CHAPTER 3: CONSENSUS SEQUENCE DETERMINATION AND ELUCIDATION OF THE EVOLUTIONARY HISTRORY OF A ROTAVIRUS Wa VARIANT REVEAL A CLOSE RELATIONSHIP TO VARIOUS Wa VARIANTS DERIVED FROM THE ORGINAL Wa STRAIN

3

3.1 Introduction

Nucleotide sequence determination has undergone a radical transformation over the last four decades, since the modest beginnings of chain terminating dideoxynucleotide sequencing of DNA in the 1970s. With drastic advances in computational hardware and software, bioinformatics and online databases, next-generation sequencers are set apart from the conventional capillary-based sequencing platforms. Modern next-generation sequencing technology grants researchers the capacity to generate and process massive amounts of sequence reads in parallel on one instrument.

Whole-genome analyses of human rotavirus strains are fundamental in studying evolutionary patterns and genetic affiliations to other strains (Ghosh and Kobayashi, 2011). Matthijnssens and co-workers suggested a novel classification system based on the whole genome sequence of all 11 rotavirus genome segments in order to obtain a more complete picture of rotavirus strain diversity (Matthijnssens et al., 2008b). Nowadays, whole genome characterization has become the sought after procedure for viral strain characterization as next generation sequencing technology becomes more widely available and affordable. The easily accessible public sequence databases contain massive amounts of sequencing data, facilitating complex analysis and strain comparisons.

The most prevalent rotavirus A strains found in humans are the genotypes G1, G2, G3, G4, G9 and G12 in combination with P[4], P[6] and P[8] (Heiman et al., 2008, Matthijnssens et al., 2010). Group A rotaviruses include the AU-1 (G3P[8]), DS-1 (G2P[4]) and Wa (G1P[8]) genogroups. The human rotavirus type A Wa strain is the prototype of rotavirus strains in the Wa-like genogroup (Heiman et al., 2008). The Wa strain (Rotavirus A strain Human-tc/USA/Wa/1974/G1P[8]) was originally isolated in the United States in 1974 from an infant with severe diarrhoea. It was also one of the first rotaviruses to be successfully adapted to cultured cells (Wyatt et al., 1980), making the Wa strain one of the best-studied human

rotaviruses to date. The Wa reference strain used today is a composite sequence of genome segments of various Wa strains and the genome segment sequences do not all originate from a single virus (Heiman et al., 2008).

This chapter describes the consensus sequence, obtained by sequence-independent genome amplification and next generation 454® pyrosequencing, of a rotavirus Wa strain (generously supplied by Dr. Carl Kirkwood from the Murdoch Children's Research Institute). The rotavirus Wa consensus strain originated from the original 1974 rotavirus Wa isolate, but the exact passage history is unknown. The evolutionary history of this strain was investigated through phylogenetic and molecular clock analyses combined with nucleotide substitution rate and evolutionary pressures analyses.

3.2 Materials and Methods

3.2.1 Rotavirus and cell culture propagation

A cell culture adapted rotavirus Wa sample was obtained from Dr. Carl Kirkwood at the *Murdoch Children's Research Institute* (MCRI), Melbourne, Australia. This strain was originally obtained by Dr. Ruth Bishop from Dr. Richard Wyatt (National Institutes of Health, USA) in 1983. This particular Wa strain is a cell culture adapted variant from the original 1974 isolate but the exact passaging history is unknown (Dr. Carl Kirkwood and Dr. Ruth Bishop, personal communication). At MCRI the strain was passaged 9 times in MA104 cells.

Following activation with 10 µg/ml porcine trypsin IX (Sigma) at 37°C for 30 minutes, the virus was passaged a further 7 times in African green monkey cells (MA104) at the North-West University (NWU), South Africa. The cells were cultured in serum free Dulbecco's modified essential medium (D-MEM; Hyclone) containing 1 µg/ml porcine trypsin (1x), 1% penicillin/streptomycin/amphotericin B (Gibco) and 1% non-essential amino acids (Lonza). Cells were cultured at 37 °C in a humidified atmosphere containing 5% CO_2.

3.2.2 Sequence-independent cDNA synthesis and genome amplification

Rotavirus double-stranded RNA (dsRNA) was isolated as described by Potgieter and co-workers (Potgieter et al., 2009). Infected cells were harvested when about 70% cytopathic

effect was reached by freeze-thawing the cell/virus suspension twice. A phenol-chloroform extraction was performed using the Trizol reagent (Invitrogen) and the single-stranded RNA was removed by precipitation with 2 M LiCl (Sigma) at 4°C for 14 h. Subsequently, the solution was centrifuged at 16 000 × g for 30 min at 4°C and the supernatant was purified using the MinElute kit (Qiagen) as described by the manufacturer. A PC3-T7 loop primer (5'p-GGATCCCGGGAATTCGGTAATACGACTCACTATATTTTTATAGTGAGTCGTATTA-OH3') (TibMolBiol) was ligated to the purified RNA and the genome was subsequently amplified as cDNA using the sequence-independent genome amplification technique free from cloning bias (Potgieter et al., 2009) with slight modifications. The purified ligated dsRNA was denatured using 300 mM methyl mercury hydroxide (Alfa Aesar). The cDNA was synthesised using AMV reverse transcriptase (Fermentas) followed by amplification of the genome with Phusion High Fidelity DNA polymerase (Finnzymes). The QIAquick (Qiagen) PCR purification kit was employed in order to purify the amplified cDNA according to the manufacturer's instructions. This rotavirus Wa-amplicon cocktail was sequenced using 454® pyrosequencing technology (GS FLX Titanium, Roche) at Inqaba Biotec (South Africa) as described before (Jere et al., 2011).

3.2.3 Sequence and phylogenetic analyses

The Lasergene™ 8.1.2 suite (DNASTAR®) was used for sequence assembly. The consensus sequence (CS) of all 11 genome segments was determined using the SeqMan module of this software suite. The nucleotide and deduced protein sequences were analysed with the **Basic Local Alignment Search Tool** (**BLAST**) and compared with Wa sequences available in GenBank. Sequences of the 11 genome segments of all rotavirus strains (***Table 3.1***) that closest resembled the WaCS were retrieved from GenBank and aligned using MEGA 5.1. The evolutionary history was determined using the Neighbour-Joining method (Saitou and Nei, 1987) conducted in MEGA 5.1 (Tamura et al., 2011) with a bootstrap value of 10 000. In order to obtain a more comprehensive phylogenetic overview, the prototype rotavirus DS-1, AU-1 and D reference strains were also included. The evolutionary distances were computed using the Maximum Composite Likelihood method (MEGA 5.1) and are in the units of the number of base substitutions per site (Tamura et al., 2004). Codon positions included were 1^{st} + 2^{nd} + 3^{rd} + noncoding. All positions containing gaps and missing data were eliminated.

Table 3.1: GenBank accession numbers of rotavirus strains used in phylogenetic analysis and pairwise comparisons.

Type	Rotavirus strain	GenBank accession numbers of different rotavirus genome segments										
		GS1 (VP1)	**GS2 (VP2)**	**GS3 (VP3)**	**GS4 (VP4)**	**GS5 (NSP1)**	**GS6 (VP6)**	**GS7 (NSP3)**	**GS8 (NSP2)**	**GS9 (VP7)**	**GS10 (NSP4)**	**GS11 (NSP5/6)**
Wa-like	RVA/Human-tc/USA/**WaCS**/1974/G1P1A[8]	DQ490539	X14942	AY267335	L20877.1	L18943	K02086	X81434	L04534	M21843	AF093199	AF306494
	RVA/Human-tc/USA/**D**/1974/G1P1A[8]	EF583021	EF583022	EF583023	EF672570	EF672571	EF583024	EF672572	EF672573	EF672574	EF672575	EF672576
	VirWa G1P[8]	FJ423113	FJ423114	FJ423115	FJ423116	FJ423117	FJ423118	FJ423119	FJ423120	FJ423121	FJ423122	FJ423123
	Wag7/8re G1P[8]	FJ423135	FJ423136	FJ423137	FJ423138	FJ423139	FJ423140	FJ423141.	FJ423142	FJ423143	FJ423144	FJ423145
	ParWa G1P[8]	FJ423124	FJ423125	FJ423126	FJ423127	FJ423128	FJ423129	FJ423130	FJ423131	FJ423132	FJ423133	FJ423134
	Wag5re G1P[8]	FJ423146	FJ423147	FJ423148	FJ423149	FJ423156	FJ423150	FJ423151	FJ423152	FJ423153	FJ423154	FJ423155
	RVA human/Bethesda/**DC5115**/1977/G4P[8]	HM773942	HM773943	HM773944	HM773945	HM773946	HM773947	HM773948	HM773949	HM773950	HM773951	HM773952
	RVA human/Bethesda/**DC2239**/1976/G3P[8]	FJ947859	FJ947860	FJ947861	FJ947862	FJ947863	FJ947864	FJ947865	FJ947866	FJ947867	FJ947868	FJ947869
	RVA/Human-wt/BGD/**Dhaka16**/2003/G1P[8]	DQ492669	DQ492670	DQ492671	DQ492672	DQ492675	DQ492673	DQ492677	DQ492676	DQ492674	DQ492678	DQ492679
	RVA/Human-tc/BRA/**IAL28**/1992/G5P[8]	EF583029	EF583030	EF583031	EF672584	EF672585	EF583032	EF672586	EF672587	EF672588	EF672589	EF672590
	RVA/Vaccine/USA/**RotaTeq-WI79-9**/1992/G1P7[5]	GU565052	GU565053	GU565054	GU565055	GU565058	GU565056	GU565060	GU565059	GU565057	GU565061	GU565062
	RVA/Human-tc/USA/**WI61**/1983/G9P1A[8]	EF583049	EF583050	EF583051	EF672619	EF672620	EF583052	EF672621	EF672622	EF672623	EF672624	EF672625
	RVA/Human-tc/GBR/**ST3**/1975/G4P2A[6]	EF583045	EF583046	EF583047	EF672612	EF672613	EF583048	EF672614	EF672615	EF672616	EF672617	EF672618
	KU G1P[8]	AB022765	AB022766	AB022767	AB222784	AB022769	AB022768	AB022771	AB022770	D16343	AB022772	AB022773
DS-1-like	RVA/Human-tc/USA/**DS-1**/1976/G2P1B[4]	HQ650116	HQ650117	HQ650118	HQ650119	HQ650120	HQ650121	HQ650122	HQ650123	HQ650124	HQ650125	HQ650126
AU-1-like	RVA/Human-tc/JPN/**AU-1**/1982/G3P3[9]	DQ490533	DQ490536	DQ490537	D10970	D45244	DQ490538	DQ490535	DQ490534	D86271	D89873	AB008656

3.2.4 Molecular clock analyses and evolutionary rate estimations

Bayesian Evolutionary Analysis Sampling Trees (BEAST) is a multifaceted evolutionary package for phylogenetic and population genetics analysis. Bayesian phylogenetic reconstructions were performed using the Markov chain Monte Carlo (MCMC) analysis contained in the BEAST software suite (1.6.2) (Drummond and Rambaut, 2007). Aligned rotavirus sequences were converted to the NEXUS format using Data Analysis in Molecular Biology Evolution (DAMBE) software 5.2.76 (http://dambe.bio.uottawa.ca/dambe). JModelTest (http://darwin.uvigo.es/software/software.html) was used to determine the most suitable nucleotide substitution model. Subsequently all strains were analysed using a HKY model with gamma distributed rate variation and a relaxed clock lognormal model with a flexible Bayesian skyline tree prior. One hundred million MCMC simulations were performed (Matthijnssens et al., 2010). Tree files of all 11 genome segments were generated and annotated with TreeAnnotator. Additionally, all 11 tree files were combined with LogCombiner 1.6.2, in order to produce a tree representing the entire genome of the rotaviruses examined. Trees were visualized by FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

Evolutionary rates were estimated for all 11 genome segments of the closely related rotavirus variants in the PAML 4.5 software package (Yang, 1997) using codon substitution models with a single non-synonymous/synonymous substitution rate (dN/dS). To elucidate general evolutionary pressures acting on protein-coding regions, non-synonymous–synonymous substitution ratios (ω) were also employed (Yang and Nielsen, 2002, Yang et al., 2000) using the PAML 4.5 software. In order to identify specific codons under diversifying conditions, three different codon-based maximum likelihood methods, SLAC, FEL and REL were utilized to estimate the dN/dS. All stop codons were removed from sequences using the CleanStopCodons function of the HyPhy 2.1.2 software package (Kosakovsky Pond et al., 2005) and analysed with the online phylogenetic analysis tool Datamonkey (Delport et al., 2010).

3.3 Results and Discussion

3.3.1 Sequence data analysis and comparison to similar rotavirus strains in GenBank

In this study, the consensus nucleotide sequence of a cell culture adapted rotavirus Wa strain, obtained from MCRI was determined. The Wa reference strain used currently is a composite sequence of various Wa strains and the genome segment sequences do not all originate from a single virus (Heiman et al., 2008). The consensus sequence of rotavirus Wa was obtained by sequence-independent genome amplification and next generation 454® pyrosequencing (GS FLX Titanium, Roche). A total amount of 9.57MB (30 507 reads) of data was generated, of approximately 400 bp per read. The complete consensus sequence for each of the 11 Wa rotavirus genome segments was attained using the Lasergene™ 8.1.2 SeqMan Pro suite (DNASTAR®). The total size of the consensus genome was 18 502 bp. Coverage of the genome segments ranged from 134-fold (for VP1) to 652-fold (for NSP2), with a 301-fold average depth of coverage (***Table 3.2***). The full genome constellation of G1-P[8]-I1-R1-C1-M1-A1N1-T1-E1-H1 was confirmed with the classification tool, RotaC (Maes et al., 2009) and was designated RVA/Human-tc/USA/WaCS/1974/G1P[8]. Eight genome segments, 1 (VP1), 2 (VP2), 3 (VP3), 4 (VP4), 6 (VP6), 5 (NSP1), 8 (NSP2) and 11 (NSP5/6), did not have any novel nucleotide changes compared to any rotavirus sequences in GenBank. A total of 4 novel nucleotide changes, which also resulted in amino acid changes, were detected in genome segment 7 (NSP3), genome segment 9 (VP7) and genome segment 10 (NSP4) (***Table 3.1 and Figure 3.1***).

Table 3.2: Summary of the WaCS data determined with 454® pyrosequencing also indicating the nature and position of novel nucleotide and amino acid changes

WaCS Genome Segment (encoded protein)	GenBank accession number	Length (bp)	Length (and position) of ORF	Average sequence coverage	Percentage similarity to Reference Wa strain[a] (Accession #)	Nature (Position) of novel[b] nucleic acid changes	Nature (position) of novel[b] amino acid changes	Protein region in which amino acid change occurred
Segment 1 **(VP1)**	JX406747	3302	3267 (19-3285)	138	100 (DQ490539)	*No novel changes*	*No changes*	-
Segment 2 **(VP2)**	JX406748	2717	2673 (17-2689)	209	100 (X14942)	*No novel changes*	*No changes*	-
Segment 3 **(VP3)**	JX406749	2591	2508 (50-2557)	438	100 (AY267335)	*No novel changes*	*No changes*	-
Segment 4 **(VP4)**	JX406750	2360	2328 (10-2338)	224	99.7 (L20877.1)	*No novel changes*	*No changes*	-
						No novel changes	*No changes*	-
Segment 5 **(NSP1)**	JX406751	1567	1460 (32 -1492)	158	99.7 (L18943)	*No novel changes*	*No changes*	- -
Segment 6 **(VP6)**	JX406752	1356	1194 (24-1217)	389	99.9 (K02086)	*No novel changes*	*No changes*	-
Segment 7 **(NSP3)**	JX406753	1059	933 (35-967)	165	98.6 (X81434)	**G to A (618)**	**Methionine to Isoleucine (206)**	Dimerization and interaction with ZC3H7B
Segment 8 **(NSP2)**	JX406754	1059	954 (47-1000)	652	100 (L04534)	*No novel changes*	*No changes*	-
Segment 9 **(VP7)**	JX406755	1062	981 (49-1029)	325	99.8 (M21843)	**T to C (378)**	**Tyrosine to Histidine (117)**	Part of a beta strand of the outer capsid glycoprotein VP7
Segment 10 **(NSP4)**	JX406756	750	528 (42-569)	325	99.7 (AF093199)	**C to T (141)**	**Leucine to Serine (34)**	H2 transmembrane domain
						C to T (154)	**Serine to Phenylalanine (38)**	H2 transmembrane domain
Segment 11 **(NSP5/6)**	JX406757	664	593 (22-615) [NSP5] 278 (80-358) [NSP6]	272	99.8 (AF306494)	*No novel changes*	*No changes*	-

[a] Wa reference strain is a composite sequence of genome segments of various Wa strains and the genome segment sequences do not all originate from a single virus (Heiman et al., 2008).

[b] Nucleotide or amino acid changes are seen as novel if they only occur in the WaCS strain in comparison to other rotavirus sequences in GenBank

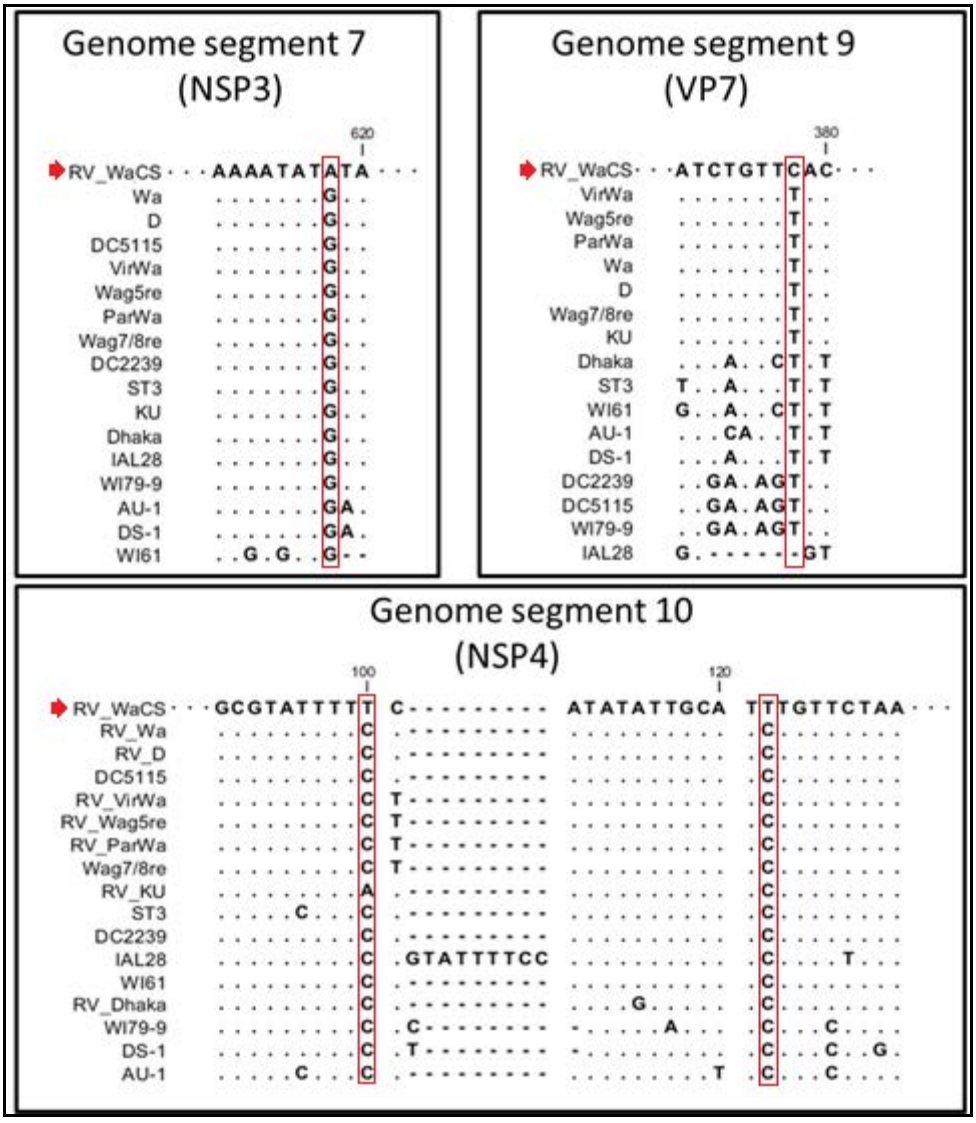

Figure 3.1: ***Nucleotide alignments of the WaCS and closely related Wa variants and rotavirus reference strains.*** *Alignments indicate the novel nucleotide changes detected in genome segments 7 (618:G to A), genome segment 9 (378: T to C) and genome segment 10 (141: C to T and 154: C to T).*

In genome segment 7 (NSP3), there was one new nucleotide change at position 618 (G to A). This resulted in an amino acid change (206: Met to Ile) in the central domain of NSP3, which is thought to form a coiled-coil structure permitting dimerization and can also interact with the zinc finger protein ZC3H7eB (Vitour et al., 2004). Upon closer examination of the alignment data, Met^{206} seems to be conserved among all rotaviruses examined (***Figure 3.2***). The NSP3 secondary structure of the WaCS and Wa reference strain was examined using the

online protein sequence analysis workbench, PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/). Amino acid position 206 in the secondary structure of both the WaCS and the Wa reference strain were predicted to be part of an alpha helix structure (***Figure 3.2 (A)***).

In genome segment 9 (VP7) only one new point mutation (378: T to C) was identified, resulting in a single amino acid change (117: Tyr to His) which is in a beta plate (Chen et al., 2009). Again, Tyr^{117} seem to be well conserved in all rotavirus strains examined. Secondary structure comparison between WaCS and the Wa reference strain indicate that both variations will form part of a beta plate (***Figure 3.2 (B)***).

In genome segment 10 (NSP4) 2 unique nucleotide changes were detected (141: C to T and 154: C to T), both resulting in amino acid changes. The two amino acid changes (34: Leu to Pro and 38: Ser to Phe) were detected in the helical transmembrane domain, H2, a region which is embedded in the endoplasmic membrane, anchoring NSP4. Amino acid position 34 does not seem to be that well conserved in the rotavirus strains examined. On the other hand, amino acid Ser^{38} is well-preserved in all strains. Structural analysis of the two variations predicted the same secondary amino acid structure and both alterations (34: Leu to Pro and 38: Ser to Phe) still formed part of an alpha helix (***Figure 3.2 (C)***).

In genome segment 4, two noteworthy nucleotide changes were also observed which were only present in one other Wa strain, the rotavirus Wag7/8re variant (***Figure 3.2 (D)***). The VP8* subunit of the WaCS strain and the Wag7/8re variant exhibited one unique nucleotide change (151: G to T), resulting in an amino acid change (51: Gly to Val). This amino acid modification occurred near the sialic acid binding domain in an amino acid region with little structural importance or known function (Kraschnefski et al., 2009) .

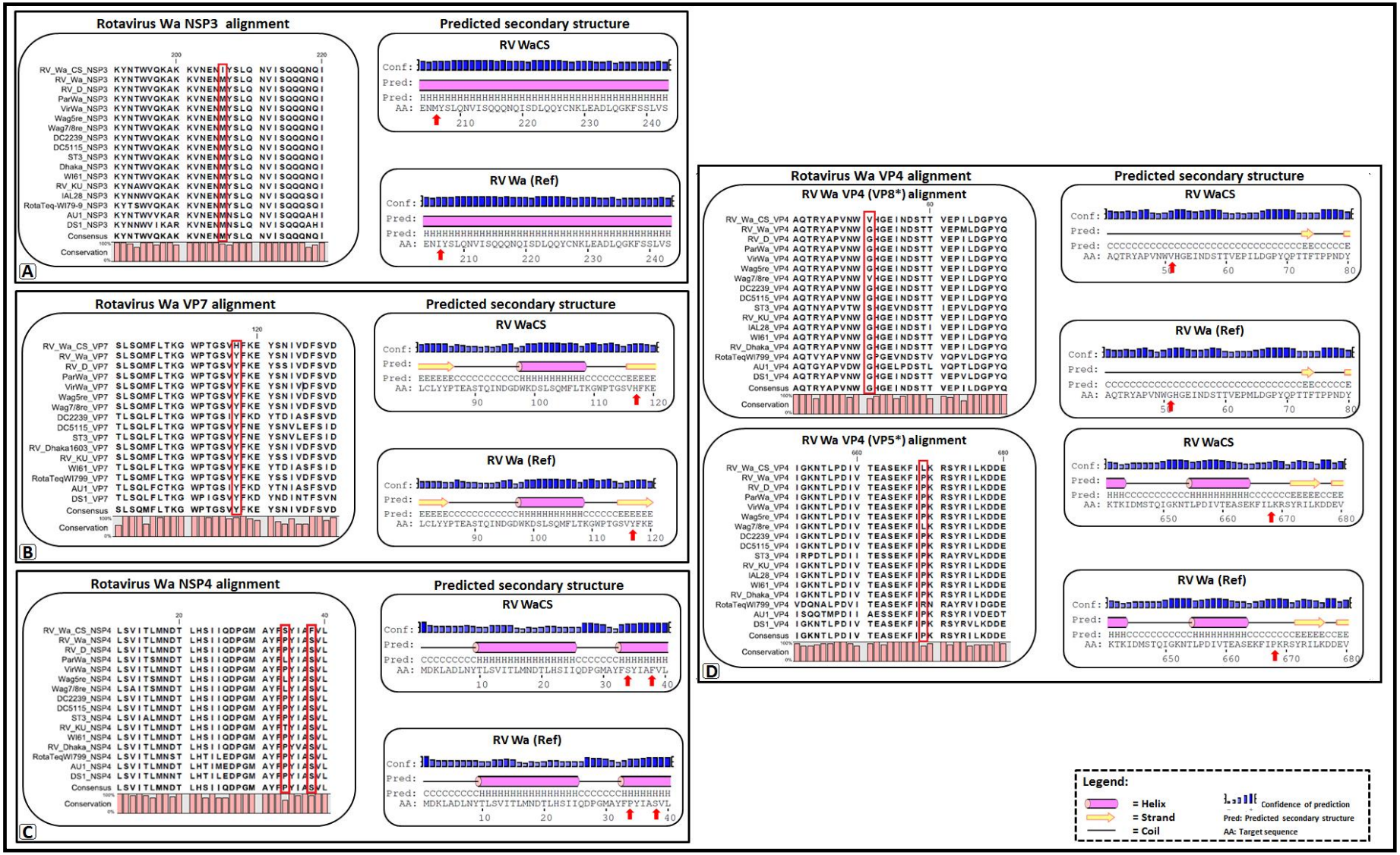

Figure 3.2: Amino acid alignments and predicted secondary structure of viral proteins exhibiting novel or noteworthy amino acid changes. *The secondary structure was predicted for NSP3 (A), VP7 (B), NSP4 (C) and VP4 (D) of the WaCS and Wa reference strain. The relevant amino acids regions containing the amino acids of interest are indicated in a red block on the alignments and by red arrows on the predicted secondary structure. Secondary protein structures predicted by the PSIPRED online protein sequence analysis workbench (http://bioinf.cs.ucl.ac.uk/psipred/).*

It must be noted that in the rhesus rotavirus strain, amino acids residues in the same general area (nucleotide position 148-150) of genome segment 4 have been known to poses limited cross-reactive neutralization (Mackow et al., 1988). The amino acid alignments indicate that Gly^{51} is conserved among the examined strains and that only WaCS and Wag7/8re did not concede to the norm. The predicted secondary structures of both WaCS and Wag7/8re did not differ from that of the Wa reference strain.

The other nucleotide change in genome segment 4 was observed in VP5* (2011: C to T). This modification results in an amino acid change (668Pro to Leu) which is located just outside the antigen domain and, as with the amino acid change in VP8*, is found in an area that does not seem to have an important influence on protein structure. No nucleotide insertions, deletions or rearrangements were detected in the WaCS strain. From *in silico* analysis, none of the detected novel nucleotide changes, and subsequent amino acid changes, is expected to have any substantial effect on the structure, or known function, of the effected viral proteins.

The evolutionary history of the WaCS strain was investigated using the Neighbour-Joining method conducted in MEGA 5.1. Rotavirus strain sequences that closely resemble the WaCS were acquired from GenBank and used for phylogenetic analysis. The reference strains AU-1, DS-1 and D were also included in the phylogenetic analysis. It was evident from phylogenetic analysis of the individual genome segments, that the RVA/Human-tc/USA/WaCS/1974/G1P[8] strain is closely related to the human rotavirus A strains VirWa, ParWa, Wag5re and Wag7/8re (***Appendix A, Figure A.1***).

From the rotavirus WaCS amino acid alignments (results not shown) and pairwise comparison data (***Appendix A, Table A.1***) of the closest Wa strains identified by BLAST, viral proteins VP1, VP2, VP6 and NSP2 are the most conserved, only eliciting one or less amino acid site change in the six closest related Wa variants (VirWa, ParWa, Wag5re, Wag7/8re, Wa and D) examined. When taking into account all rotavirus strains examined, including the reference strains AU-1 and DS-1, a similar picture emerges. Over 85% of the amino acid sites of the structural proteins VP1, VP2 and VP6 are conserved over the broad range of rotaviruses studied. On the other hand, VP7 appears to be the viral protein that is the least

conserved, with only about 1% of the amino acid sites preserved in all rotavirus strains examined. It is interesting to note that the KU strain, which has a G1P[8] genotype similar to the WaCS, exhibited little phylogenetic similarities to the Wa genotype. Genome segments 1 (VP1), 2 (VP2), 3 (VP3) and 4 (VP4) of the KU strain's pairwise comparison analysis indicated major amino acid differences in comparison to the other Wa strains. The rotavirus KU strain was in Japan at approximately the same time as the original Wa strain (USA). It would be interesting to further analyse the divergence between these two variants.

3.3.2 Description of rotavirus Wa variants and molecular clock analyses

It was evident from phylogenetic analysis (***Appendix A, Supplementary Figure A.1***) and pairwise comparisons (***Appendix A, Supplementary Table A.1***) of the individual genome segments, that the rotavirus WaCS strain is closely related to the human rotavirus A strains VirWa, ParWa, Wag5re and Wag7/8re. Although nucleotide and amino acid sequences of these rotavirus Wa variants were available on GenBank, no publications could be found on the origins or passage history of these rotaviruses. In order to determine the passage history of our rotavirus Wa strain, it was important to know the passage history of these closely related Wa variants. We contacted Dr. Lijuan Yuan (Virginia Tech, Blacksburg, Virginia, USA) and she kindly provided detailed passaging information on the ParWa, VirWa, Wag5re, Wag7/8re and the unpublished AttWa variants. From her information it was clear that these related rotavirus Wa variants were derived from the Wa human rotavirus strain (HRV), originally isolated from an infant with diarrhoea, obtained by Dr. R. G. Wyatt (NIH, Bethesda, Maryland) and passaged in gnotobiotic pigs (Wyatt et al., 1980) (***Figure 3.3***). Double-stranded RNA was isolated from a pooled intestinal sample of the 21st passage of virulent Wa variant (VirWa), and sequenced. After 11 passages of the original Wa HRV in gnotobiotic pigs, the virus was adapted to MA104 cells. The cell culture adapted Wa HRV was plaque purified and subsequently cloned 6 times in MA104 cells by limiting dilution in the laboratory of Dr. Wyatt.

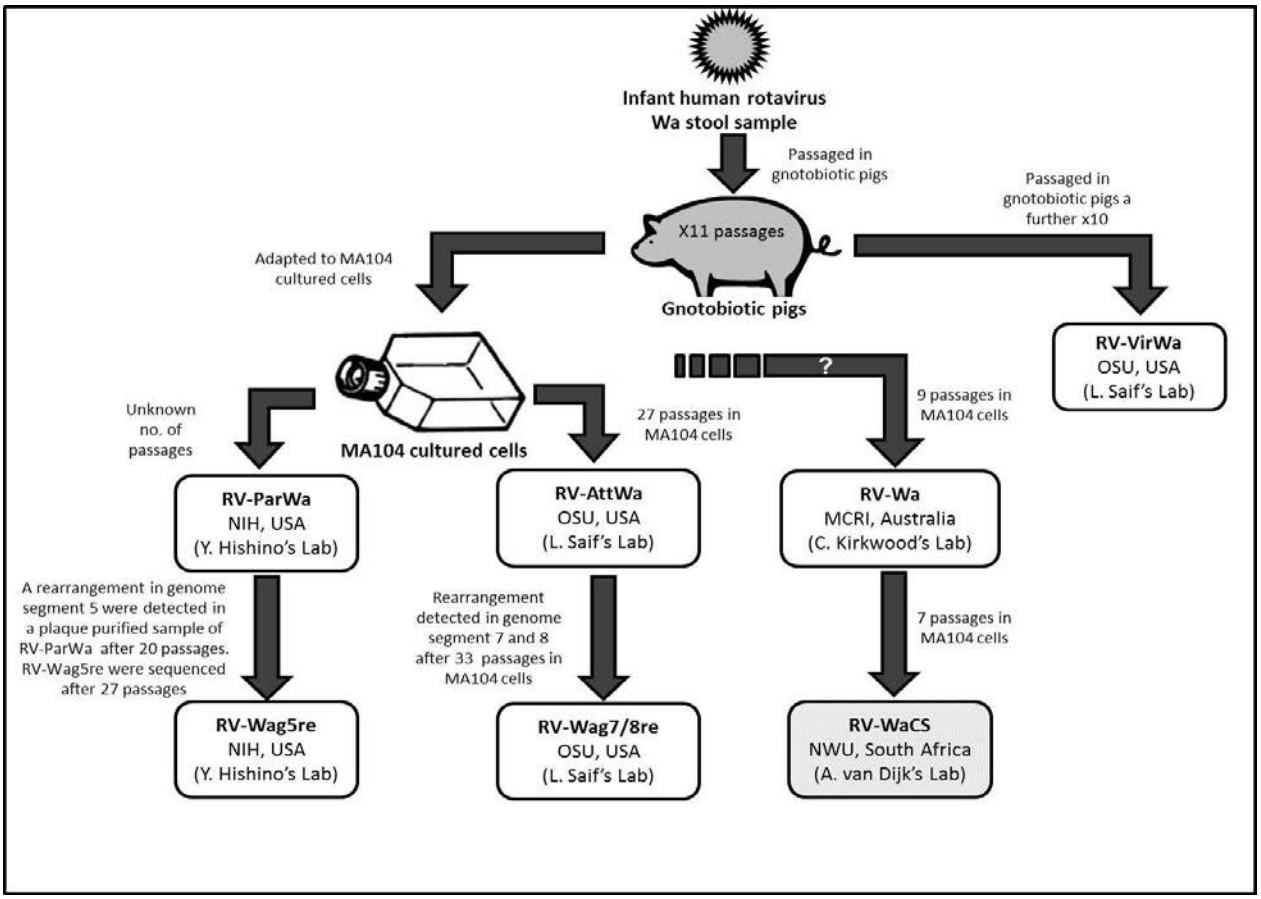

Figure 3.3: ***Schematic diagram summarising the known passage history of the rotavirus Wa variants originating from the 1974 infant human rotavirus Wa stool sample isolate.***

The following information on the closely related variants were kindly provided by Dr. Lijuan Yuan*: Resulting from the identification of an attenuated version of the cell culture adapted Wa HRV, AttWaHRV, this virus was further passaged continuously in MA104 cells in the laboratory of Dr. Linda Saif at Ohio State University. After 27 passages in MA104 cells, the electropherotype of the AttWaHRV was still indistinguishable from the initial Wa HRV isolate (see Supplementary Figure B1 for the electropherotype patterns of the different rotavirus Wa variants). Rearrangements in genome segments 7 and 8 were detected after the 33rd passage and became dominant at the 35th passage. This variant was designated Wag7/8re. At NIH, Dr. Y. Hoshino serially plaque purified a cell culture adapted Wa HRV strain with an unknown passage history. The 33rd purified plaque was purified once more (thus 33-1) to produce the parent virus (ParWa) which was passaged serially in MA104 cells. One of the plaques was passaged further in MA104 cells at high MOI. At passage 20, a rearrangement in genome segment 5 (NSP1) became evident with PAGE analysis and at passage 24 this*

mutant (designated as Wag5re) became dominant. This variant was plaque purified three times by Dr. Y. Hoshino and was sequenced at the 27th passage (see ***Appendix B*** for electropherotype patterns of the different rotavirus Wa variants**.**).

Phylogenetic analysis indicated, furthermore, that the WaCS is also related to the rotavirus Wa- and D reference strains (Heiman et al., 2008) (***Appendix A, Figure A.1***). Since no clear ancestor of the WaCS could be identified with phylogenetic analysis, the Markov chain Monte Carlo (MCMC) molecular clock models were employed to estimate rooted phylogenetic in the Bayesian Evolutionary Analysis Sampling Trees (BEAST) software. BEAST is a popular and multifaceted evolutionary package for phylogenetic and population genetics analysis. The MCMC in the Bayesian skyline plot model estimates the distribution of a specific population size through time straight from a gene sequences using a specified nucleotide-substitution model. A feature that set Bayesian skyline plots, aside from other phylogenetic analysis, is its inclusion of credibility intervals for the estimated population size at every single point in time, all the way to the most recent common ancestor of particular gene sequences. Afford mentioned credibility intervals epitomise both phylogenetic and coalescent uncertainty (Drummond et al., 2005). All 11 genome segments were individually analysed and combined to produce a universal phylogenetic tree (***Figure 3.4***). This analysis confirmed that the consensus sequence is indeed evolutionarily related to the human rotavirus A strains Wag5re, Wag7/8re, ParWa and VirWa (see ***Appendix A, Figure A.2 for*** MCC trees of all 11 genome segments). Although these analyses confirmed the close genetic relationship between the rotavirus WaCS and two rotavirus Wa variants (ParWa and VirWa), as indicated by the phylogenetic investigation, no clear-cut ancestral strain could be identified between all 11 genome segments. As indicated by Dr. Yuan, nucleotide insertions are present in two of the closely related Wa variants (Wag5re and Wag7/8re) (***Figure 3.5***) which is not present in the WaCS. The genome segment 5 (NSP1) of the Wag5re variant contains a large insert of 965 bp. The *in silico* generated ORF for Wag5re indicated that there is 21 bp missing at the 5' end which resulted in an altered amino acid chain compared to the other three variants. The C-terminal of NSP1 is thought to be crucial for binding interferon regulatory factor 3 (IRF3) (Barro and Patton, 2007) and this modification may cause a non-functioning NSP1 of this strain. The virulence of Wag5re was compared to that of ParWa in gnotobiotic pigs. Although the Wag5re was able to infect neonatal gnotobiotic

pigs, its infectivity was similar than the tissue-culture adapted ParWa HRV (Wag5wt 33-1) which has the wild type genome segment 5 as measured by fecal and nasal virus shedding and sero-conversion (L. Yuan personal communication). Virulence of the Wag5re was also reduced compared to the Wag5wt in gnotobiotic pigs as measured by diarrhoea rate and mean cumulative diarrhoea scores. No infectious virus was detected in the intestinal contents by CCIF in the Wag5re or Wag5wt animal group, indicating a very low rate of virus replication in the intestine. The immunogenicity of Wag5re was lower than the Wag5wt as measured by IgM antibody titers on post-inoculation day 8 in the serum, small and large intestinal contents (L. Yuan personal communication).

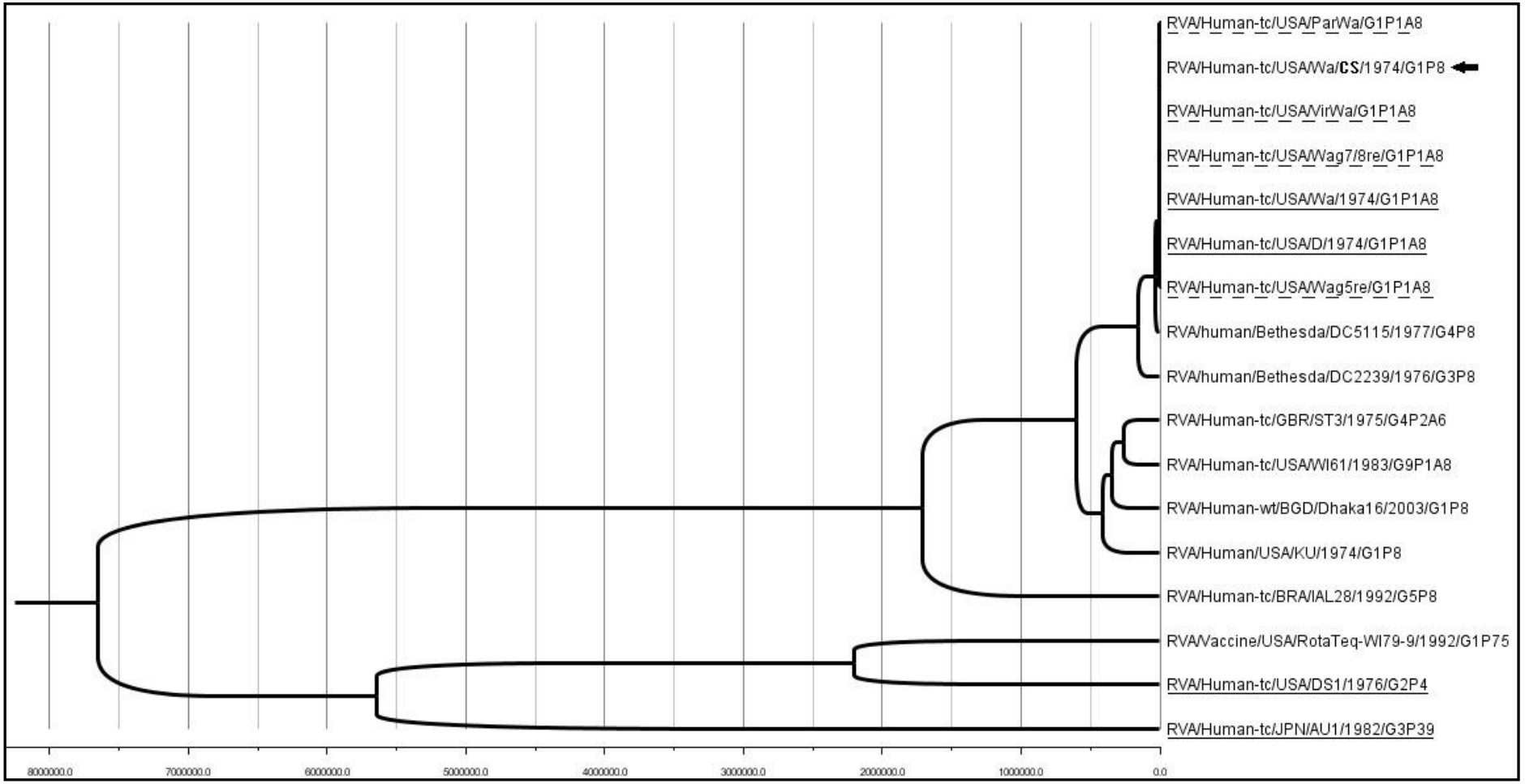

Figure 3.4: ***Combined maximum clade credibility (MCC) tree of all 11 genome segments of the 17 rotavirus sequences analysed using the Bayesian MCMC framework.*** *The sequenced consensus strain (WaCS) is indicated with an arrow, the reference strains (Wa, D, DS-1 and AU-1) are underlined and the closely related variants (ParWa, VirWa, Wag5re and Wag7re) are underlined by a dotted line.*

It has been shown that NSP1 is not crucial for rotavirus replication (Silvestri et al., 2004) and thus it is not surprising that the Wag5re variant could be propagated with a defective NSP1 protein. Nucleotide insertions were also detected in genome segment 8 (NSP2) in the Wag7/8re variant. The nucleotide sequence between positions 83-1009 is repeated in the subsequent 927 bp, but with no effect on the ORF or protein sequence. A similar insertion was detected in genome segments 7 (926 bp in Wag7/8re) and 8 (927 bp in Wag7/8re). These nucleotide segment insertions did not affect the ORF of the particular genome segments, nor were the proteins affected. The insertion in genome segment 7 (NSP3) is a duplication of its own 50 – 976 bp region, and although it contains a large portion of the original ORF, it lacks the first 15 base pairs.

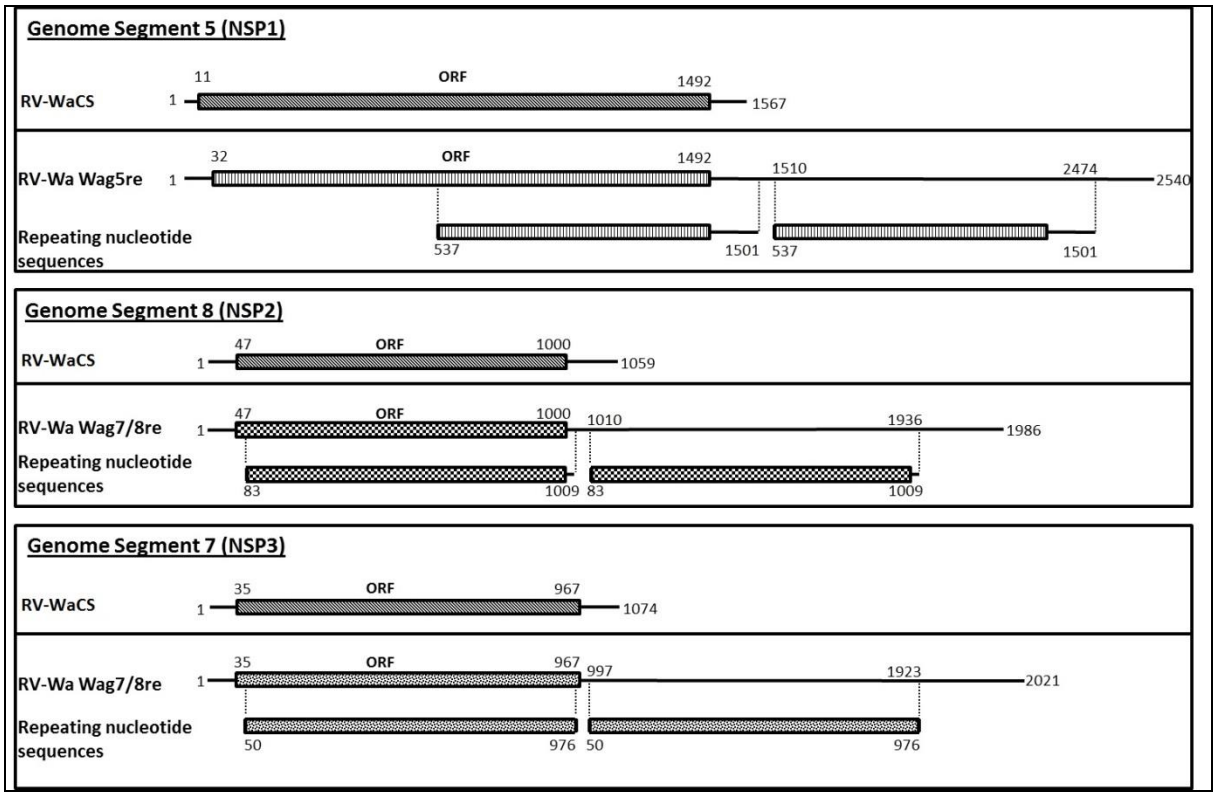

Figure 3.5: ***Comparison between genome segment 5, 7 and 8 of the WaCS and related variants exhibiting nucleotide repeats.*** *Genome segment 5 (NSP1) of the Wag5re variant contains a large insert of 965 bp, the Wag7/8re variant contains nucleotide inserts in genome segment 7 (NSP3) and 8 (NSP2) of 926 bp and 927 bp respectively. The figure compares the genome segments of the WaCS sequence to that of the variants exhibiting nucleotide inserts and also indicates the position of these repeats.*

In silico analysis suggests that there may be formation of a partial second NSP3 protein, lacking the first 6 amino acids at the N-terminal (Taniguchi et al, 1996). The region containing the first 6 amino acids has no known secondary structure or function (Deo et al., 2002) and may therefore not have such a significant influence on the overall protein structure or function. Considering that this large nucleotide insertion of almost 1000 bp would place considerable added strain on genome replication, it is possible that this partial, additional NSP3 protein may have had an advantageous effect for Wag7/8re. The N-terminal of NSP3 binds 3'-mRNA while the C-terminal interacts with the cellular translation initiation factor, eIF4G (Piron et al., 1998a, Poncet et al., 1993). This binding of eIF4G is a similar function as the native poly(A) binding protein, but with a much higher affinity, effectively shutting down the translation of host cell mRNAs. NSP3 is also thought to circularize mRNA which is important for translation initiation of the viral genome (Vitour et al., 2004). Being able to produce a larger amount of NSP3 may have had advantages for viral replication by impairing host protein synthesis more effectively.

3.3.3 Substitution rates and evolutionary pressures analyses

The nucleotide substitution rates were calculated for the closest related members of the rotavirus Wa lineage, spanning more than 50 passages. Individual amino acid sites in a protein are under different selective pressures at any given point in time. By comparing relative fixation rates of silent (synonymous) and protein altering (non- synonymous) substitutions, we are able to identify specific regions under selection. In order to examine the selecting pressures acting on protein-coding regions of the Wa lineage, the non-synonymous–synonymous substitution ratios (ω) were also calculated. Evolutionary rates were estimated for all 11 genome segments of the closely related rotavirus variants (Wa, D, ParWa, VirWa, Wag5re, Wag7/8re) using PAML 4.5. Individual amino acid sites in a protein are under different selective pressures at any given point in time. Non-synonymous/synonymous rate ratio (ω = dN/dS) is one of the most popular statistical methods used to quantify selection pressures on protein-coding regions. Populations with a ω ratios smaller than 1 are considered as under negative selection, ω = 1 as neutral, and ω larger than 1 as under positive selection. PAML determines the general nucleotide substitution rate of a specific sequence, thus giving an idea of the general selection

tendency in a protein coding region. Negative selection (dN/dS<1) was evident in most of the genome segments indicating that very few nucleotide substitutions caused changes in the amino acid sequence. The amino acid selection rate of the examined rotavirus Wa lineage was low and seemed to be quite conserved between the different variants. Nucleotide substitution rates (***Table 3.3***) varied between 1.00×10^{-4} and 6.03×10^{-3} substitution rates persite per year for the different genome segments. Genome segment 6 (VP6) and genome segment 8 (NSP2) had the lowest substitution rates (1.00×10^{-4}), while genome segment 10 (NSP4) had the most nucleotide substitutions per site per year (6.03×10^{-3}).

Table 3.3: Summary of the nucleotide substitution rates and possible sites under diversifying selection of the genome segments of the Wa rotavirus lineages

Genome segment (viral protein)	Evolutionary rate (substitutions per year per site)	dN/dS	Individual sites under diversifying selection
GS1 (VP1)	1.09×10^{-4}	0.327	Negative selection codon 632
GS2 (VP2)	1.96×10^{-4}	0.135	None detected
GS3 (VP3)	1.86×10^{-4}	0.135	None detected
GS4 (VP4)	3.05×10^{-4}	1.491	Positive selection codon 471
GS5 (NSP1)	1.66×10^{-4}	0.092	None detected
GS6 (VP6)	1.00×10^{-4}	0.998	None detected
GS7 (NSP3)	2.94×10^{-4}	0.190	None detected
GS8 (NSP2)	1.00×10^{-4}	0.999	None detected
GS9 (VP7)	1.07×10^{-4}	0.997	None detected
GS10 (NSP4)	6.03×10^{-3}	0.254	None detected
GS11 (NSP5/6)	1.94×10^{-4}	0.998	None detected

Genome segments 1 (VP1), 2 (VP2), 3 (VP3), 5 (NSP1), 7 (NSP3) and 10 (NSP4) have dN/dS<1 and seem to be under general negative selection. Genome segment 4 (VP4) is the only segment with a dN/dS value higher than 1, indicating general positive selection across the sequence. Considering that VP4 is crucial for cellular attachment and penetration, it is not surprising that it is exhibiting positive selection when serially adapted to at least two

different hosts over an extended period. In order to find specific sites that may undergo diversifying selection, the **single likelihood ancestor counting** (SLAC), fixed effects likelihood (FEL) and random effects likelihood (REL) methods were utilized. These three models assemble the distribution of substitution rates across sites and then calculate the rate at which individual sites may evolve (Kosakovsky Pond and Frost, 2005). REL is similar to the codon-based selection analyses found in the PAML software suit and is suitable for analyzing low divergence alignments and small datasets. This method allows synonymous rate variation and is probably best suited for our dataset, but it must be noted, that in some cases, this method is known to be vulnerable to high rates of false positives (Kosakovsky Pond and Frost, 2005). SLAC is usually used to acquire substitution maps at each individual site for large datasets. FEL is considered to be a well-balanced method and is usually applied to medium to large datasets and to ascertain site-by-site substitution rate estimations. Due to the relatively small number of sequences from the same population, the results generated from these three methods were combined, to ensure a more reliable analysis.

A likely negative selected site was detected in VP1 (codon 632) and in VP4 a positive selected site was detected (codon 471) (***Table 3.3***). The individual site under positive diversifying selection identified in VP4, is located in a beta strand region on the VP5* subunit. Although non-synonymous/synonymous ratio tests have been applied to single populations (Holt et al., 2008, Jones et al., 2003, Plikat et al., 1997), the dN/dS ratio was originally developed for genetic analysis of divergent species (Kryazhimskiy and Plotkin, 2008). The data set used for evolutionary date determination and site selection is quite small and derived from a single population. Despite these shortcomings, the evolutionary analysis still gives an indication on the different evolutionary rates of the individual genome segments of a related population.

3.4 Summary

This is the first report of a consensus sequence of a rotavirus Wa strain using sequence-independent cDNA synthesis and amplification combined with 454® pyrosequencing. These results show the complete consensus genome sequence for the cell culture adapted 1974 Wa strain, free from cloning bias and the limitations of sequencing which was used for most

of the Wa sequencing to date. Phylogenetic, pairwise comparisons and a combined molecular clock for rotavirus Wa, indicated that the Wa consensus sequence determined in this study is most closely related to ParWa and VirWa. Since ParWa was passaged 11 times in gnotobiotic pigs followed by cell culture adaption (MA104) and VirWa was passaged 21 times in gnotobiotic pigs, it can therefore be assumed that the WaCS determined in this study was obtained from an early cell culture adapted Wa variant from the initial gnotobiotic pig passaged Wa isolate. Evolutionary rates and genetic adaption of rotavirus strains are important factors to consider for future vaccine development. It is interesting to note that from *in silico* analysis, none of the detected nucleotide changes, and consequent amino acid variations, had any significant effect on viral structure. Despite serial passaging in animals as well as cell culture, including several selection methods, the Wa genome seems to be stable. The WaCS sequence is derived from a single rotavirus Wa population using next generation 454® pyrosequencing. Next generation sequencing is free from cloning bias and has a far better depth of coverage than the Sanger sequencing techniques used to determine the previous rotavirus Wa sequences. Considering that the current rotavirus Wa reference strain is a composite sequence of various Wa variants, the rotavirus WaCS may be a more appropriate reference sequence. The WaCS sequence will be used as template for synthesising all 11 rotavirus Wa genome segments in an effort to create a rotavirus Wa plasmid set (see ***chapter 5***).

Part of the work presented in this chapter was published in the journal Infection Evolution and Genetics (see ***Appendix B*** for article).