns in the consensus genome segment 11 sequences and M33608 (boxed). This identically folding region spans nucleotide 199–492 in M33608 and nucleotide 199–498 in the consensus genome segment 11. The bracket in the consensus sequence structure indicates nucleotides 625-772 that are not present in M33608. The secondary structure for the consensus sequence had more stems and hairpin loops than M33608. **B**, Part of the consensus genome segment 11 secondary structure (nucleotide 1-86 and 625-821) and M33608 from nucleotide 1-83 and 512-667 **(C)** indicating NSP5/6 start sites, three 5'-CUUC motifs and the 5'-GGGAGCUCCC- palindrome. Only the 5'-CUUC- motif at nucleotide 49-52 was base-paired with GAAG in both the consensus genome segment 11 and M33608. In the RNA secondary structures, the 5'-GGGAGCUCCC- palindrome folded into a conserved structure where 5'-GGGAG- formed a stem and 5'-CUCCC- was part of a loop structure. The folding of the 5' and 3' terminal ends of the consensus genome segment 11 was in a stem form while that of M33608 was an open loop.

3.3 Discussion

The objective of this study was to determine the whole genome consensus sequence of the prototype rotavirus DS-1 strain. The consensus whole genome sequence was needed for the attempt to recover rotavirus using reverse genetics. Application of the consensus sequence would eliminate the presence of minor variants with unfavourable characteristics in genetic material which could interfere with the efficiency of rotavirus recovery. The use of sequence-independent genome amplification allowed the determination of the exact 5'- and 3'-teminal end sequences. The 5'- and 3'-terminal end sequences are highly conserved in rotaviruses and the sequences obtained with 454® pyrosequencing were the same as those described before for group A rotaviruses (Wentz *et al*., 1996b, Wentz *et al*., 1996a, Tortorici *et al*., 2006). The 3'-terminal end sequence -UAUGACC-3', which was observed in genome segments 2 (VP2) and 10 (NSP4), was also reported before for the DS-1 strain in genome segment 2 (Matthijnssens *et al*., 2008a). It was also possible to describe the first 34 nucleotides at the 5'- terminus and last 30 nucleotides at the 3'-terminal end for genome segment 10 (figure 3.7). The 5'- and 3'-temini are important for the efficient synthesis of the minus RNA strand (Chen *et al*., 2001).

The use of the next generation, 454® pyrosequencing, assured that a sufficient depth of genome sequencing coverage would be obtained. The variation in the depth of coverage of 22–458-fold observed in this study (Table 3.2) was a result of several factors. For instance, the rotavirus genome segments are of variable length (751 bp–3302 bp) and the relative quantity of each genome segment would be expected to vary after PCR amplification (Cha and Thilly, 1993). Therefore, more reads of the most abundant genome segment were expected to be captured. During library preparation, high AT or high GC regions have been shown to amplify minimally or not at all and this may cause variations in the depth of sequencing coverage (Oyola *et al*., 2012, Harismendy *et al*., 2009). While higher depth of coverage allows detection of low frequency mutations, for the purpose of determining a consensus sequence of the rotavirus DS-1 strain, the average depth of coverage of 204-fold obtained was considered adequate.

Although the depth of coverage of 22-fold was lowest for genome segment 1 (VP1; Table 3.2), the absence of sequence differences between the consensus genome segment 1 sequence and genome segment 1 sequences in GenBank was expected. This expectation was based on the functional role of VP1 i.e., RNA-dependent RNA polymerase (RdRp) and that high structural conservation levels were observed in the RdRp of rotaviruses (Vasquez-del Carpio *et al*., 2006). Similarly for genome segment 3 (VP3), the amino acid sequence was highly conserved despite the 11 nucleotide differences observed (Table 3.2).

The highest number of nucleotide and amino acid sequence variations was observed in genome segment 4 (VP4). This could be attributed to high immunologic pressure since VP4 induces neutralising antibodies (Arias *et al*., 1989). However, remarkable amino acid sequence conservation was observed in the deduced sequence of VP7 (genome segment 9) when the consensus sequence was compared to VP7 sequences in GenBank. VP7 is also expected to be under significant mutational drift but high amino acid sequence conservation was also described before, following the study of 19 (serotype 1) rotavirus strains (Flores *et al*., 1988). The novel I397 observed in VP4 increases hydrophobicity in the loop region and may increase infectivity since hydrophobicity of the VP5* apex is required for membrane disruption during rotavirus cell entry (Kim *et al*., 2010a). The 30% translated sequence reads

containing N397 (Figure 3.5) suggests that a minor population variant exists within the analysed DS-1 strain. Since the amino acid N397 is hydrophilic, the virus particles containing this variation could be less inefficient at cell penetration compared to the particles with I397 in VP4. However, this assumption requires separation of these sub-species by plaque purification followed by growth curve analyses.

VP8* is an important target for neutralising antibodies and P serotype specificity (Dormitzer *et al*., 2004). The region from amino acid 106–159 is generally a variable region (Gorziglia *et al*., 1986). The probable biological effect of variations at amino acid residue positions 142 and 144 may be related to antigenicity since these residues are located on the surface of the protein structure in a hyper-variable region. According to information at the Universal Protein Resource (http://www.uniprot.org/uniprot/P11196), residue 150 is an N-linked glycosylation site in the rotavirus DS-1 strain. However, a 150N→D change observed in this study indicates that this site is not glycosylated in the DS-1 strain analysed. This change may affect the efficiency of cell entry, antigenicity and virulence (Chattopadhyay *et al.*, 2010) and should be investigated *in vivo*.

Trypsin cleavage of VP4 occurs between R230 and N231 (Arias et al., 1996). However, in this study a 230R→S change was observed. A second trypsin cleavage site present between R246 and A247 was unchanged in the consensus sequence. Cleavage at either or both the cleavage sites results in the conformational changes required for entry into cells (Yoder *et al*., 2009). The cleavage site between R246 and A247 may be preferred for cleavage in the DS-1 strain. Cleavage site preference was observed in the SA11 rotavirus strain where R247 is preferred to R241 (R230 equivalent in DS-1 strain) (Gorziglia *et al*., 1986). While the biological effect of differences at trypsin cleavage sites is not clear, a second trypsin cleavage site in VP4 has been correlated with virulence (Estes and Cohen, 1989).

The insertion and deletions in genome segment 5 (NSP1) suggest possible sequencing errors in GenBank sequences L18945 and EF672578. However variations within genome segment 5 nucleotide sequences of DS-1-like strains such as the TB-Chen, M69 and DRC88 have been reported (Matthijnssens *et al*., 2006b,

Chen *et al*., 2008, Xu *et al*., 1994). Therefore, the seven amino acid residues observed in the consensus sequence of NSP1 (Table 3.2), and in other DS-1-like strains such as GER-09, N26-02 and B1711 (Figure 3.5) suggests that the DS-1 genome segment 5 might have an alternative start codon. In addition, NSP1 is known to be the least conserved protein and varies in length among different rotavirus strains (Mitzel *et al*., 2003). For instance, NSP1 of the Wa strain contains 486 amino acids, that of SA11 contains 495 amino acids (Nakagomi and Kaga, 1995) and NSP1 of the DS-1 strain in this study was composed of 493 amino acid residues. The biological significance of short and long NSP1 proteins is not known and will need to be investigated *in vivo*. However, the most important functional feature of NSP1 is the highly conserved cysteine-rich zinc finger motif (Graff *et al*., 2007) which was conserved as expected (Figure 3.5).

The large size of genome segment 11 (821 bp; Figure 3.9) was consistent with the short electrophoretic mobility pattern for the rotavirus DS-1 strain (Nakagomi *et al*., 1989). The rotavirus DS-1 genome segment 11 sequence reported by Matsui and co-workers is missing 148 bp in the 3'-UTR (Matsui *et al*., 1990). The DS-1-like rotavirus VMRI and M69 strains contain long 3'-UTRs due to duplications of region 328–618 in the 3'-UTR (Matsui *et al*., 1990). In their report, Matsui and co-workers (1990) suggested that the long 3'-UTR may have been exogenous but they could not ascertain the source. However, intra- and intergenic recombination has been reported as a mechanism of genome segment rearrangement at the 3'-UTR (Desselberger, 1996, Cao *et al*., 2008, Donker *et al*., 2011). In this study, no duplications were observed in the 3'-UTR and the 148 bp region was confirmed to be an inherent part of the genome segment 11 nucleotide sequences. The 148 bp region may modify the formation of stem loops or cruciform structures in the dsRNA that have structural or functional roles (Figure 3.10) (Li *et al*., 2010). RNA modelling suggests that presence of the 148 bp region increase the number of stem loops formed by genome segment 11 RNA (Figure 3.10A). The additional stem loops could improve efficiency of intermolecular interactions by exposing some RNA regions or reduce the efficiency by steric hindrance. However, RNAfold predictions may not accurately predict the 2D structure and should be confirmed by a biochemical method. Although rearrangement of the 3'-UTR does not confer any growth advantage (Troupin *et al.*, 2011), the effect of a long 3'-UTR in the rotavirus DS-1

strain is not known and could be studied once a reverse genetics system is developed.

In conclusion, the application of the sequence-independent genome amplification in combination with 454® pyrosequencing allowed the determination of the consensus sequence for the rotavirus DS-1 genome strain free from cloning-bias and the limitations of Sanger sequencing. In the process, sequence differences in genome segments 2 to 11 that were identified during previous sequencing efforts, were observed and presented. The results suggest the occurrence of minor population variants in some genome segments of the cell culture-adapted rotavirus DS-1 strain such as genome segment 4. It is possible that some of the differences reported here resulted from cell culture propagation of the rotavirus DS-1 strain in different laboratories, introducing point mutations into the genome. This study made it possible to use a whole genome consensus sequence of a prototype rotavirus strain in the attempt to develop a rotavirus reverse genetics system.