Chapter 4

Characterisation of a South African rotavirus SA11 sample with 454® pyrosequencing and molecular clock phylogenetic analyses

4.0 Introduction

The simian agent 11 (SA11; RVA/Simian-tc/ZAF/SA11/1958/G3P[2]) was isolated from an overtly healthy monkey in 1958 by Dr Hubert Malherbe at the National Institute of Virology, Johannesburg, South Africa (Malherbe and Harwin, 1963, Malherbe and Strickland-Cholmley, 1967). Amongst all the known SA11 derivatives, the SA11-H96 strain is the original 1958 isolate (Malherbe and Strickland-Cholmley, 1967, Small *et al*., 2007, Matthijnssens *et al*., 2010b). Due to its ability to propagate very well in cell culture, rotavirus SA11 became a model for rotavirus biological studies, such as investigating the replication cycle and determining the function of proteins encoded by the genome segments (Estes *et al.*, 1979a, Matthijnssens *et al*., 2010b). As a result, the strain was distributed to various laboratories worldwide (Small *et al.*, 2007, Lopez and Arias, 1992, Pereira *et al*., 1986, Patton and Stacy-Phipps, 1986, Estes *et al*., 1979a). Subsequently, genome heterogeneity was described for some of the genome segments, namely genome segments 4 (VP4), 5 (NSP1), 8 (NSP2) and 7 (NSP3) (Lopez and Arias, 1992; Pereira *et al*., 1986; Small *et al*., 2007). Heterogeneity was observed in electrophoretic mobility patterns (Pereira *et al*., 1984, Pereira *et al*., 1986) and was subsequently shown to be the result of nucleotide sequence variations (Small *et al*., 2007). The heterogeneity of the rotavirus SA11 genome suggests diverse evolutionary paths for different rotavirus SA11 strains.

Strategies for recovering dsRNA viruses using core-derived wild-type transcripts were reviewed in chapter 2 (Boyce and Roy, 2007, Boyce *et al*., 2008, Matsuo *et al*., 2010). In this study, a similar approach was pursued and the rotavirus SA11 strain was selected for use in the attempt to recover rotavirus from wild-type transcripts due to its prolific propagation properties in cell culture. It was thought that the rapid propagation of rotavirus SA11 could enhance the probability of rotavirus recovery through efficient protein synthesis, replication and packaging of rotavirus SA11

particles. Furthermore, the plasmid constructs in the two rotavirus single genome segment reverse genetics systems contained genome segments from the rotavirus SA11 strain (Komoto *et al.*, 2006, Trask *et al*., 2010b). This suggests that the rotavirus SA11 strain may be a key strain for developing a rotavirus reverse genetics system. This hypothesis was also based on observations that, for the measles reverse genetics system, only certain strains such as the Edmonston strain (an over-attenuated laboratory strain that is adapted to non-lymphoid cells) could be used in the development of a reverse genetics system (Takeda *et al*., 2000).

In the event that rotavirus SA11 was recovered by reverse genetics, a comparison of the sequence of the recovered rotavirus to that of the original strain would be required. To date, similar to the rotavirus DS-1 strain (section 3.0), the nucleotide sequences of rotavirus SA11 strains deposited in GenBank since the early 1980s were all determined using Sanger sequencing. The sequences were either generated directly from PCR amplicons of the genome segments or from amplicons that were first cloned into plasmids followed by Sanger sequencing (Both *et al*., 1983, Small *et al*., 2007, Liu and Estes, 1989). No local sequencing of a rotavirus SA11 strain has been performed in South Africa before. Therefore, it was decided to determine the whole genome consensus sequence of cell-culture adapted rotavirus SA11 stored in South Africa, of which the passage history was unknown, with sequence-independent genome amplification (Potgieter *et al*., 2009) and 454® pyrosequencing (Roche). Following this, the consensus sequence was compared to all the rotavirus SA11 sequences held in GenBank. In addition, the molecular clock phylogenetic analysis tool would be applied to further characterise the South African rotavirus SA11 strain to demonstrate the strain's evolutionary relationship to other known rotavirus SA11 strains.

4.1 Materials and methods

4.1.1 Cells, virus samples and propagation

MA104 cells were maintained as described in section 3.1.1. A sample of cell culture-adapted rotavirus SA11 was received from Mrs. I. Peenze of the Diarrhoeal Pathogens Research Unit, University of Limpopo (Pretoria, South Africa). The sample was received with an unknown passage history. The cell culture medium-

containing vials were only labelled “rotavirus SA11” (Figure 4.1). The cell line in which the virus had been propagated before was also not known. No additional information such as the origin of the sample was available. Rotavirus in an aliquot of the sample was activated with 10 µg/ml porcine trypsin IX for 30 minutes at 37 ^{o}C. Adsorption onto MA104 cells was performed by incubating the trypsin-activated rotavirus on a monolayer of MA104 cells in a 25 cm^2 flask (Nunc^{TM}) for 30 minutes at 37 ^{o}C. The infected MA104 cells were incubated in DMEM (Hyclone) containing 1 µg/ml trypsin and supplemented with 1% NEAA (Lonza) and 1% PSA (Lonza) (section 3.1.1). Following the observation of CPE, a second propagation in MA104 cells in a 75 cm^2 flask (NuncTM) was performed for 3 days and extraction of dsRNA was performed as described previously (section 3.1.2).

Figure 4.1. A photograph of the rotavirus SA11 samples obtained from the Diarrhoeal Pathogens Research Unit indicating the only information received for these samples.

4.1.2 454® pyrosequencing and sequence data analyses

Oligo-ligation and sequence-independent genome amplification was carried out as described in sections 3.1.2–3.1.5, except that cDNA synthesis was performed with AMV reverse transcriptase (Fermentas). The 454® pyrosequencing was performed at Inqaba Biotec^{TM} (section 3.6). Some nucleotide sequence alignments were performed with the *mVISTA* visualisation module (Frazer *et al.*, 2004). Distance matrix inference of phylogeny was conducted using the Maximum Composite

Likelihood model with MEGA software v 5.05 (Tamura *et al*., 2011, Tamura *et al*., 2004). Accession numbers of the consensus sequences of the SA11 genome segments determined in this study and those retrieved from GenBank are listed in Table 4.1.

Table 4.1. List of accession numbers of the SA11 rotavirus consensus sequences determined in this study, and sequences retrieved from GenBank. The abbreviated common strain names were used for simplicity.

Genome segment	SA11-N2 [a]	SA11-N5 [a]	SA11-H96	SA11-Both	SA11-5N	SA11-5S	SA11-30/19	SA11-30/1A	O Agent	Other SA11 variants
1(VP1)	JN827244	JQ688673	DQ838640 [b]	X16830	DQ838636	DQ838637	DQ838638	DQ838639	NC	DQ457016 DQ838601 AF015955
2(VP2)	JN827245	JQ688674	DQ838635 [b]	X16831	DQ838631	DQ838632	DQ838633	DQ838634	NC	L33364 AF474406 L20123
3(VP3)	JN827246	JQ688675	DQ838645 [b]	X16387	DQ838641	DQ838642	DQ838643	DQ838644	NC	DQ838600 X16062
4(VP4)	JN827247	JQ688676	DQ841262 [b]	X14204	DQ838602	DQ838603	DQ838604	DQ838605	DQ838596	M23188 D16345 D16346
5(NSP1)	JN827248	JQ688677	DQ838599 [b] JF791801 [c]	X14914	NS	AF290884	NS	AF290882	NC	L18944 AF290883 AF290881
6(VP6)	JN827249	JQ688678	DQ838650 [b] JF791806 [c]	AY187029	DQ838646	DQ838647	DQ838648	DQ838649	NC	L33365
7(NSP3)	JN827250	JQ688679	DQ838610 [b] JF791803 [c]	X00355	DQ838606	DQ838607	DQ838608	DQ838609	NC	AY065843 GU550506 EF460843 M87502
8(NSP2)	JN827252	JQ688680	DQ838615 [b] JF791802 [c]	J02353	DQ838611	DQ838612	DQ838613	DQ838614	DQ838597	L04531 L04532 L20901
9(VP7)	JN827253	JQ688681	DQ838620 [b]	V01190	DQ838616	DQ838617	DQ838618	DQ838619	NC	K02028
10(NSP4)	JN827254	JQ688682	DQ838625 [b] JF791804 [c]	K01138	DQ838621	DQ838622	DQ838623	DQ838624	NC	AF087678
11(NSP5/6)	JN827255	JQ688683	DQ838630 [b] JF791805 [c]	X07831	DQ838626	DQ838627	DQ838628	DQ838629	NC	AF306493 M28347

[a] Sequences determined in this study
[b] Sequences of RVA/Simian-tc/ZAF/SA11-N2/1958/G3P[2] determined by Small and co-workers (2007)
[c] Sequences of RVA/Simian-tc/ZAF/SA11-N2/1958/G3P[2] determined by Dutta and co-workers (2011)
NC: Sequences not compared
NS: No sequence data available

4.1.3 Molecular clock phylogenetic analyses

The nucleotide sequences of the open reading frames (ORF) of the consensus genome segments determined in this study, and corresponding nucleotide sequences retrieved from GenBank were used to determine genome segment divergence based on molecular clock evolutionary analysis. ORF nucleotide sequences were aligned using BioEdit sequence alignment editor v7.1.3 (Hall, 1999). Bayesian phylogenetic reconstructions were performed with Markov chain Monte Carlo (MCMC) in Bayesian Evolutionary Analysis by Sampling Trees software v1.6.1 (BEAST) (Drummond and Rambaut, 2007). Molecular clock evolutionary analysis in BEAST was performed with the Hasegawa Kishino Yano (HKY) substitution model (Hasegawa *et al*., 1985) with gamma distributed rate variation, uncorrelated lognormal relaxed clock model and a coalescent constant size tree prior. The HKY substitution model was selected by testing with jModelTest v 0.1.1 (Posada, 2008). Four separate MCMC analytical runs, at 100 million generations per run, were performed for each genome segment. Data from the four runs were combined using LogCombiner v1.6.1. The combined analysis was diagnosed using Tracer v1.5 (http://tree.bio.ed.ac.uk/tracer). Maximum clade trees were annotated using TreeAnnotator v1.6.1 and visualised with FigTree v1.3.1 (http://tree.bio.ed.ac.uk/figtree/).

4.2 Results

4.2.1 Determination of the consensus sequence of the rotavirus SA11 genome

The consensus sequence of the viral genome represents the predominant and most genetically fit sequence of a virus population (Domingo *et al.*, 2006, Domingo *et al*., 2012, González-López *et al*., 2004). No rotavirus SA11 nucleotide sequence has been determined locally in South Africa. The whole genome consensus sequence of cell-culture adapted rotavirus SA11 stored in South Africa was determined with sequence-independent genome amplification (Potgieter *et al*., 2009) and 454® pyrosequencing (Roche). Pyrosequencing of the sequence-independent-amplified cDNA genome of the rotavirus SA11 sample generated 29849 reads of approximately 400 bp each which were assembled into contigs. Twelve full-length consensus genome segment sequences were obtained. A single consensus

sequence was present for each genome segment, except for genome segment 8 (NSP2) for which two distinct consensus sequences were present. All the twelve genome segments contained the typical rotavirus group A 5'-terminal sequence, 5'-GGC(U/A)$_7$-. Genome segments 1 (VP1), 2 (VP2), 3 (VP3), 4 (VP4), 6 (VP6), 8 (NSP2) and 9 (VP7) contained the typical -UGUG**A**CC-3' 3'-terminal sequence. However, genome segments 5 (NSP1) and 7 (NSP3) contained the 3'-terminal sequences -UGUGA**A**CC-3' and -UGUG**G**CC-3' respectively.

One consensus genome segment 8 sequence (GenBank ID: JN827252) was DS-1-like and assigned the N2 genotype with RotaC (Maes *et al*., 2009). The second consensus genome segment 8 sequence (GenBank ID: JQ688680) was assigned the N5 genotype and it was 100% identical to that of SA11-H96 which was sequenced in the USA (Small *et al*., 2007). The N2-genotyped genome segment 8 sequence (JN827252) was 100% identical to that of SA11-Both. Two sets of genome segment sequences encoding VP6 and NSP1–NSP5 of SA11-H96 available in GenBank (Table 4.1) were sequenced in by Dutta *et al*. (2011) and the Small *et al*. (2007). Visualisation of the alignment of the two genome segment 8 sequences obtained in this study with *mVISTA* showed that the two sequences were substantially different (Figure 4.2). This suggested the presence of two rotaviruses in the original sample but, for each virus, 10 of the 11 genome segments could not be separated. The viruses were named RVA/Simian-tc/ZAF/SA11-N5/1958/G3P[2] (SA11-N5) and RVA/Simian-tc/ZAF/SA11-N2/1958/G3P[2] (SA11-N2). The full genotypes assigned using the RotaC v2.0 web tool (Maes *et al*., 2009) were G3-P[2]-I2-R2-C5-M5-A5-Nx-T5-E2-H5 where x was either 2 or 5. The average depth of pyrosequencing coverage obtained was 367-fold. Genome segment 1 had the lowest coverage of 125-fold, while genome segment 8 of SA11-N2 (JN827252) displayed the highest coverage of 610-fold. Only genome segments 1 and 3 had an average depth of coverage lower than 200-fold (Table 4.2). SA11-N5 and SA11-N2 were present in the virus sample in a ratio of approximately 1:2 based on the ratios of their respective genome segment 8 sequence reads in the pyrosequence data.

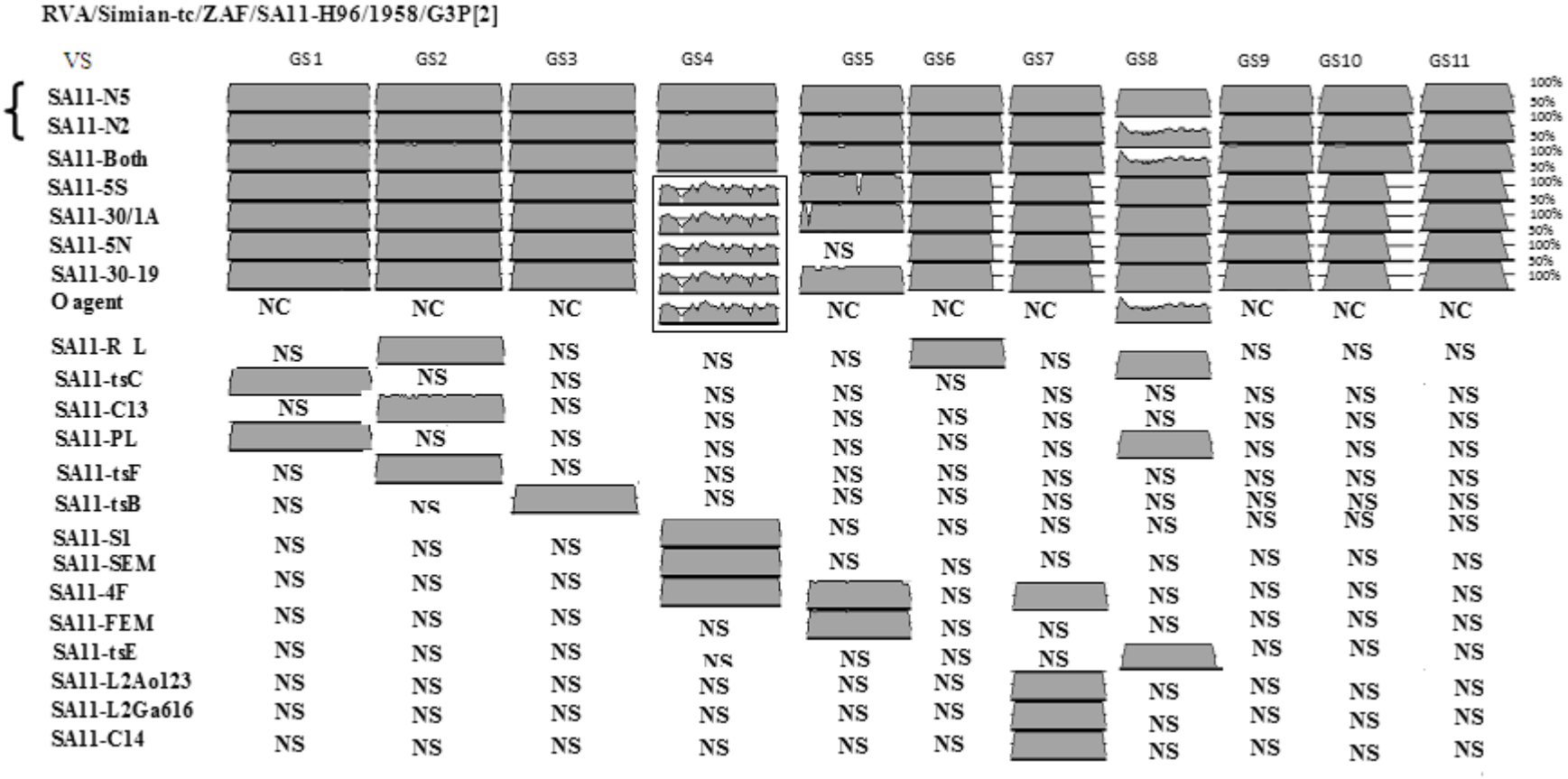


Figure 4.2. The *mVISTA* visualisation alignment comparing the nucleotide sequences of SA11-N2 and SA11-N5 (indicated by the bracket at the left side of the figure) to SA11 sequences in GenBank. Genome sequence is abbreviated with GS. The SA11-H96 used for comparison was sequenced by Small et al., 2007. Genome segment 4 (VP4) sequences with a P[1] type are boxed. NC represents no comparison performed while NS indicates no sequence in GenBank. The GenBank accession numbers were used for some genome segments where no specific strain names were available.

4.2.2 Comparison of the rotavirus SA11 consensus sequence to rotavirus SA11 sequences in GenBank

It is an established fact that nucleotide sequences are divergent among the rotavirus SA11 variants (Small *et al*., 2007, Matthijnssens *et al*., 2010b). Following the determination of the consensus whole genome sequence of SA11-N2 and SA11-N5, the consensus sequences were compared to the nucleotide sequences of all rotavirus SA11 variants in GenBank. Nucleotide differences were observed in all genome segments when the consensus genome segment sequences of SA11-N5, SA11-N2 and nucleotide sequences of genome segments of SA11 derivatives available in GenBank were compared to each other. For instance, RotaC nucleotide sequence similarity profiling indicated that the closest nucleotide genome segment sequences in GenBank to those of SA11-N5 and SA11-N2 were as follows: With the exception of genome segments 4 (VP4) and 5 (NSP5), the consensus nucleotide sequence of the other nine genome segments of SA11-N5 and SA11-N2 were identical to the sequence of a corresponding genome segment of one of 5 different SA11 derivatives (Table 4.2). Furthermore, distance matrices showed that the consensus sequence of eight genome segments of SA11-N5 and SA11-N2, were most similar to those of SA11-H96 (Small *et al.*, 2007). Genome segments 1 (VP1), 2 (VP2), 6 (VP6) 7 (NSP3), 8 (N5 typed NSP2) 9 (VP7) 10 (NSP4) and 11 (NSP5/6) were identical to those of SA11-H96 sequenced by Small and co-workers (2007; Appendix 1). Genome segment 5 (NSP1) was 99.7% identical to that of SA11 H-96 (Table 4.2). Genome segment 3 (VP3) was 100% identical to that of SA11-5S and 99.96% identical to that of SA11-H96 (Figure 4.1; Table 1; Appendix 1 Figure 1C). Genome segment 4 (VP4; JN827247/JQ688676) was 99.6% identical that of SA11-Both (Figure 4.1; Table 1; Appendix 1 Figure 1D). A 99% nucleotide identity was observed between the consensus genome segment 4 (VP4) sequence of SA11-N5, SA11-N2 and the corresponding nucleotide sequences of SA11-H96 and other SA11 strains with a P[2] genotype (Fig. 4.2; Appendix 1 Figure 1D). In contrast, only 77% nucleotide identity was observed between the consensus genome segment 4 and SA11 strains with a P[1] genotype (Fig. 4.2).

Table 4.2. Comparison of the consensus nucleotide and deduced amino acid sequences of SA11-N2 and SA11-N5 to sequences of SA11-H96 (Small *et al*., 2007).

Genome segment (N2/N5)	Size (bp)	Amino acids in coding region	Average depth of coverage (-fold)	NT similarity to rotavirus SA11 variants (RotaC)	Molecular clock rate (NT/site/yr)	Nucleotide differences			Amino acid differences (H96→N2/N5)
						Point changes	Insertions	Deletions	
1 (VP1)	3302	1088	125	SA11-H96 (100%)	1.3×10^{-4}	None	None	None	None
2 (VP2)	2693	882	609	SA11-H96 (100%)	7.5×10^{-5}	None	None	None	None
3 (VP3)	2591	835	156	SA11-5S (100%)	1.4×10^{-6}	2	None	None	650L→H
4 (VP4)	2362	775	358	SA11-Both (99.6%)	1.4×10^{-5}	9	None	None	72T→M, 157P→S, 187A→G, 261F→L, 332Y→Ser, 366V→M, 388S→R[a]
5 (NSP1)	1614	496	225	SA11-H96 (99.7%)	4.8×10^{-5}	8	None	None	36E→A, 84-86QQLu→RTV, 96L→Q, 137K→L, 188E→D
6 (VP6)	1356	397	585	SA11-5N (100%)	2.1×10^{-5}	3	None	None	None
7 (NSP3)	1105	313	305	SA11-H96 (100%)	1.3×10^{-4}	2	None	None	None
8(NSP2)^{N2}	1059	317	610	SA11-Both (100%)	ND	205	None	None	45 amino acid differences
8(NSP2)^{N5}	1059	317	265	SA11-H96 (100%)	1.9×10^{-4}	None	None	None	None
9 (VP7)	1063	326	487	SA11-30/1A (100%)	2.9×10^{-5}	2	None	None	29→T[a], 309V→A[a]
10 (NSP4)	751	175	495	SA11-H96 (100%)	ND	4	None	None	33F→L[a], 93G→D, 114D→G[a]
11 (NSP5/6)	667	198/93	429	SA11-Both (100%)	8.5×10^{-5}	1	None	None	None

[a] Novel amino acid residues detected in this study.
NT: nucleotide
ND: not determined due to too few nucleotide sequences (genome segment 8) or zero-length interior branches (genome segment 10).

The deduced amino acid sequences from pyrosequence data revealed for the first time the occurrence of five amino acids encoded in the ORFs of 3 genome segments. A novel 388R was found in the antigenic domain of VP4 in 11% (29/264) of pyrosequence reads that indicated G, instead of T(U), at nucleotide position 1996 (Table 4.2; Figure 4.3A). The nucleotide T(U), in the consensus sequences, results in a serine residue. Amino acid position 388 is located on the surface of the predicted VP5* structure (data not shown) inferred using PDB ID: 1SLQ (Dormitzer *et al*., 2004). Novel amino acids 29T (in the signal peptide located in the N-terminal region of the deduced VP7 sequence) and 309A (in the C-terminal region of VP7) were also detected. Amino acid 29T was encoded by 10.6% (57/537) pyrosequence reads that contained nucleotide C at position 135, while amino acid 309A was encoded by 7.2% (36/501) pyrosequence reads that contained the nucleotide C at position 970 (Figure 4.3B). The consensus sequence contained nucleotides T(U) and A at positions 135 and 970, resulting in the amino acids isoleucine and valine, respectively. For genome segment 10 (NSP4; JN827254/JQ688682), the two novel amino acids 33L and 114G were encoded by 28.4% (76/268) pyrosequence reads that had nucleotide C at position 138, and 24.5% (108/333) of pyrosequence reads that indicated nucleotide G at position 382, respectively (Figure 4.3C). The consensus NSP4 (genome segment 10) sequence had amino acids 33F and 114D. No amino acid differences were observed in the deduced protein sequences of NSP2 and NSP5/6 sequences of SA11-N5 when compared to the respective deduced amino acid sequences of SA11-H96 (Small *et al*., 2007). For SA11-N2 and SA11-H96 (Small *et al*., 2007), no amino acid sequence differences were observed between their deduced NSP5/6 sequences.

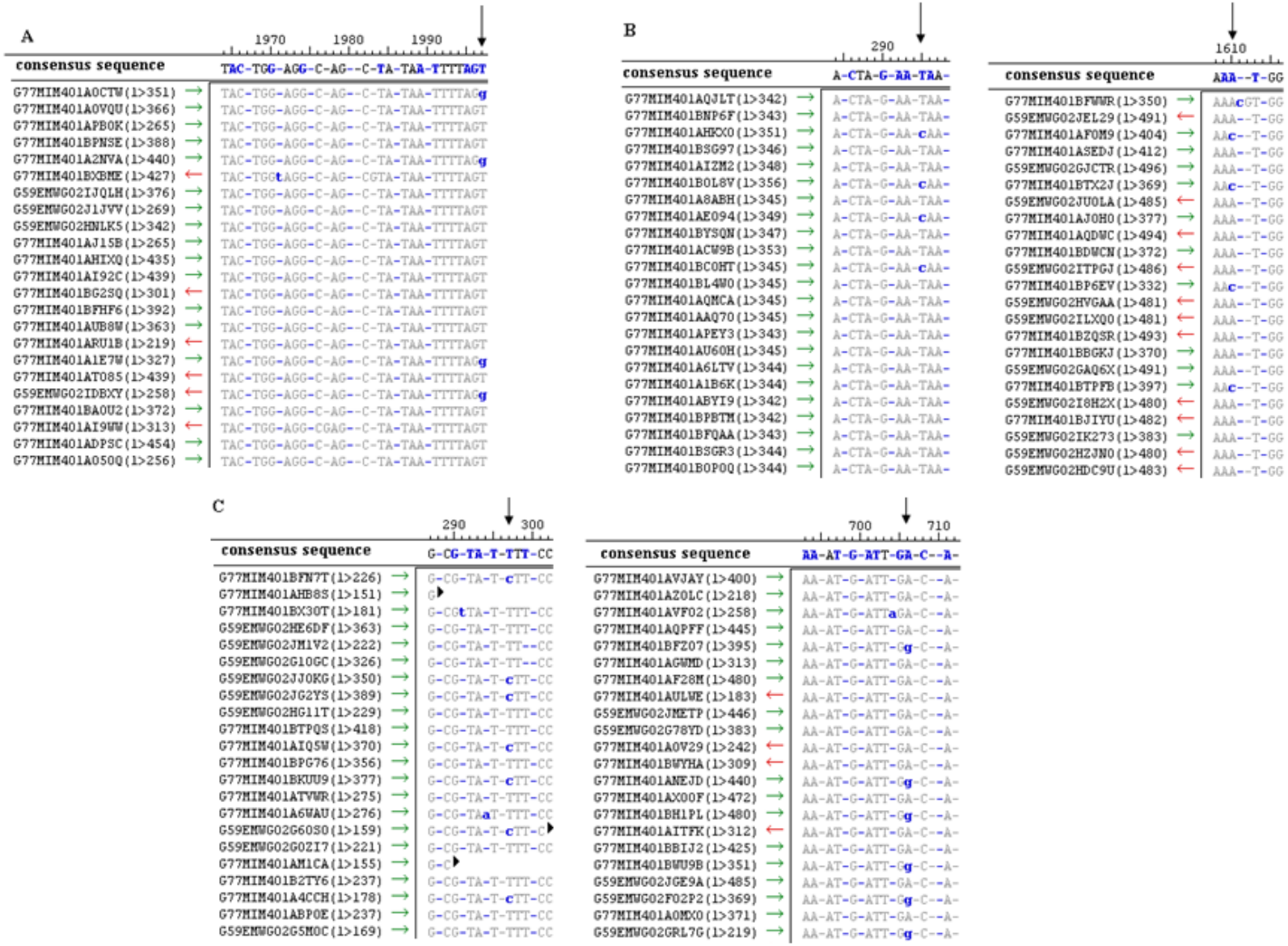

A
1970
1980
1990
consensus sequence
TAC-TGG-ACG-C-AG--C-TA-TAA-TTTTAGT
G77MIM401A0CTW(1>351)
G77MIM401A0VQU(1>366)
G77MIM401APB0K(1>265)
G77MIM401BPNSE(1>388)
G77MIM401A2NVA(1>440)
G77MIM401BXBME(1>427)
G59EMWG02IJQLH(1>376)
G59EMWG02J1JVV(1>269)
G59EMWG02HNLK5(1>342)
G77MIM401AJ15B(1>265)
G77MIM401AHIXQ(1>435)
G77MIM401AI92C(1>439)
G77MIM401BG2SQ(1>301)
G77MIM401BFHF6(1>392)
G77MIM401AUB8W(1>363)
G77MIM401ARU1B(1>219)
G77MIM401A1E7W(1>327)
G77MIM401AT085(1>439)
G59EMWG02IDBXY(1>258)
G77MIM401BA0U2(1>372)
G77MIM401AI9WW(1>313)
G77MIM401ADPSC(1>454)
G77MIM401A050Q(1>256)
B
290
consensus sequence
A-CTA-G-AA-TAA-
G77MIM401AQJLT(1>342)
G77MIM401BNP6F(1>343)
G77MIM401AHKX0(1>351)
G77MIM401BSG97(1>346)
G77MIM401AIZM2(1>348)
G77MIM401B0L8V(1>356)
G77MIM401A8ABH(1>345)
G77MIM401AE094(1>349)
G77MIM401BYSQN(1>347)
G77MIM401ACW9B(1>353)
G77MIM401BC0HT(1>345)
G77MIM401BL4W0(1>345)
G77MIM401AQMCA(1>345)
G77MIM401AAQ70(1>345)
G77MIM401APEY3(1>343)
G77MIM401AU60H(1>345)
G77MIM401A6LTV(1>344)
G77MIM401A1B6K(1>344)
G77MIM401ABYI9(1>342)
G77MIM401BPBTM(1>342)
G77MIM401BFQAA(1>343)
G77MIM401BSGR3(1>344)
G77MIM401B0P0Q(1>344)
1610
consensus sequence
AAA--T-GG
G77MIM401BFWWR(1>350)
G59EMWG02JEL29(1>491)
G77MIM401AF0M9(1>404)
G77MIM401ASEDJ(1>412)
G59EMWG02GJCTR(1>496)
G77MIM401BTX2J(1>369)
G59EMWG02JU0LA(1>485)
G77MIM401AJ0H0(1>377)
G77MIM401AQDWC(1>494)
G77MIM401BDWCN(1>372)
G59EMWG02ITPGJ(1>486)
G77MIM401BP6EV(1>332)
G59EMWG02HVGAA(1>481)
G59EMWG02ILXQ0(1>481)
G77MIM401BZQSR(1>493)
G77MIM401BBGKJ(1>370)
G59EMWG02GAQ6X(1>491)
G77MIM401BTPFB(1>397)
G59EMWG02I8H2X(1>480)
G77MIM401BJIYU(1>482)
G59EMWG02IK273(1>383)
G59EMWG02HZJN0(1>480)
G59EMWG02HDC9U(1>483)
C
290
300
consensus sequence
G-CG-TA-T-TTT-CC
G77MIM401BFN7T(1>226)
G77MIM401AHB8S(1>151)
G77MIM401BX30T(1>181)
G59EMWG02HE6DF(1>363)
G59EMWG02JM1V2(1>222)
G59EMWG02G10GC(1>326)
G59EMWG02JJ0KG(1>350)
G59EMWG02JG2YS(1>389)
G59EMWG02HG11T(1>229)
G77MIM401BTPQS(1>418)
G77MIM401AIQ5W(1>370)
G77MIM401BPG76(1>356)
G77MIM401BKUU9(1>377)
G77MIM401ATVWR(1>275)
G77MIM401A6WAU(1>276)
G59EMWG02G6OSO(1>159)
G59EMWG02G0ZI7(1>221)
G77MIM401AM1CA(1>155)
G77MIM401B2TY6(1>237)
G77MIM401A4CCH(1>178)
G77MIM401ABP0E(1>237)
G59EMWG02G5M0C(1>169)
700
710
consensus sequence
AA-AT-G-ATT-GA-C--A-
G77MIM401AVJAY(1>400)
G77MIM401AZ0LC(1>218)
G77MIM401AVF02(1>258)
G77MIM401AQPFF(1>445)
G77MIM401BFZ07(1>395)
G77MIM401AGWMD(1>313)
G77MIM401AF28M(1>480)
G77MIM401AULWE(1>183)
G59EMWG02JMETP(1>446)
G59EMWG02G78YD(1>383)
G77MIM401A0V29(1>242)
G77MIM401BWYHA(1>309)
G77MIM401ANEJD(1>440)
G77MIM401AX00F(1>472)
G77MIM401BH1PL(1>480)
G77MIM401AITFK(1>312)
G77MIM401BBIJ2(1>425)
G77MIM401BWU9B(1>351)
G59EMWG02JGE9A(1>485)
G59EMWG02F02P2(1>369)
G77MIM401A0MX0(1>371)
G59EMWG02GRL7G(1>219)

Figure 4.3. Contig alignments depicting the identification of novel minority coding sequences in rotavirus SA11 in genome segments 4 (VP4), 9 (VP7) and 10 (NSP4). The position where nucleotide variants were observed is indicated with an arrow. **A**, Minority variants in genome segment 4 contain the nucleotide G at position 1996, resulting in encoding of a novel 388R in the antigenic region of the VP5* region of VP4. The consensus sequence contained the nucleotide T(U) at the same position which resulted in the encoding of amino acid 388S. **B**, Nucleotide sequence variations in genome segment 9 (VP7). The consensus sequence contained the nucleotide T(U) at position 135 and A at 970 which encoded the amino acids 29Ile and 309Val, respectively. A minority population variant contained C at nucleotide position 135 and C at nucleotide position 970. Therefore, novel amino acid residues 29T and 309A were encoded. **C**, Novel minority nucleotide sequences detected in genome segment 10 (NSP4). The consensus sequence contained the nucleotides T(U) and A at positions 138 and 382 to result in the encoding of amino acids 33F and 114D respectively. A minor population variant contained the nucleotides C and G at positions 138 and 382 resulting in the respective encoding of novel amino acids 33L and 114G.

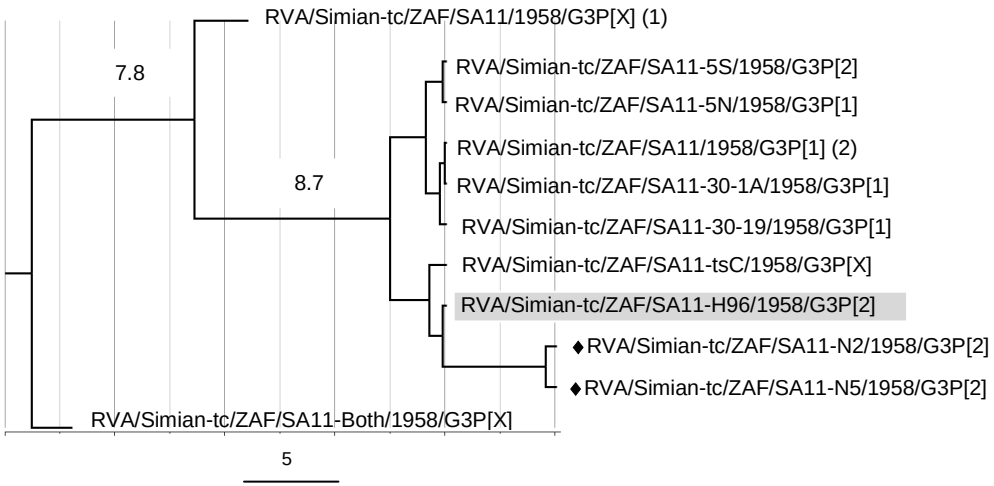
4.2.3 Molecular clock analyses and phylogenetic relationships

The molecular clock hypothesis is based on the idea that molecular evolution occurs at a uniform rate over time (Kumar, 2005, Gojobori *et al*., 1990). Molecular clock analysis indicates whether evolutionary rates vary among strain variants and viral genome segments (Hayashida *et al*., 1985, Schierup and Hein, 2000). To further characterise the rotavirus SA11 sequenced in this study by determining its evolution in relationship to other SA11 derivatives, molecular clock evolutionary analysis was carried out for all 12 consensus genome segment sequences of SA11-N5 and SA11-N2. However, the molecular clock analysis failed for 2 of the genome segments. In the case of genome segment 8 (NSP2), there were only two nucleotide sequences available in GenBank for the N2-genotype. It was not possible to do molecular clock analysis with so few sequences. The genome segment 10 (NSP4) molecular clock results were excluded from the analyses because zero-length interior branches were obtained (Coddington and Scharff, 1994). For the other genome segments, rates of evolution were in the range 1.4×10^{-6} – 1.9×10^{-4} nucleotide substitutions/site/year. The range between the lowest and highest substitution rates differed by a magnitude of ~100-fold. The lowest evolutionary rate was observed for genome segment 3 (VP3), while the highest was observed for genome segment 1 (VP1; Table 4.2). For genome segment 1 (VP1), analysis of the nucleotide substitution rate for the region which spans nucleotide position 778–2610, encoding amino acids at position 260–

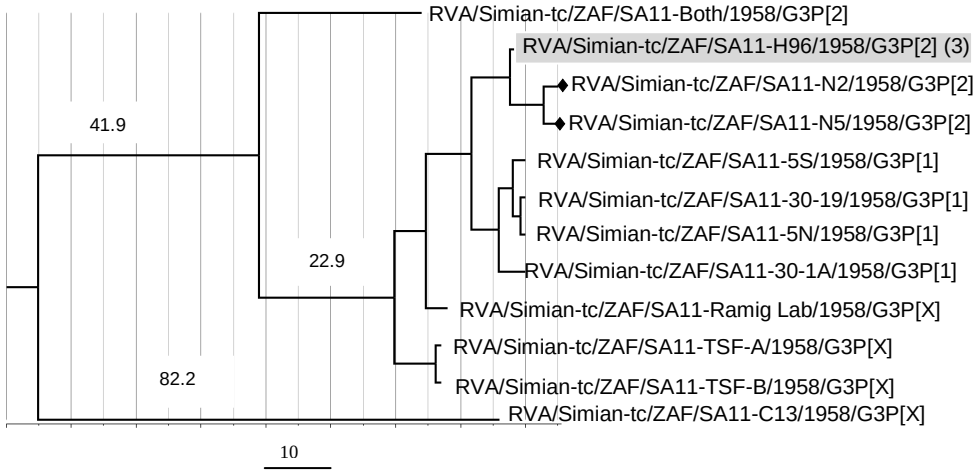
870 was performed. This part of VP1 contains a unique RdRP region at the N-terminal end followed by the fingers, palms and thumb regions of the polymerase domain (Vasquez-del Carpio *et al*., 2006, Ogden *et al*., 2012, Lu *et al*., 2008). The nucleotide substitution rate obtained for this polymerase-encoding region was 9.9 x 10^{-5} nucleotide substitutions/site/year with a CV of 0.6. This nucleotide substitution rate was lower than the overall rate of 1.3 x 10^{-4} nucleotide substitutions/site/year. However, the Ka/Ks ratio of non-synonymous to synonymous nucleotide substitution rates (Hurst, 2002, Yang and Bielawski, 2000) obtained for the genome segment 1 ORF was 0.6.

The maximum clade credibility trees (MCC) trees showed that all genome segments encoding the double-layered particle, except genome segment 3 (VP3), were most closely related to SA11-H96 (Figure 4.4A–C and F). Genome segments encoding NSP1, NSP3 and the viroplasm-forming NSP2 (SA11-N5; JQ688680) and NSP5/6 were also closely related to SA11-H96 (Figure 4.4E, G, H and J). MCC trees constructed showed that genome segment 4 was most similar to that of SA11-SEM (Figure 4.4D). Genome segments 3 and 9 (VP7) were closely related to those of SA11-5S and SA11-30-19 respectively (Figure 4.4A and I).

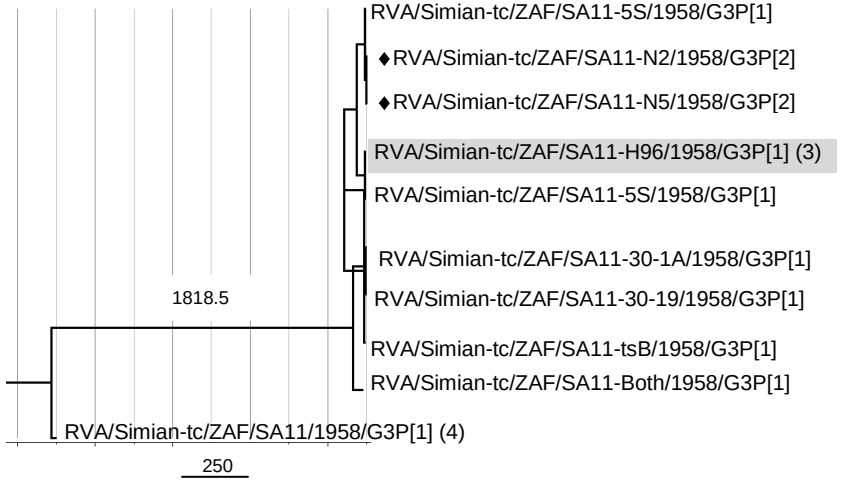
A. Genome segment 1 (VP1)



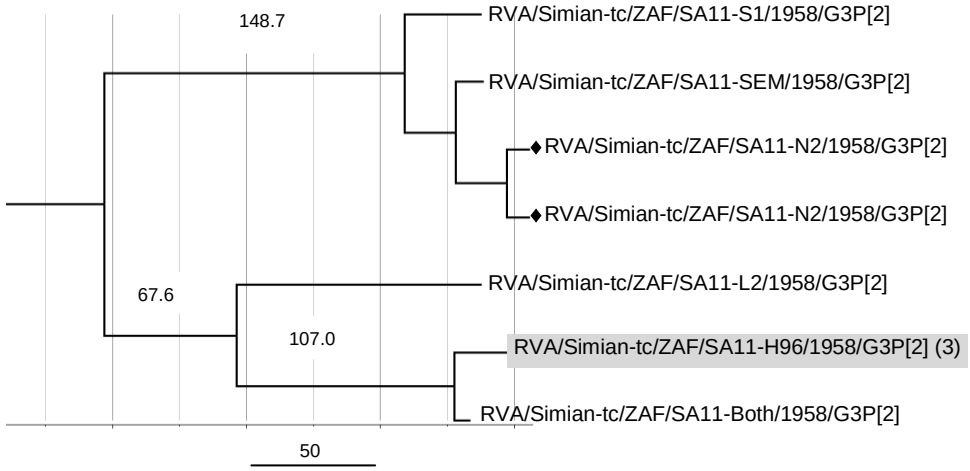
B. Genome segment 2 (VP2)



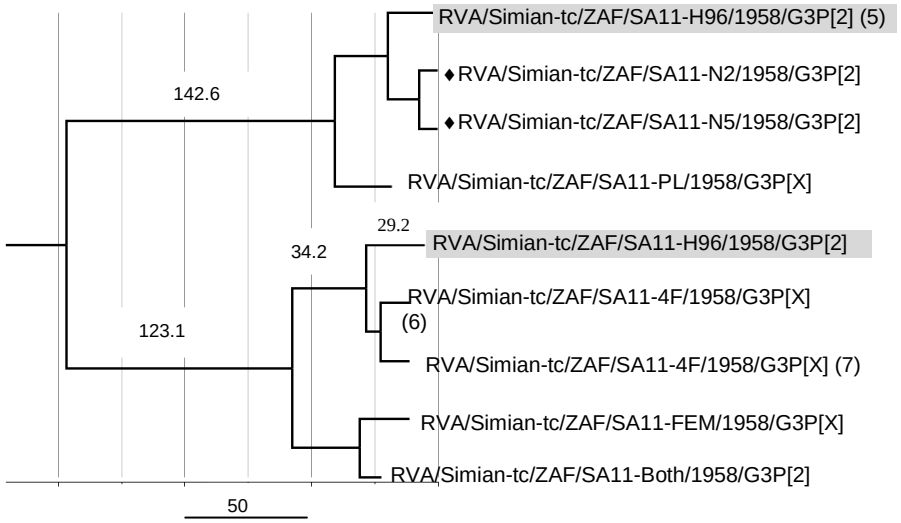
C. Genome segment 3 (VP3)



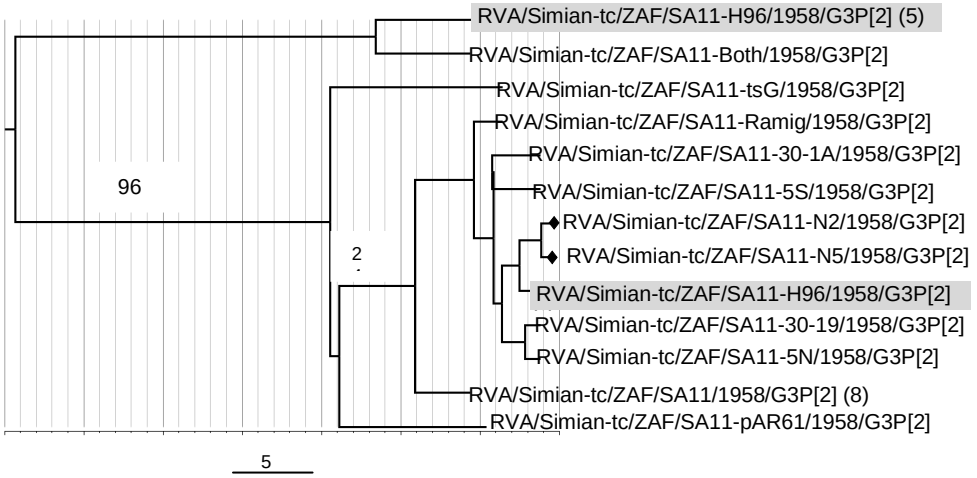
D. Genome segment 4 (VP4)



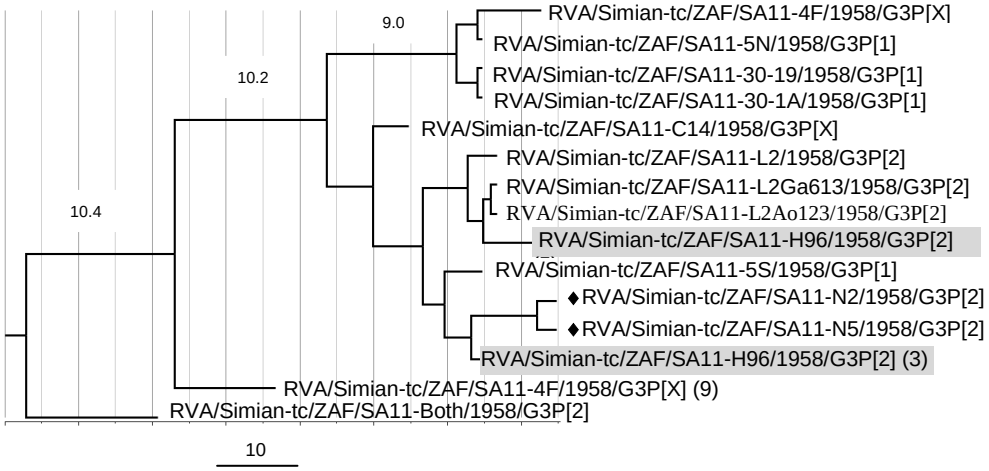
E. Genome segment 5 (NSP1)



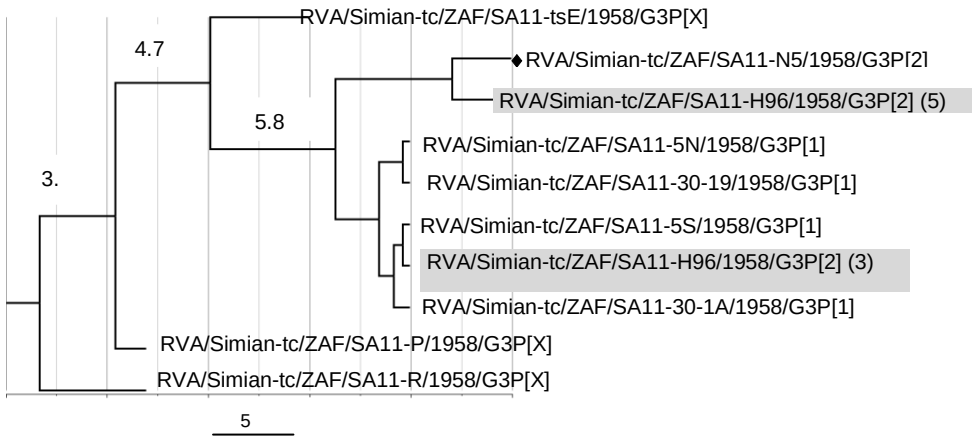
F. Genome segment 6 (VP6)



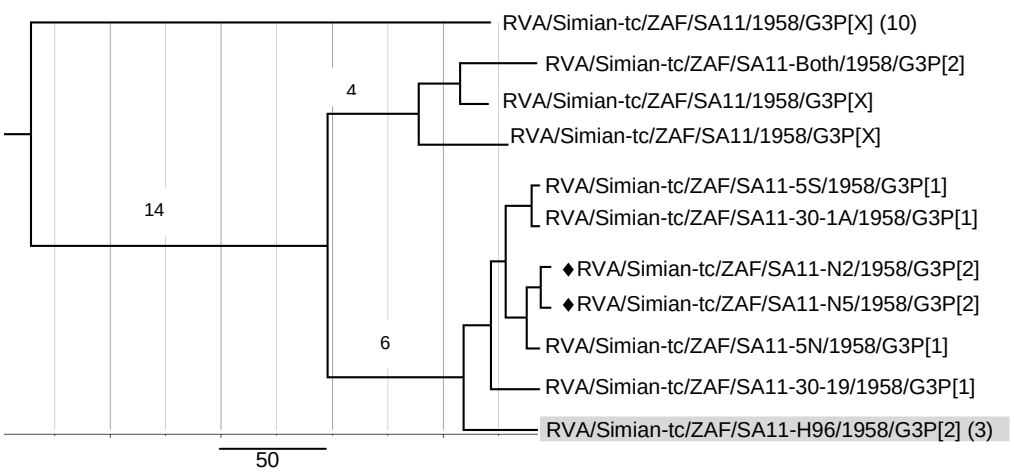
G. Genome segment 7 (NSP3)



H. Genome segment 8 (NSP2; N5 genotype)



I. Genome segment 9 (VP7)



J. Genome segment 11 (NSP5/6)

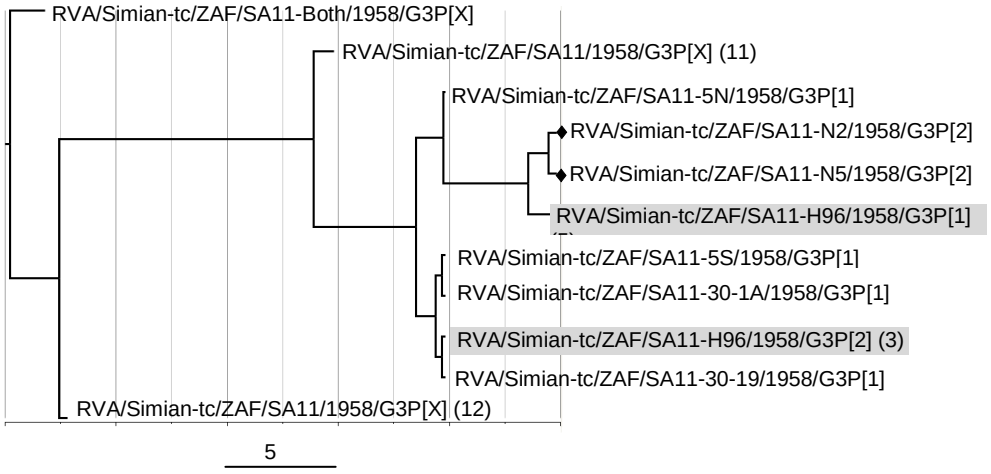


Figure 4.4. MCC trees constructed with Bayesian MCMC framework in BEAST software, to depict the molecular clock evolutionary relationships between SA11-N2, SA11-N5 and rotavirus SA11 sequences obtained from GenBank. Branch times (years) are indicated for some selected branches. SA11-N2 and SA11-N2 are indicated with a black diamond while SA11-H96, which is considered the prototype of the SA11 strains, is shaded grey. Two SA11-H96 sequences were available in GenBank for genome segments 5 (**E**), 6 (**F**), 7 (**G**), 8(**H**) and 11 (**J**). Sequences indicated with (3) where sequenced by Small and co-workers (Small *et al*., 2007) and those indicated with (5) were sequenced by Dutta and co-workers (2011). Accession numbers of sequences associated with unspecified strains are as follows: (1): AF015955; (2): DQ457016; (4): X16062; (6): AF290881; (7): AF290883; (8): X00421; (9): GU550506 (10): M87502; (11): AF306493 and (12): M28347.

The evolutionary rate obtained using genome segment 7 (NSP3) ORF sequences was 1.3 x 10^{-4} nucleotide substitutions/site/year with a coefficient of variation (CV) of 0.3. The substitution rate calculated for the SA11 rotavirus determined in this study is in accordance with the estimated average rate of 7.9 x 10^{-4} nucleotide substitutions/site/year in all genome segments except segments 1, 3 and 10 reported for a human group B rotavirus (Yang *et al*., 2004c). The MCC tree constructed, for genome segment 7 indicated that SA11-N2 and SA11-N5 were both closely related to strain SA11-H96 (Figure 4.4G).

The rate of evolution for genome segment 9 (VP7), was 2.9 x 10^{-5} nucleotide substitutions/site/year with a CV of 0.43. The frequency of appearance of monoclonal antibody resistant variants *in vitro* has been estimated to be approximately 10^{-5} (Taniguchi and Urasawa, 1995). The MCC tree generated, based on ORF nucleotide sequences for genome segment 9, showed that the SA11 strains clustered into two evolutionary groups (Figure 4.5I). SA11-N2 and SA11-N5 were closely related to SA11-5N and clustered with SA11-H96, SA11-5S, SA11-30-1A and SA11-30-19.

4.3 Discussion

The rotavirus SA11 strain is an important model for studying the biology of rotaviruses (Small *et al*., 2007). The whole genome consensus nucleotide sequence of rotavirus SA11 has not been determined before. Therefore, the aim of this study was to determine the consensus sequence of a rotavirus SA11 strain which was stored in a South African laboratory. Since the sample was received without any passage history, it was also important to determine the evolution of the strain, using molecular clock phylogenetics, in comparison to known rotavirus SA11 strains in the GenBank database.

The 5'-terminal end sequences obtained for all the genome segments of the rotavirus SA11 sequenced in this study were the expected sequences for group A rotaviruses (Tortorici *et al*., 2006). The 3'-terminal sequences obtained for genome segment 1 (VP1), 3–6 (VP3–VP6), 8 (NSP2) and 9 (VP7) were the typical 3'-terminal end sequences (Wentz *et al.*, 1996b, Wentz *et al*., 1996a). Genome segment 2 (VP2)

contained the sequence -U**A**UGACC-3' at the 3'-terminal end. This 3'-terminal end sequence was also described for genome segments 2 (VP2) and 10 (NSP4) of the rotavirus DS-1 strain (Matthijnssens *et al*., 2008a, Mlera *et al*., 2011). The atypical -UGUGA**A**CC-3' and UGUG**G**CC-3' 3'-terminal sequences obtained for genome segments 5 (NSP1) and 7 (NSP3) were identical to the 3'-terminal terminal sequences determined previously for these genome segments in SA11 or rotaviruses in general (Patton *et al.*, 2001, Small *et al*., 2007, Mitchell and Both, 1990, Mossel and Ramig, 2002). The atypical -UGUGA**A**CC-3' 3'-terminal sequence was found to reduce the efficiency of dsRNA synthesis and genome segment expression (Patton *et al*., 2001).

The occurrence of two distinct genome segments encoding NSP2 (Figure 4.2), with different genotypes, led to the conclusion that there were two rotaviruses in the virus sample. However, it was not technically possible to separate the other ten genome segments and assign them to a specific virus (SA11-N5 or SA11-N2) because the SeqMan Pro assembler in Lasergene® v8.1.2 (DNASTAR™) cannot assign individual genome segment contigs to a specific virus in the case of a mixed infection such as found in this sample. Distance matrices showed that the ten genome segments which could not be differentiated were very closely related to SA11-H96 sequenced by Small and co-workers (2007) (Appendix 1). Therefore, the virus identities could only be assigned based on the nucleotide sequence differences in genome segment 8. The genome segment 8 of SA11-N2 was identical to that of SA11-Both. The SA11-Both strain acquired its genome segment 8 from the bovine rotavirus O (Offal) agent as a result of reassortment (Small *et al*., 2007). The bovine rotavirus O agent was isolated from the sewage of an abattoir processing cattle and sheep (Malherbe and Strickland-Cholmley, 1967). This suggests that the rotavirus sample was contaminated with the bovine rotavirus O agent prior to, or during, propagation in cell culture. Therefore, the contamination resulted in reassortment between SA11 and the bovine rotavirus O agent. It is possible that reassortment could have involved other genome segments of the bovine rotavirus O agent. Since no other distinct genome segments were found and no passage history was available, any other reassortant could have been lost during passage in cell culture or was not detected. Furthermore, the detection of five novel amino acids in VP4, VP7 and NSP10 (Figure 4.3A, B and C) suggests either the occurrence of minor population variants or

sequence differences between SA11-N2 and SA11-N5 but their biological significance was not clear and should be investigated *in vivo* in follow-up studies after the two viruses have been separated by several rounds of plaque purification.

Alignment of genome segment 4 (VP4) nucleotide sequences indicated that the consensus nucleotide sequences of SA11-N2 and SA11-N5 were most closely related to the genome segment 4 of SA11-H96 (Figure 4.2). Therefore, the bovine rotavirus O agent was not the source of the genome segment 4 of SA11-N2 and SA11-N5 as is the case in some SA11 strains such as SA11-5S, SA11-5N and SA11-30-1A (Small *et al*., 2007). The likely effect of the novel 388R, observed in VP4, could be the altering of antigenicity such as conferring an antibody-escape phenotype, a suggestion which will need to be confirmed experimentally. For genome segment 5 (NSP1), the 99.7% sequence identity of SA11-N5 and SA11-N2 to SA11-H96 results in the segments grouping to a phylogenetic branch of which genome segment 5 (NSP1) SA11-H96 has been the sole representative to date. This confirms a previous observation that genome segment 5 does not seem to co-evolve with the other genome segments of rotavirus SA11 derivatives (Small *et al*., 2007).

Although there was a wide variation in the molecular clock substitution rates (~100-fold) between genome segments, the rates were within the expected ranges for dsRNA viruses ~1 x 10^{-4} – 1 x 10^{-6} (Duffy *et al*., 2008, Sanjuán *et al*., 2010, Barr and Fearns, 2010, Jenkins *et al*., 2002). However, the exact cause of such a variation could not be established. Interestingly, higher nucleotide substation rates of 1.36 x 10^{-3}–4.78 x 10^{-3} were recently reported for the genomes of group B rotaviruses (Lahon *et al*., 2012). The high rate of evolution of genome segment 1 (VP1) was unexpected since VP1 is a RNA-dependent RNA polymerase (RdRP) which is vital for ensuring viral replication. However, as with other RdRPs, sequence analyses of rotavirus genome segment 1 (VP1) revealed that region encoding the basic right hand structure motif of RdRPs is highly conserved and that the variations in the sequences are generally located close to the C- and N-terminal regions (Vasquez-del Carpio *et al*., 2006). A report describing an analysis of the evolution of the VP1 of ovine rotaviruses in comparison to other rotavirus group A strains also showed that amino acid variations mainly occurred in the N- and C-terminal regions of the RdRP (Chen *et al*., 2009b). The Ka/Ks ratio of non-synonymous to synonymous nucleotide

substitution obtained for the genome segment 1 ORF of 0.6 suggests synonymous substitutions and that genome segment 1 is under pressure to conserve the VP1 sequence

An apparent higher degree of divergence was observed following molecular clock evolutionary analysis in comparison to distance matrices (Figure 4.4; Appendix 1). This difference is attributed to the different approaches of the two methods for reconstructing phylogenetic trees. The divergences also highlight sequence differences which could have been introduced during cell culture in the many laboratories that sequenced the rotavirus SA11 genome segments. However, Bayesian phlyogenetic analyses in MCMC are more statistical (probabilistic) and samples trees in proportion to their likelihood thereby producing the most credible or consensus tree (Drummond *et al*., 2005, Drummond and Rambaut, 2007). On the other hand, distance matrix calculations do not assume a molecular clock in the analysis (Felsenstein, 1988).

Few molecular clock studies involving rotaviruses have been reported. In one of these few reports, genome segment 4 was found to evolve at a high rate of 0.58 x 10^{-3} nucleotide substitutions/site/year (Jenkins *et al*., 2002). A lower rate of 1.4 x 10^{-5} nucleotide substitutions/site/year was obtained in this study for cell culture-adapted SA11 strains. For genome segment 9, the evolutionary rate obtained in this study of 2.9 x 10^{-5} substitutions/site/year was lower than the rate estimated for human G9 rotavirus strains (Matthijnssens *et al*., 2010a). Rates of 1.87 x 10^{-3} substitutions/site/year for G9 (VP7) rotaviruses, and 1.66 x 10^{-3} substitutions/site/year for G12 (VP7) rotaviruses were determined (Matthijnssens *et al*., 2010a). Sequences analysed in the reports of Jenkins and co-workers (2002) as well as Matthijnssens and co-workers (2010a) were obtained from wild-type field strains. Matthijnssens and co-workers attributed the high evolutionary rate they observed in G9 and G12 rotaviruses to the immunological pressure on genome segment 9 (Flores *et al*., 1988, Taniguchi and Urasawa, 1995, Matthijnssens *et al*., 2010a). The same conclusion can also be drawn for the high rate observed for genome segment 4 by Jenkins and co-workers.

Genome segment 3 (VP3) of SA11-5S was closely related to genome segment 3 of SA11-N2 and SA11-N5 (Figure 4.4C). SA11-5S was isolated from limiting dilutions following 26 passages of SA11-4F in MA104 cells (Patton *et al*., 2001). However, the SA11-4F strain followed a separate and distinct evolutionary path from that of SA11-5S. No precise passage history was available for strains SA11-4F from which SA11-5N was obtained. The SA11-FEM strain has a genome segment 4 that was closely related to the consensus genome segment 4 sequences of SA11-N2 and SA11-N5 (Figure 4.4D). No passage history is available for SA11-5N except that the strain was a plaque isolate from a stock of strain SA11-4F (Dr John T. Patton, personal communication). SA11-30-1A and SA11-30-19 were isolated by triple-plaque purification after 30 passages of SA11-FEM in MA104 cells (Patton *et al*., 2001). The genome segment 7 nucleotide sequences of SA11-N2 and SA11-N5 were closely related to that of SA11-H96.

To conclude, sequence-independent genome amplification and 454® pyrosequencing revealed the presence of two distinct genome segment 8 nucleotide sequences in the same sample. One of the two genome segment 8 sequences was genotype N2 and the other N5. The other ten sequences could not be separated due to limitations in the assembling module of Lasergene®. Therefore, the sample was a mixture of two rotaviruses which were named SA11-N2 and SA11-N5. SA11-N2 acquired a bovine rotavirus genome segment 8 (NSP2), due to reassortment with the bovine rotavirus O agent in a similar manner as SA11-Both. Although SA11-N2 obtained genome segment 8 by reassortment like SA11-Both, distance matrices showed that only genome segment 8 (NSP2) and 11 (NSP5/6) of SA11-N2 and SA11-Both were identical. Distance matrices also showed that genome segment 11 (NSP5/6) of SA11-N2 was identical to that of SA11-H96 sequenced by Small and co-workers (2007). Therefore, SA11-N2 is not the same as SA11-Both. Furthermore, based on evolutionary divergences observed in distance matrices, it is concluded that SA11-N2 and SA11-N5 are close derivatives of the SA11-H96 strain. However, pyrosequencing the SA11-H96 strain of Small and co-workers (2007) could be useful in understanding its relatedness to the SA11 strains sequenced in this study. Traditional Sanger sequencing could have only detected the presence of one of these viruses. Therefore, the consensus whole genome sequences of two rotaviruses in a sample that was stored in South Africa were determined. Genome

segment 4 (VP4) sequences of SA11-N2 and SA11-N5 were not a result of reassortment with a bovine rotavirus O agent as in other rotavirus SA11variants. Although rotavirus SA11 strains are very closely related, MCCs indicated that genome segments 5 (NSP1), 6 (VP6) and 11 (NSP5/6) of SA11-N2 and SA11-N5 followed the evolutionary path similar to that of SA11-H96 sequenced by Dutta and co-workers (2011). The genome segments 1 (VP1), 2 (VP2) and 8 (NSP2 of SA11-N5; JQ688680) followed an evolutionary path similar to that of SA11-H96 sequenced by Small and co-workers (2007). This suggests a relative instability of the rotavirus SA11 genome. However, inter-laboratory variations in cell lines and culture conditions used for propagating SA11-H96 could also contribute to apparent evolutionary path differences. Finally, a well-characterised rotavirus SA11 now became available for the generation of wild-type transcripts for use in the attempt to recover rotavirus by reverse genetics.