MOLECULAR CHARACTERIZATION AND THE *IN VIVO* ESTIMATION OF MUTATION RATE OF THE ALEUTIAN MINK DISEASE VIRUS ISOLATE FROM CAPE BRETON, NOVA SCOTIA

By

Priyanka P Rupasinghe

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
December 2019

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The sequence of 91.6% (4397 bp) of the genome of one Aleutian mink disease virus from Cape Breton (NS-CB) was determined. The comparison of the NS-CB with global strains revealed 10 hypervariable regions at the nucleotide level, of which five were novel. Eleven recombination events were detected between the NS-CB and 31 global strains, and phylogenetic analysis of the NS-CB and 23 global strains with no recombination revealed that than NS-CB was closely related to 19 Danish strains. Twenty mink which were inoculated with the NS-CB were sampled after 16 and 176 weeks post-inoculation (wpi) and NS-CB descendants were sequenced. The number of nucleotide substitutions at 176 wpi was greater than that at 16 wpi (113 vs 36) whereas the estimate of mutation rate at 176 wpi was lower than that at 16 wpi (2.85×10^{-3} vs 9.13×10^{-4} substitutions/site/year), showing a decreasing trend in mutation rate per unit of time.

LIST OF ABBREVATIONS USED

aa	Amino acid
AD	Aleutian disease
ADE	Antibody-dependent enhancement of infection
ADRC	Aleutian Disease Research Centre
AMDV	Aleutian mink disease virus
ATP	Adenosine triphosphate
BC	Binary character
BIC	Bayesian Information Criterion
BLV	Bovine Leukemia provirus
bp	Base pair
BEAST	Bayesian evolutionary analysis sampling trees
CIEP	Counter- immunoelectrophoresis
CPV	Canine parvoviruses
CRFK	Crandell Rees feline kidney
DMSO	Dimethyl sulfoxide
dNTP	Deoxynucleotide
dN	Non-synonymous substitution
dsDNA	Double strandard DNA
ECR	Entire coding region
ELISA	Enzyme-link immunisorbent assay
ExPASy	Expert Protein Analysis System
FPV	Feline panleuopenia virus
GFAV	Gray fox amdovirus
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HIV	Human influenza virus
HVR	Hypervariable region
ICTV	International Committee on the Taxonomy of the Virus
IUPAC	International Union of Pure and Applied Chemistry
kb	Kilo base
LORF	Left open reading frame
LTR	Long Terminal Repeat
mb	Mega base
MDCK	Madin-Darby Canine Kidney
MEGA	Molecular Evolutionary Genetics Analysis
MD	Missing data
MEV	Mink enteritis virus
ML	Maximum likelihood
MP	Maximum parsimony
MPR	Mid-point rooting
MSA	Multiple sequence alignment
MUSCLE	Multiple Sequence Comparison by Log-Expectation
MVM	Minute virus of mice

NGS	Next-generation sequencing
NJ	Neighbor-joining
NS	Nova Scotia
NS1	Non-structural protein 1
NS2	Non-structural protein 2
NS3	Non-structural protein 3
NS-CB	Nova Scotia-Cape Breton
nt	Nucleotide
PBS	Phosphate buffer saline
wpi	Weeks post- inoculation
PAUP	Phylogenetic Analysis Using Parsimony
PPV	Procine parvovirus
RDP	Recombination Detection Program
RF	Replicative form
RFFAV	Red fox fecal Amdovirus
RHR	Rolling Hairpin Replication
RORF	Right open reading frame
SARS	Severe Acute Respiratory Syndrome
ssDNA	Single-stranded DNA
SDT	Sequence Demarcation Tool
UPGMA	Unweight Pair Group Method with Arithmetic mean
UTR	Untranslated region
VP1	Virion protein 1
VP2	Virion protein 2

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Dr. Hossain Farid for giving me the opportunity to work on this exciting research project as well as for his valuable guidance throughout my study period. His enthusiastic passion for research has provoked me to expand my knowledge in genomics and develop the skills in critical thinking and paying attention to the details. I would also like to extend my appreciation to Dr. Svetlana Yurgel and Dr. Bruce Rathgeber for serving my advisory committee and providing their advice throughout the study.

I wish to thank Dr. Hussain Irsahd and Irin Irju, who conducted the inoculation experiment that provided research materials for my thesis research. I gratefully acknowledge the assistance given by the ADRC staff for animal management and collecting my research samples.

I would like to recognize the scholarship awarded by the Canadian Mink Breeders Association. This research was supported by the funds of Canada Mink Breeders Association, Nova Scotia Mink Breeders Association, Fur Farmers of Canada Marketing Association, American Legend Cooperatives, North American Fur Auctions, American Mink Council, American Mink Farmers Research Foundation, Agriculture and Agri-Food Canada through the CARD Councils of Ontario, British Columbia and Nova Scotia (Agri-Futures NS) and Atlantic Innovation Fund of the Atlantic Canada Opportunity Agency.

This achievement would not have been possible without the support of my husband Vasantha, son Kalana, and daughter Viraji. Last but not least, my heartfelt thank goes to my mother, sisters and their families for their encouragement from thousands of miles away.

CHAPTER 1. INTRODUCTION

Aleutian disease (AD) is a highly contagious chronic illness of mink (*Neovison vison*) which exists on commercial mink farms in all mink producing countries (Farid *et al*., 2012; Jahns *et al*., 2010; Knuuttila *et al*., 2009; Oie *et al*., 1996; Olofsson *et al*., 1999; Wang *et al*., 2014) except Iceland (Gunnarsson, 2001). AD causes significant economic losses in the 97 million-dollar Canadian mink industry (Statistics Canada, 2017) by increasing adult and kits mortality and decreasing reproductive success (reviewed by Canuti *et al*., 2015; Reichert and Kostro, 2014). In addition, AD causes the development of undesirable white hair fibers on pelts which reduces their market value (Farid and Ferns, 2011).

The causative agent of AD is the Aleutian mink disease virus (AMDV), a non-enveloped autonomous parvovirus with a 4801nt long single-stranded DNA (ssDNA) genome (reviewed by Bloom *et al*., 1994). In adult mink, AD is characterized by hypergammaglobulinemia, plasmacytosis, glomerulonephritis and an immune complex disorder (Bloom *et al*., 1994). Furthermore, AMDV infection causes acute, rapidly progressing pneumonia, with high mortality rates in neonatal mink kits (Bloom *et al*., 1994). Even though host factors, including genotype and age, contribute to the severity of AMDV infection, the key determinants of disease variability and severity are encoded in the viral genome (Bloom, *et al*., 1994; Oie *et al*., 1996). Several strains of AMDV have been identified, ranging from the non-pathogenic to highly virulent strains (Bloom *et al*., 1994; Oie *et al*., 1996). Highly virulent strains cause disease in all mink color types and the moderately pathogenic strain, the Pullman strain causes low mortality in kits and mild

disease symptoms in the non-Aleutian genotype (royal pastel). The Pullman strain is able to cause severe disease in an Aleutian genotype (sapphire) (Hadlow *et al*., 1983).

` There is no cure or vaccine against AD (Aasted *et al*., 1998; Castelruiz *et al*., 2005). The current strategy for controlling the disease is testing of mink for antibodies against AMDV by counterimmunoelectrophoresis (CIEP) (Cho and Ingram, 1972) or recently by Enzyme-link immunosorbent assay (ELISA) (Dam-Tuxen *et al*., 2014; Farid and Segevall, 2014; Knuuttila *et al*., 2014) and culling seropositive animals. Test and kill strategies have not brought a permanent eradication of the virus in many ranches in Nova Scotia (NS) (Farid *et al*., 2012) or in other countries like Denmark (Canuti *et al*., 2015). Persistence and re-appearance of AMDV infection on mink ranches are common due to several factors; such as false-negative CIEP test results (Farid *et al*., 2012), persistence of the virus in the environment (Hussain *et al*., 2014), or transmission of the virus from neighboring farms or infected wildlife (Farid *et al*., 2013; Oie *et al*., 1996). It is important to identify the sources of persistent infection or re-contamination of the ranches in order to develop strategies to remove the virus from the premises and prevent its introduction to the ranch. The sequencing of the viruses on a ranch and comparison of the sequences with existing sequence databases are the only ways to identify the source of contamination. This approach is hampered by the unusually high degree of genetic diversity of AMDV as a result of its high mutation rate (Gottschalck *et al*., 1991; Olofsson *et al*., 1999). So far, there has not been a timescale measurement of the mutation rate of the AMDV genome, because most of the pathogenic strains of the AMDV cannot be propagated in cell-culture except the AMDV-G (Bloom *et al*., 1994), and the SL3 (Schuierer *et al*., 1997) strains. Therefore, the mutation rate of AMDV can only be estimated *in vivo*, where a uniform

group of animals needs to be inoculated with a single AMDV isolate and followed by the nucleotide (nt) variations over time. The mutation rate of the AMDV over time will be estimated in this study using the nt changes of the nearly complete genome sequence of the AMDV isolate circulating in Cape Breton, NS, which has not been performed before. The findings will provide benefits to mink ranches by understanding the origin of the virus on a newly infected ranch through comparing of the existing AMDV sequence(s) on a farm and previously prepared databases.

CHAPTER 2. LITERATURE REVIEW

2.1. Parvoviruses

Parvoviruses, as the name suggested from the Latin ('parvum' meaning small) are the smallest of known viruses (Cotmore *et al*., 2014). Parvoviruses genome is a linear non-segmented single-stranded DNA molecule, and an average genome size of approximately 5000 nucleotides (nt). Parvoviruses infect a wide variety of hosts, ranging from insects to primates. The family *Parvoviridae* is divided into two subfamilies *Parvovirinae* and *Densovirinae,* based on the taxonomic group of the host species. The sub-family *Parvovirinae,* which includes viruses that infect vertebrate hosts. The sub-family *Densovirinae* contains a large group of viruses, which contain viruses that infect arthropod hosts. The subfamily *Parvovirinae* is further divided into eight genera: *Protoparvovirus, Amdoparvovirus, Aveparvovirus, Bocaparvovirus, Dependovirus, Erythrovirus, Copiparvovirus, and Tetraparvovirus* according to the genome organization (Cotmore *et al*., 2014). The diversity of the genus *Amdoparvovirus* has increased after the International Committee on Taxonomy of Viruses (ICTV) 2011 classification. For that reason, genus *Amdoprvovirus* has been subjected to reclassification (Cotmore *et al*., 2014). So far, five variants have been identified in genus *Amdoparvovirus,* in which AMDV is the only member of Carnivore *Amdoparvovirus* 1. Another member of this genus is gray fox *Amdoparvovirus* (GFAV, Li *et al*., 2011), which belongs to Carnivore *Amdoparvovirus* 2, and the raccoon dog and fox *Amdoparvovirus* (RFAV, Shao *et al*., 2014) which is classified under Carnivore *amdoparvovirus* 3. The red fox fecal *Amdovirus* (RFFAV) belongs to Carnivore *amdoparvovirus* (Bodewest *et al*., 2014). Recently identified parvovirus from striped skunk was added to the Carnivore *amdoparvovirus* 5 (Canuti *et al*., 2017).

2.2. Genome organization

2.2.1. Genome and genes

The genome structure of AMDV was elucidated using the AMDV-G, a cell-cultured derivative of highly pathogenic Utah1 strain (Bloom *et al*., 1994). The single-stranded 4801 nt long genome of AMDV is encapsulated in a non-enveloped icosahedral capsid as a single "minus'' sense polarity molecule. Similar to other parvoviruses, AMDV possesses self-complementary palindromic sequences at both ends of the genome. The left 3' end contained 117 nt palindrome with the Y shaped configuration, but not identical to any other parvoviruses. While the right 5' end does not possess a hairpin structure, it ends with a 25 nt A &T rich repeat. AMDV is characterized as heterotelomeric due to the differences in size and the sequences of the termini. The virus has a promoter (P3) located at nt 151-160 (GTATATAAGC), -29 nt from the initiation of transcription. The presence of another weak promoter (P36) at nt 1729 (TATTAA) has also been reported (Alexandersen *et al*., 1988). There is an upstream CAT enhancer element at nucleotides 97-99 (-83 nt from initiation of transcription) with the sequence CATTA. Two polyadenylation signals (AATAAA) are located at nt 2564 and 4394. There are two left and right ORF's. The left ORF encodes for three non-structural proteins (NS1, NS2, and NS3), and the right ORF encodes the capsid proteins VP1 and VP2 (Huang *et al*., 2014). Also, AMDV has one or two smaller middle ORFs (Bloom *et al*., 1994). In addition to the complete genome sequence of the AMDV-G, the complete genome sequence of pathogenic AMDV-BJ (Xi *et al*., 2016) is also available on GenBank. The majority of the nearly complete AMDV genome available on GenBank (n=32) have been annotated, and locations of the five genes have been identified, except SL3, LM, Beijing, LN2, and LN3 strains.

2.2.2. AMDV replication

Parvoviruses do not encode the DNA polymerase, because of the limited coding capacity. The autonomous parvoviruses only replicate in the nucleus of the host's dividing cells where they can utilize components of the host cell DNA replication machinery. The formation of a complementary DNA strand of the single-stranded genome using the host DNA polymerases is essential for the initial step of parvovirus DNA replication. Therefore, S-phase of the host cell cycle is generally critical for parvovirus replication. Similar to other parvoviruses, AMDV replicates its genome by rolling palindromic sequences located at the genomic extremities. The Y shaped 3' end hairpin functioning as primers in viral DNA replication. The single-stranded genome is first converted into replicative forms (RF) after the host polymerase binds to the 3' termini of the hairpin (Bloom *et al*., 1994). Therefore, the presence of the RFs of AMDV in an organ is a strong indication of virus replication at that site (Alexandersen *et al*., 1988).

2.2.2.1. Virus replication *in vivo*

DNA replication is an essential part of the pathogenesis of AMDV infection in neonatal and older mink. For example, AMDV causes acute interstitial pneumonia in neonatal mink, destroying alveolar cells and leading to fatal respiratory distress syndrome due to permissive viral replication compared to the chronic disease in the adult minks resulting from restricted replication (Alexsanderson *et al*., 1998; Bloom *et al*., 1994). The permissive replication of the AMDV is characterized by the low levels of anti-AMDV antibody in neonatal mink, and high levels of viral DNA in the infected cells. The infection of adult mink with AMDV leads to the persistent infection or chronic form of the disease

associated with the excessive synthesis of specific anti-AMDV antibodies (hypergammaglobulinemia). Furthermore, the antiviral antibody combines with the virus to make antibody-antigen complexes that are deposited in multiple organs. These complexes also facilitate the entry of the virus to macrophages, the target cells for persistent infection. The infection of macrophages is mediated through the Fc-receptor-dependent mechanism. Therefore, AMDV infection of macrophages was identified as antibody-dependent enhancement of infection (ADE) (Bloom *et al*., 1994; Canuti *et al*., 2015).

2.2.3. Nucleotide and amino acid hypervariable regions

Parvovirus genomes are highly conserved (Chapman *et al*., 1993), but AMDV isolates show a high degree of variability in both nt and aa sequences of the structural and non-structural genes (Bloom *et al*., 1988; Gottschalck *et al*., 1994; Oie *et al*., 1996). The nt and aa variabilities are particularly high in some regions of the genomes, known as hypervariable regions (HVR). For the first time, Bloom *et al*. (1988) identified a 36 nt long segment of the VP2 gene with higher variability (44.4%) compared with other parts of 3456 nt sequences of AMDV-G and Utah1 (Table 2.1). Four HVRs were identified by Olofsson *et al*. (1999) in the NS1 gene of 39 strains of AMDV, and Oie *et al*. (1966) tested the variability of the HVR which was already identified by Bloom et al. (1988) in six strains of AMDV and reported almost the same level of nt variability. Sequence comparison of the NS1 proteins of the AMDV-G, K, SL3, United, Utah1, and 35 other AMDV isolates revealed a high aa variability in four sections of this protein (Olofsson *et al*., 1999). High aa variability is also shown by Oie *et al*., 1996 (Table 2.2). It has been shown that the aa

and carboxy termini of the NS1 proteins showed higher variability than the middle region of the protein (Gottschalck *et al*., 1994).

Viral genome analyses often focus on conserved segments of nt or aa residues, because a high level of aa homology is required for maintaining the structure and function of protein (reviewed by Valdar, 2002). The biological importance of HVRs in the AMDV genome has not been studied in detail. It has been suggested that nt substitutions in the HVR of AMDV have a tendency to be non-synonymous (Olofsson *et al*., 1999). Amino acid changes in HVR of VP2 has been identified as a determinant of host specificity of related parvoviruses, such as canine parvovirus (CPV) (Chang *et al*., 1992) and feline parvovirus (FPV) (Truyen *et al*., 1995). The aa sequence of the HVR1 of the hepatitis C virus (HCV) at the N terminus of the viral glycoprotein E2 contains epitopes that are recognized by patients' antibodies (Kato *et al*., 1993). The aa residues of the VP2 HVR are not related to the AMDV pathogenicity (Bloom *et al*., 1993; Oie *et al*., 1996). X-ray crystallography of capsid atomic structure revealed that aa residues of the VP2 HVR are mostly exposed at the mounds of the capsid, whereas the tissue tropism and disease pathogenicity determinant are located near the threefold axes in the dimple or the wall of the dimple (McKenna *et al*., 1999).

The most common uses of the HVR of the AMDV genome are for the identification of related viral isolates using phylogenetic techniques. The regions flanking the HVRs of the NS1 or VP2, identified by Olofsson *et al*. (1999) and Bloom *et al*. (1988), respectively, have been PCR amplified and sequenced for genome analysis (Table 2.1). It should be noted that Oie *et al*. (1996) sequenced the HVR of the VP2 gene, originally identified by Bloom *et al*. (1988), and other researchers referred to this region as the one identified by

Oie *et al*. (1996) (Table 2.2). The information on the VP2 HVR of AMDV has been used in analyzing the ferret parvovirus (ADVF) (Saifuddin and Fox, 1996).

The number of different nts or aa residues are the measures of variability, which are calculated in multiple sequence alignment (MSA). Bloom *et al*. (1988) and Olofsson *et al*. (1999) who reported HVR of AMDV for the first time, by aligning different sequences and manually counting the number of variants at each position. Several approaches for the determination of conserved and variable regions of proteins have been proposed (Valdar, 2002), of which Shannon entropy score is widely used for measuring variability in aa residues (Troesch *et al*., 2006) and nts (Batista *et al*., 2011). The Shannon entropy measures the degree of diversity at a given position in a set of aligned sequences, for which a few different formulas have been used. The frequency scores for the protein sequences have a sum of symbols up to 21 terms and nt sequences up to four (Valdar, 2002). Batista *et al*. (2011) used the software DAMBE (Data Analysis in Molecular Biology and Evolution, http://dambe.bio.uottawa.ca/) to identify the most conserved regions of the Papillomavirus. This software uses the equation $Hi=-(\sum^{4}_{j=1} P_{ij} \log_2 P_{ij})$, where Hi is the entropy of each site i, j is equal to 1, 2, 3 and 4, corresponding to the A, C, G, and T nucleotides, respectively, and p_{ij} is the proportion of the nucleotide j in the site i. The values range between zero (no variability) and 2.0 (1/4 each nt). The entropy measurement has been used to identify a novel HVR in the hepatitis C virus (HCV) and the E2 envelope glycoprotein (Troesch *et al*., 2006). They used Entropy-One Webtool (Korber *et al*., 1994). It should be noted that even though the number of published AMDV sequences have dramatically increased in recent years, no attempt has been made to identify HVRs of the new sequences using entropy.

Table 2.1. Nucleotide variability of the inside and outside of the hypervariable regions

HVR	Region compared				Hypervariable region				
1	Strains	Length	Location	Variant Number (%)[a]	Location	Gene	length	Variants number (%)[b]	Ref.[c]
1	G & Utah1	3456	721-4176	57 (1.7)	3094-3130	VP2	36	16 (44.4)	1
1	39 strains	336	587-923	44 (20.3)	652-698	NS1	47	10 (21.2)	2
2					687-721		35	11 (31.4)	
3					827-851		25	11 (44.0)	
4					879-890		12	5 (41.6)	
1	6 strains	1944	2406-4349	125(6.4)	3099-3131	VP2	33	26(78.78)	3

[a]Number and percent variable sites (in bracket), excluding the HVR.
[b]Number and percent variable sites (in bracket) in the HVR.
[c]References: 1. Bloom *et al*., 1988; 2. Olofsson *et al*., 1999; 3. Oie *et al*., 1996

Table 2.2. Amino acid variability of the inside and the outside of the hypervariable regions

HVR	Region compared				Hypervariable region				
1	Number of strains	Length	Location	Variant Number (%)[a]	Location	Gene	length	Variants number (%)[b]	Ref.[c]
1	39	112	128-239	4(3.4)	128-150	NS1	23	8(34.7)	1
2					169-181		13	7(53.8)	
3					209-217		9	5(55.5)	
4					225-228		4	4(100)	
1	15	53	211-263	3(5.6)	232-242	VP2	11	9(81.8)	2

[a]Number and percent variable sites (in bracket), excluding the HVR.
[b]Number and percent variable sites (in bracket) in the HVR, [c]References 1. Olofsson *et al*., 1999; 2. Oie *et al*., 1996

Table 2.3. AMDV genome analyses using the hypervariable regions

Reference	Gene and positions[a]		Number and type of positions	HVR reference
Saifuddin and Fox, 1996	VP2[b]	3043- 3407	365 nt	Oie *et al*., 1996
Knuuttila *et al*., 2009	NS1[c]	563-952	390 nt	Olofsson *et al*., 1999 (HVR 1-4)
Christensen *et al*., 2011	NS1	578- 951	328 nt	Olofsson *et al*., 1999 (HVR 1-4)
Jensen *et al*., 2011	NS1	373-747	374 nt	Olofsson *et al*., 1999 (HVR 1-2
Nituch *et al*., 2012, 2015	NS1	601-922	322 nt	Olofsson *et al*., 1999 (HVR 1-4
Nituch *et al*., 2012, 2015	VP2	2725-3255	531 nt	Oie *et al*., 1996
Leiman *et al*., 2015	NS1	563-952	390 nt	Olofsson *et al*., 1999 (HVR 1-4)
Leiman *et al*., 2015	VP2	2587-3252	665nt	Oie *et al*., 1996
Persson *et al*., 2015	NS1	373-747	374 nt	Olofsson *et al*., 1999 (HVR 1-2)
Ryt-Hansen *et al*., 2017	NS1	578- 951	328 nt	Olofsson *et al*., 1999 (HVR 1-4)
Mañas *et al*., 2001	VP2	70-180	111 aa	Oie *et al*., 1996
Jahns *et al*., 2010	VP2	211-250	40 aa	Oie *et al*., 1996
Jakubczak *et al*., 2016	VP2	172-353	182 aa	Oie *et al*., 1996

[a] Positions are based on the AMDV-G.
[b] Virion protein 2.
[c] Non-structural protein 1.

2.2.4. Transcription profile, proteins and functions

2.2.4.1. Transcription profile

Similar to other viruses, AMDV employs an extensive overlapping transcript and the use of multiple reading frames to increase the genetic information included in its small genome. The virus generates a single pre-mRNA from the P3 promoter which undergoes alternative splicing and polyadenylation to generate six different mRNA transcripts (Huang *et al*., 2014; Qiu *et al*., 2006). The splicing process is controlled by specific sequences, known as splice-donors and splice acceptors. The three donor sites are located at nt 384, 1961 and 2213, and the three splice acceptor sites are located at nt 1737, 2042 and 2287 of the AMDV genome, as shown in Figure 2.1 (Huang *et al*., 2014; Qiu *et al*., 2006).

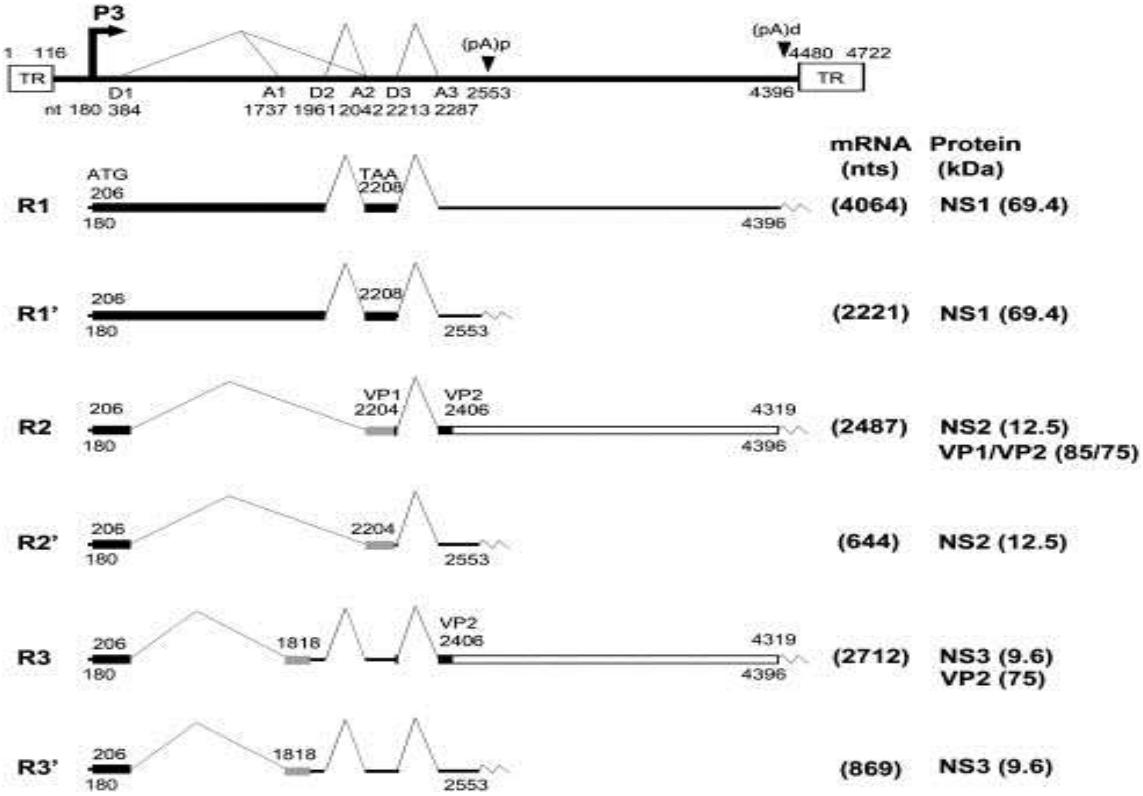

Figure 2.1. Transcription profile of AMDV-G (AC#- JN040434). Splice donor sites are (D1, D2, and D3), Splice acceptor site are (A1, A2, and A3), and the six mRNA transcripts and polyadenylation sites are shown (adapted from Haung *et al*., 2014).

2.2.4.2. Non-structural proteins and their functions

AMDV is coding for three non-structural proteins, NS1, NS2, and NS3. All three proteins share 59 aa residues at the N-terminus of the polypeptide and process a unique region for each NS protein at the C- terminus. The NS2 and NS3 contain 54 and 27 unique aa residues in the C terminal end, respectively (Huang *et al*., 2014; Qiu *et al*., 2006). The NS1 plays different roles in genome replication in parvovirus infection by possessing motifs for DNA binding, and ATPase and helicase activities (Nüesch and Rommelaere, 2006). The ATP-binding pocket of AMDV consists of six aa residues from 435 to 440 (GTGKTL), which is completely conserved in AMDV-G, Utah 1, United and K strains (Gottschalck *et al*., 1994).

The NS1 also facilitates virus replication by possessing two caspase recognition sequences which were mapped to the aa residues 224INTDS228 and (282DQTDS286), with recognition sites 227D↓S and 285D↓S. During the AMDV replication, caspases cleave the NS1 at these two sites (Best *et al*., 2003). The majority of the full-length NS1 and the C-terminal caspase-cleaved fragment of NS1 (aa 286 to 641) mainly found in the nucleus during replication (Best *et al*., 2003). It has also been shown that when one of its cleavage sites is mutated, the full NS1 remains in the cytoplasm, indicating that caspases cleavage is required for nuclear localization of full-length NS1. The absence of NS1 protein in the nucleus also leads to the decreased expression of NS1-dependent viral capsid protein (Best *et al*., 2003). There are some differences in caspase cleavage sites among the AMDV strains. The first caspase recognition sequence in NS1 (224INTDS228) is variable among some isolates, whereas the second one (282DQTDS286) is conserved among isolates of AMDV (Best *et al*., 2003; Leng *et al*., 2018). This conservation of caspase recognition site

implies that caspase activity and NS1 cleavage are essential in the replication of both pathogenic and non-pathogenic AMDV strains and supports the idea that caspase activity regulates virus replication *in-vivo* as well as *in vitro* (Best *et al*., 2003). High levels of NS2 protein and its localization with NS1 in the nucleus were reported (Huang *et al*., 2014; Qiu *et al*., 2006). In contrast, NS3 was expressed at low levels and did not localize with NS1 in the nucleus. Huang *et al*. (2014) also investigated the role of NS1, NS2, and NS3 in AMDV replication and viral progeny production and demonstrated that all non-structural proteins play important roles in AMDV replication.

2.2.4.3. Structural proteins and their functions

The AMDV capsid is formed by assembling of 60 VP1and VP2 protein subunits, with approximately 1: 9 ratio, which is arranged to produce a negatively charged surface. The VP1 contains the entire sequence of the VP2 and has an additional 43 aa residues on its N-terminus (Bloom *et al*., 1994). AMDV capsid proteins have an ability to self-assemble into empty capsids when expressed using recombinant vaccinia or baculovirus systems in invertebrate cells (Bloom *et al*., 1994). These empty capsids have similar physicochemical and antigenic properties of native virions. For that reason, attempts been made for developing the particle-based vaccine for AMDV (Castelruiz *et al*., 2005). In addition to building the capsid, VP2 protein determines the cellular host range and the pathogenicity of the virus (Bloom *et al*., 1994).

Differences among AMDV strains for pathogenicity for mink have been reported. The Utah1 and TR are highly pathogenic, and Pullman and SL3 are moderately pathogenic and cause AD symptoms in adult mink (Bloom *et al.,* 1994; Haas *et al*., 1991; Hadlow *et*

al., 1983, 1984; Oie *et al.,* 1996). These strains do not replicate in cell-culture, except for the SL3, which can be replicated in cell culture to some degree (Haas *et al*., 1991). On the contrary, AMDV-G, which was derived from pathogenic Utah1, is non-pathogenic for mink but replicates in Crandell feline kidney cells (CRFK) (Bloom *et al*., 1994).

There have been several attempts to clearly identify genomic regions and variants that control pathogenicity. Bloom *et al*. (1993) replaced seven segments of the AMDV-G with the corresponding portions from pathogenic Utah1 strain to identify particular sequence variation that determines replication capacity in cell culture and pathogenicity in mink. The chimeric constructions were transfected into CRFK cells to check their ability to propagate and yield infectious viruses. Several of the clones replicated in cell culture, and the ability of AMDV to replicate in CRFK cells were assigned to a 531-bp fragment spanning nt 2553 to 3084 of the capsid protein. There are 14 nt differences in this segment, suggesting that one or more of these differences determine the replication capacity of AMDV in CRFK cells. The *in vivo* infectivity and pathogenicity of chimeric viruses were determined by inoculating adult sapphire mink. Serum anti-AMDV antibodies were tested by CIEP on days 10, 30, and 60 post-inoculation (pi), and viremia on day 10 pi was used as the measure of virus replication. Animals were inspected for the typical lesions of AD at necropsy 60 to 90 days pi. None of the constructs showed antibody, viremia or signs of AD in mink. It was concluded that that the presence of the Utah1 hypervariable region in an AMDV-G virus was not enough to cause pathogenicity. They concluded that more research was needed to clarify how the determinants function *in vitro* and whether the same genomic region governs *in vivo* replication and pathogenicity.

Comparisons between sequences of the non-pathogenic AMDV-G and pathogenic strains led researchers to propose that the pathogenicity of AMDV isolates and their ability to propagate in cell culture are regulated by five aa residues in the VP2 protein, namely 352, 395, 434, 491, and 534 (Bloom *et al*., 1994; Oie *et al*., 1996). Three of the five aa residues in AMDV-G, which is non-infective to mink but can grow in cell-culture (Bloom *et al*., 1994) are different from those in the five pathogenic strains (H395Q, N491D, E, and H534D) (Table 2.4). The aa at position 352 (I) is also different from that in other strains, except in the SL3 (Schuierer *et al*., 1997), and AMDV-G and TR share the same aa at position 434 (N) (Oie *et al.,* 1996).

Table 2.4. Positions and amino acid of in vivo determent in VP2 of AMDV isolates

Isolate	Positions of amino acids[a]				
	352	395	434	491	534
AMD-G	I	H	N	N	H
Utha1	V	Q	H	E	D
TR	V	Q	N	D	D
ZK8	V	Q	H	E	D
Pullman	V	Q	E	E	D
SL3	I	Q	H	E	D

[a] D-Aspartate, E-Glutamate, H-Histidine, I- Isoleucine, N-Asparagine, Q-Glutamine, V-Valine

In a study (Fox *et al*., 1999), all five aa residues of the AMDV-G were replaced by those in the Utah1 strain (I352V, H395Q, N434H, N491E, and H534D) using the site-specific mutagenesis. These constructs were used individually to transfect CRFK cells in order to assess their ability to propagate. Three of the mutant clones (352V, 491E and 534D) and AMDV-G, had a similar replication titer (10^7 FFU/ml), and 395Q and 434H did not grow in cell culture. It was concluded that aa residues at positions 395 (H) and 434 (N) play important roles in AMDV replication *in vitro*. It was stated that the reason for single mutations at positions 395 and 434 being replication-defective in cell culture was not clear,

and the interaction of these single mutants with other aa residues could determine virus replication *in vitro.* In the same study (Fox *et al*., 1999), sapphire mink were inoculated with the three constructs which showed replication *in vitro* (352V, 491E, and 534D), along with the chimeric molecular clone G/U-10 (Bloom *et al*., 1988) and Utah1 strain. The 395Q and 434H clones were not included in the *in vivo* study. The mink inoculated with 491E did not show viremia, detectable levels of antibody or AD lesions. The viremia or AD lesions were not detected in animals inoculated with 352V either, but it showed detectable antibody response. The G/U-10 and Utah1 strain showed antibodies and viremia, but 534D showed low antibody titer compared to the G/U-10 and Utah1 and did not cause pathological lesions. Although 534D did not cause AD lesions, it was concluded that 534D is the determinant of pathogenicity because it showed both antibody and viremia.

In another study, full-length chimeric molecular clones between AMDV-G and Utah1 strains were constructed to identify the regions of AMDV capsid protein that modulate pathogenicity and ability to replicate in CFRK cells (Bloom *et al*., 1998). Twelve chimeras were constructed and tested, and two (G/U-8 and G/U-10) were pathogenic to mink. The capsid sequences of G/U-8 and G/U-10 differed from AMDV-G at five and four amino acid residues, respectively. The G/U-8 included Utah1 aa residues at positions 352 (V), 395 (Q), 434 (H), 491 (E) and 534 (D), but the G/U-10 had those aa residues, except had AMDV-G aa at position 352 (I). To determine pathogenicity, five adult Aleutian mink were inoculated with each of G/U-8 or G/U-10 clones. The antibody was measured by CIEP and viremia was determined by PCR on 10, 30, 60 and 90-days pi, and animals were subjected to histopathology. These two clones developed viremia, antiviral antibody, and histopathological lesions of AD. It was also found that two regions of the AMDV genome

determines the ability of a virus to replicate in cell culture. The results of the above studies confirmed that the ability to propagate in cell culture and pathogenicity of AMDV is regulated by sequences in the capsid protein gene.

Capsid proteins also possess a motif (DLLDG) from aa 417 to 421 (McKenna *et al*., 1999), which can be recognized by caspases and cleave after aspartic acid D420 (Cheng *et al*., 2010). It has been shown that caspase cleavage of capsid protein reduces the production of virions by limiting the capsid protein availability to encapsulate the single-stranded AMDV DNA. It was concluded that the cleavage of capsid protein during AMDV infection may affect in regulating persistent infection of AMDV (Cheng *et al*., 2010). Sang *et al*. (2012) found that caspase recognition site 417DLLD↓G421 was conserved in 19 full-length VP2 aa sequences included in their study.

2.5. AMDV host range

In addition to farm-raised mink, free-ranging American mink populations are infected in many countries, including Canada (NS and ON), USA, France, Spain, UK, Denmark, Finland, Estonia, Ireland, the Netherlands, Germany, Russia and China (Canuti *et al*., 2015). The highest prevalence of AMDV (93.7%) in wild American mink was reported in NS (Farid, 2013). In addition to American mink, members of the *Mustelidae* family (European mink, ferrets, weasels, stone martens, otters, and fishers) are also susceptible to AMDV infection. Some carnivores (foxes, skunks, raccoons, and bobcats) were also reported to be infected with AMDV (Farid, 2013; Oie *et al*., 1996). The phylogenetic analysis suggests that wild Mustelid species in Finland carried AMDV isolates closely related to those found in farmed mink as well as those which were different

from farmed mink isolates (Knuuttila *et al*., 2015). Published information on AMDV infection of wild animal species indicates that the virus can be transmitted from wild animal species to farmed mink (Oie *et al*., 1996), causing health problems and contravene the AMDV control strategies (Farid, 2013).

2.6. Genetic variability and sequence relatedness

2.6.1. Nucleotide and amino acid sequence variation

The two published complete AMDV genomic sequences including the 3' and 5' ends of the virus, are AMDV-G (Bloom *et al*., 1994), and AMDV-BJ (Xi *et al*., 2016). There are no nt differences in 3' end of the above two genomes. However, structural differences appear at 5' end due to the deletion of 4 nts (285-288) in AMDV- BJ strain. The sequence diversity of the left ORF of several AMDV isolates from Finland, Sweden, the Netherlands, Denmark, Canada, and China has been reported. Partial sequence analysis of 336 bp fragment of the AMDV NS1 gene showed that 86 % to 100% differences in nt and 78%-100% differences in aa in the same fragment of virus strains isolated from farm-raised mink in Finland (Knuuttila *et al*., 2009). Coding sequence analysis of AMDV NS1 protein of four AMDV strains revealed that TATTAA sequence was conserved among the AMDV-G, Utah1 and United strains, while AMDV-K had a substitution at nt 1729 (CATTAA). This transitional mutation may have led to the AMDV-K with different promotor function in relation to the pathogenesis (Christensen *et al*., 1993). Sequence comparison of non-structural genes (123-2208 nt) of four AMDV strains (G, Utah1, K, and United) also shows an unusually high degree of variability (Gottschalck *et al*., 1994). Sequence similarities of the NS genes (123-2208 nt) of AMDV-G with Utah1, K and

United were 99.2%, 88.6%, and 94.8%, respectively. The sequence comparison of the AMDV-BJ left ORF with seven other AMDV strains originated from three different continents revealed that the AMDV-BJ strains had the highest number of nt substitutions leading to the twenty-two aa substitutions. Those non-synonymous nucleotide substitutions had not been reported in other strains before. The insertion of 15 nucleotides at the end of the left ORF (C terminal of NS3 protein) corresponds to the residues ETEDG.

The genomic region coding for the VP2 has been shown to have up to 11 % nt difference, reflecting that the VP2 gene is more conserved among different AMDV isolates (Bloom *et al*., 1988; Schuierer *et al*., 1997). The N-terminal region of the VP2 gene has been identified as a possible region for detecting AMDV strain variation. The nt sequence can vary between 8–39 nt near the N-terminus in some isolates due to the nt deletions between nt 2469 and 2516 in GT repeat region. Therefore, the number of Glycine residues in poly-G region on the VP2 gene is variable among different AMDV strains (Li *et al*., 2012; Wang *et al*., 2014; Xi *et al*., 2016). Strain-specific C→T transition at nt 513, C→A transversion at 846, and C → A, mutation at position 949 of VP2 were identified in the Chinese isolates (Wang *et al*., 2014). Jakubczak *et al*. (2017) compared fragment of the VP2 gene sequence of AMDV strains circulating Polish farmed mink with AMDV-G strain and identified 16 nt substitutions between nts 3395 and 3827, in which transitions occurred in six positions. Ten ambiguous (S, R, W, and Y) positions were among those 16 nt substitutions. Six of the sixteen nt substitutions were non-synonymous. The presence of an ambiguous position in AMDV isolates circulating in Newfoundland was also documented (Canuti *et al*., 2016). It was suggested that the chronic infection of AMDV in farmed mink

could be a factor favoring the ambiguous positions in the AMDV genome (Canuti *et al*., 2016).

2.6.2. Effect of recombination on phylogenetic analysis

Recombination is one of the mechanisms contributing to virus evolution. Recombination in viruses creates chimeric molecules from parental genomes. The main advantage of recombination for a virus is combining advantageous properties from various genomes into a new one by eliminating deleterious mutations. Therefore, recombination has been considered as an important mechanism for the emergence of viruses with new properties (Lam *et al*., 2010; reviewed by Som, 2014). However, recombination influences the outcome of phylogenetic analyses because it violates the assumption of a single evolutionary history in molecular phylogenetic analyses (Martin *et al*., 2015; Som, 2014). The differences in the evolutionary histories between the exchanged genome regions as a result of the recombination lead to incongruent topologies in inferred phylogenetic trees. The exclusion of the recombinant segments of the genomes from the dataset before phylogenetic analysis or identify the recombination breakpoints and the use of the regions that are not interrupted from breakpoints could prevent the inconsistencies in phylogenies (Martin *et al*., 2015).

Several bioinformatics tools have been developed to detect recombination events, which listed in http://bioinf.man.ac.uk/robertson/recombination/programs.shtml. The widely used method RDP4 is available for free download from http://web.cbio.uct.ac.za/_darren/rdp.html. The first report on the recombination of AMDV was on 1533 bp portion of the VP2 gene, corresponds to the positions 2643–4175, of 15

AMDV isolates (Shackelton *et al*., 2007). The recombination events were detected using six recombination detection programs (Bootscan, Chimeric, GENECONV, MaxChi, RDP, and SiScan) implemented in RDP2 software. Two recombinant regions were identified near nts 543 and 1051. Three ML phylogenetic trees were constructed to interfere in evolutionary histories using the method available in PAUP software (Phylogenetic Analysis Using Parsimony) (Swofford, 2003). The most appropriate model of nt substitution for the tree construction, the GTR+I+T was determined by MODELTEST (Posada & Crandall, 2001a). Bootstrap resampling was performed based on 1000 replicates and nodes with ≥70 % bootstrap support were considered as significant. All three phylogenies showed significant incongruences the clades associated with TH5/SL-3/ADV-G/TR and Utah1/MDPMVT6/Utah1 kit/Far East isolates.

Knuuttila *et al*. (2009) performed recombination analysis on a 336 bp fragment of the NS1 gene, from bp 587 to 922, of 54 AMDV strains using SplitsTree4 software (www.splitstree.org). No significant recombination breakpoint was detected, and it was concluded that recombination analysis using short sequences was not sufficient for detecting recombination events (data not shown). Canuti *et al*. (2016) investigated recombination events in 14 complete coding sequences of AMDV (six from GenBank, one from wild mink, seven from Newfoundland farm mink) using seven methods included in the RDP 4.55 software package (Martin *et al.*, 2015), namely RDP, GENECONV, BootScan, MaxChi, Chimera, SiScan and 3Seq. Eleven recombination breakpoints were detected, and four breakpoint events at nts 903, 1926, 3201 and 3755 were selected based on detection order, probability scores, and phylogenetic evidence for subsequent evaluation for phylogenetic inconsistencies. Five trees were constructed using the genome sections

between the four recombination breakpoints. The additional sixth tree was built using the sequence of the entire VP2 gene for comparison. The trees constructed from sub-portions of the VP2 ORF did not give statistical support for the occurrence of recombination events. It was concluded that the low genetic variability in small sub-portions was not enough for the clear identification of phylogenetic inconsistencies by resolving tree structure. However, two of the NS1 ORF phylogenies and complete VP2 ORF phylogeny resolved the inconsistencies in phylogenetic trees especially strains associated with the strains from China) LN1, LN2, LN3), Newfoundland (WM25), and Utah1.

2.6.3. Phylogenetic analysis

2.6.3.1. Uses of phylogenetic analysis

Phylogenetic analysis using phenotypic or genetic parameters have been used to answer several types of biological questions, such as evolutionary relationship among taxa, the origin, and spread of microorganisms, migration patterns of living organisms, to name a few (reviewed by Yang and Rannala, 2012). Molecular phylogenetic analysis has been extensively used in virology, including AMDV, to identify the origin of virus outbreaks, virus transmission or evolutionary relationships among viral isolates. For instance,

Ryt-Hansen *et al*. (2017a) used phylogenetic analysis to trace the source of a recent AMDV outbreak in Denmark, where many farms became infected in 2015. In this study, a total of 137 mink from 73 farms were sampled during the outbreaks, and 13 retrospective samples from five farms were used. Phylogenetic analyses were performed using a 328 bp segment of the NS1 gene, and that of the sequences of Danish isolates from GenBank (number of sequences were not reported), to track the source of the outbreak. It was shown

that each outbreak cluster originated from a single source, including feed sources, and the geographical locations of outbreaks were highly correlated to the clustering of the corresponding isolates. It was also concluded that partial NS1 sequences were useful for grouping sequences into major clusters but were not adequate to track virus spread within clusters due to the high levels of intra-cluster sequence identity.

Phylogenetic analyses have been widely used for finding the genetic relationship among strains of AMDV. For example, Kowalczyk *et al*. (2018) studied the relationships among AMDV isolates from two mink farms in Poland using 1138 bp fragment of the VP2 gene of 10 mink from each farm. The phylogenetic analysis included 21 global isolates from GenBank and revealed that the AMDV isolates from these farms grouped in the same cluster with three of the four Irish strains, the Pullman strain, Bel 2 from Russia and the Beijing strain from China but were far from the isolate from Finland. The phylogenetic relationships between nine Polish AMDV isolate from two farms and 37 global AMDV isolates from China, Ireland, Russia, Canada, USA, Belarus, Germany, and Finland were determined using a 203 aa fragment of the VP2 protein (Jakubczak *et al*., 2016). The Polish strains occupied one of the five clades in two different branches, each included individual from one of the farms, and the same clade contained the Irish and all Chinese strains. They suggested that the closer relationships between Polish and Chines strain were because of their common ancestry, which is logical because American mink were imported from Europe to China (Li *et al*., 2012). A 336 nt fragment of the NS1 gene from 14 isolates from five farms in Finland and 40 global AMDV sequences from GenBank were subjected to phylogenetic analysis to identify the origin of the isolates and their distribution (Knuuttila *et al*., 2009). The Finnish samples included four previously sequenced isolates. All samples

were classified into three clades, and Finish samples were scattered into all three clades. They concluded that AMDV strains did not group according to the year of isolation, pathogenicity or the geographical origin. The distribution of AMDV between farmed mink and wild animals has been studied by phylogenetic analysis. Knuuttila *et al*. (2015) investigated the genetic relationship of AMDV isolates circulating in free-ranging mustelids (17 feral mink, 2 badgers, one polecat) from nine regions in Finland, ten wild mustelids (2 feral mink, 4 badgers, 3 pine marten, one polecat) from Estonia, 12 farmed mink from Finland and five from Estonia. Phylogenetic analysis was performed using the 373 nt fragments of the NS1 gene. The 47 sequences were classified into five clades. The AMDV sequences from farmed mink were clustered in four of the five clades. Two of the five clades contained the majority of the wild mink and one clade comprised of AMDV sequences from wild mustelids with the exception of wild mink. It was concluded that some of the free-ranging mustelids in Finland carried AMDV isolates similar to those present in farmed mink. Persson *et al*. (2015) studied the phylogenetic relationships between 10 free-ranging mink samples from Sweden and 74 AMDV sequences from GenBank, including 22 farmed mink samples from Sweden. The analysis was based on a 374 bp fragment of the NS1 gene. Eighty-two of the sequences clustered into four groups, but two of the sequences from wild mink which originated from the north of Sweden made an additional cluster. The results indicated that closely related AMDV strains existed in farmed and free-ranging mink in Sweden. It was also concluded that the two sequences that did not group with others were not related to the previously described global AMDV strains including the Swedish strains. In another study, to find genetic changes that had to happen in the AMDV genome over time, samples from 18 mink from a single farm in Newfoundland,

Canada, were collected in 2007 and 2014 (Canuti *et al*., 2016). Extracted DNA samples were cloned, and 127 clones were sequenced using a 483 bp fragment of the NS1 gene and were subjected to phylogenetic analysis. It was concluded that mink sampled in 2014 had higher genetic variability than those samples collected in 2007. Phylogenetic analyses have been used to answer other types of questions in a variety of viruses. For example, phylogenetic analysis was perform using the gene coding for viral enzyme *pol* to find the efficacy of the anti-viral drug for human influenza virus (HIV), because a mutation in *pol* has shown to be resistant to antiviral drugs (reviewed by McCormack and Clewley, 2002).

2.6.3.2. Target genome for the phylogenetic analysis

Usually, small segments of viral genomes, including AMDV, have been used in phylogenetic analysis, primarily to save sequencing cost. Table 2.5 summarises the sizes and the locations of AMDV fragments that have been used in 24 phylogenetic analyses which have been performed in 18 studies. Table 2.5 shows that only two phylogenetic analyses were conducted using aa sequences. Partial NS1 gene was widely used (13 of the 24), compared to the partial VP1 (one report), partial or complete VP2 gene (five), and nearly complete genome (one report). The NS1 region has been used more often because it is more variable than VP2 segment, as shown in the current (Section 2.6.3.2) and other studies (Gottschalck *et al*., 1994; Hagberg *et al*., 2016), and is thus more suitable for phylogenetic analysis of closely related viral isolates. The largest number of analyses (11) used NS1 fragments between 321 and 374 nt long, located between nts 578 and 951, except two which were located between nts 1854 and 2230. These regions were used primarily because they included HVR, as shown in Sections 5.4. Leimann *et al*. (2015) found low

confidence levels at the nodes of the phylogenetic tree using the VP2 fragment (526 bp), but the combined NS1 and VP2 fragments improved the confidence levels. Hagberg *et al*. (2017) compared the nearly complete AMDV genome with 400 bp windows (179 windows) and calculated the relative resolution of each window using (i) the number of internal nodes in the tree (the larger the better) and (ii) the number of fully resolved internal nodes (those which have only two branch). They found that the best short window had 60% of the relative resolution of the nearly complete genome, those with high relative resolutions were located at the 3' end of the NS1 gene, and short fragments, particularly the VP2 segments, did not generate accurate phylogenetic trees. It was suggested that phylogenetic tree construction using nearly complete genome sequences was more accurate than the partial NS1 gene sequences for elucidating transmission pathways of AMDV between mink farms. The use of nearly complete genome sequences resolved the polytomy (branching out from single node) associated with four farms with high node confidence levels. Different segments of a genome may generate different phylogenetic trees because of differences in the degree of variability, which is the result of differences in mutation rate, as well as recombination. Streck *et al*. (2011) conducted global phylogenetic analyses and nucleotide substitution rate analyses for complete NS1 and VP1 genes and partial VP1 gene of porcine parvovirus (PPV) to investigate the appearance of new PPV isolates worldwide.

Table 2.5. Summary of the published phylogenetic analyses of the AMDV strains

Genomic location	Genomic region[a]	Size	Number of isolates	Country	Outgroup[b]	Reference
2753- 3973	Combined 2 VP2 fragments	1138 nt	41	Poland	RFAV	Kowalczyk *et al*., 2018
98-4467	Near complete genome	4369 nt	61	Denmark	GFAV	Hagberg *et al*., 2017
578-951	Partial NS1	328 nt	150	Denmark	.	Ryt-Hansen *et a*l., 2017a
578-951	Partial NS1	328 nt	564	Denmark	.	Ryt-Hansen *et al*., 2017b
3519-3972	Partial VP2	453 nt	34	Poland	.	Jakubczak *et al*., 2017
1207-1690	Partial NS1	483 nt	131	Canada	RFAV & GFAV	Canuti *et al*., 2016
602-922	Partial NS1	321 nt	179	Canada	.	Canuti *et al*., 2016
1859-2208	Partial NS1	348 nt	56	Canada	.	Canuti *et al*., 2016
2949-3228	Partial VP1	280 nt	128	Canada	.	Canuti *et al*., 2016
2753-3562	Partial VP2-HVR	203 aa	46	Poland	.	Jakubczak *et al*., 2016
578-951	Partial NS1	374 nt	84	Sweden	.	Persson *et al*., 2015
587-922	Partial NS1	328 nt	151	Estonia	.	Leimann *et al*., 2015
587- 922 + 2730-3255	Partial NS1+ Partial VP2	871 nt	40	Estonia	.	Leimann *et al*., 2015
1854-2230	Partial NS1	373 nt	47	Finland	GFAV	Knuuttila *et al*., 2015
1208-1637	Partial NS1	429 nt	16	Poland	MVM	Reichert *et al*., 2014
2336-4474	VP2 near complete	1941 nt	19	China	MEV	Sang *et al*., 2012
578-951	Partial NS1	370 nt	84	Denmark	.	Jensen *et al*., 2012
2406-4349	VP2 complete	1944 nt	46	Denmark	.	Jensen *et al*., 2012
2668-3412	Partial VP2	785 nt	11	China	.	Li *et al*., 2012
601-922	NS1 HVR	322 nt	127	Canada	.	Nituch *et al*., 2012
2725-3255	VP2 HVR	531 nt	75	Canada	.	Nituch *et al*., 2012
578-951	Partial NS1	328 nt	51	Denmark	.	Christensen *et al*., 2011
587-922	Partial NS1	336 nt	54	Finland	B19 & PVP	Knuuttila *et al*., 2009
587-922	Partial NS1	112 aa	40	Sweden	.	Olofsson *et al*., 1999

[a]HRV=Hypervariable region, [b] RFAV-Raccoon dog amdo virus, GFAV- Gray fox amdo virus, MVM- Minute virus of mice, MEV- Mink enteritis virus, PVP-Porcine parvovirus.

The phylogenetic analyses suggested that partial VP1gene can be used to represent the full length of the VP1 gene because of the similarities in the topology of the two phylogenies. However, topology differences were observed in phylogenies constructed using NS1 and VP1 data sets. The calculated substitution rates for the NS1 and VP1 genes were 5×10^{-4} and 3×10^{-4} per site per year, respectively, and it was concluded that the differences in substitution rates could be a reason for the phylogenetic incongruences between NS1 and VP1 in PPV.

2.6.3.3. Construction of phylogenetic trees

The phylogenetic tree construction involves a few steps, namely acquiring the sequences, MSA, estimate the phylogenetic tree using the selected method, and graphical representation of the inferred tree (Hall, 2013). An accurate sequence alignment, i.e., the correct identification of homologous positions of nts or aa residues, is essential for constructing a tree that can accurately show the evolutionary relationships among sequences that have a common ancestor (Hall *et al*., 2009). A major issue in MSA is the presence of indels, resulting in sequences with different lengths, and MSA software inserts gaps during alignment. The issue of gaps in aligned sequences has two parts: the number of gaps that are generated by the MSA programs, and the method that phylogenetic tree construction programs handle the gaps. The MSA programs insert gaps in the alignments to maximize the number of homologous positions of nts or aa residues. This is performed by assigning gap opening and gap extension penalties to keep the number of gaps to a minimum (Gotoh, 1982). The default setting for penalty scores available in all MSA

programs are often quite adequate, and the users have a choice to adjust those penalties to improve the alignment quality (Hall *et al*., 2009; reviewed by Hall and Barlow 2006). There are more than 50 phylogenetic analysis software programs (https://en.wikipedia.org/wiki/List_of_phylogenetics_software), using either distance-based methods, such as Neighbor-Joining (NJ) or character-based methods, including maximum parsimony (MP), Bayesian phylogenetic inference, maximum likelihood (ML) (Yang and Rannala, 2012). Each software treats indel-induced alignment gaps using one or more of several methods, includes a binary character (BC), missing data (MD) and maximum likelihood criterion (MLε). The combination of different phylogenetic tree construction software and gap treatment methods created a complex situation, making it difficult to identify a single best software. Attempts have been made to shed light on a segment of this complex situation by computer simulation. For instance, Dwivedi and Gadagkar (2009) used computer simulation to investigate the effects of two alignment gaps, insertion to deletion of 1:1 and 1:3, three gap treatment methods (MD, BC and ML) and four phylogenetic analysis software (NJ, MP, likelihood (PhyML) and Bayesian) on the accuracy of phylogenetic analysis. They found that:

i- there was a strong relationship between the amount of gap in the data and the level of phylogenetic accuracy when the alignments contained a large number of gaps,

ii- gaps resulting from deletions contributed more to the inaccuracy of phylogenetic accuracy compared to insertions,

iii- the probabilistic methods (Bayesian, PhyML & "MLε", a method implemented in DNAML in PHYLIP) performed better at most levels of gap percentage when compared to MP and NJ methods, with Bayesian analysis being clearly the best,

iv- methods that treat gapped sites as MD yielded less accurate trees when compared to those that attribute phylogenetic signal to the gapped sites (BC or MLε),

v-The accuracy of the phylogenetic inference depends on alignment gaps created by insertion and deletion,

vi- more gaps have been introduced by all the alignment when the ratio of insertion to deletion was 1:1, compared to the insertion to deletion ratio of 1:3.

The tree construction algorithms use one of the gap-handling methods (BC, MD, or MLε), or provide user-control options. As an example, the widely used phylogenetic tree building software Molecular Evolutionary Genetics Analysis (MEGA) has complete and partial deletion options for handling gaps in constructing the phylogenetic tree (Kumar *et al*., 2016). The phylogenetic tree construction method using the Bayesian approach implemented in MrBayes (nbisweden.github.io/MrBayes/) considers gaps as missing data and ignores them; therefore, gaps and missing characters are not contributing to phylogenetic inference (Ronquist and Huelsenbeck, 2005). Dwivedi and Gadagkar (2009) investigated the accuracy of phylogenetic inference under varying proportions of indel-induced alignment gaps for three different gap treatment methods (BC, MD, ML) implemented in four different phylogenetic tree constriction methods (PhyML, MP, NJ and Bayesian analysis) by computer simulation. The results indicated that all methods performed well, showing between 90 to 100% accuracy when the number of gapped sites per sequence was 20% or lower. The analysis also indicated that the probabilistic methods, particularly Bayesian analysis, were more accurate than other methods when the alignment had a higher percentage of gaps. In addition, compared to MD method of treating gaps. a much higher accuracy was observed when the gaps were coded separately as in the BC

treatment in conjunction with the Bayesian and MP methods (Dwivedi and Gadagkar, 2009). It could be concluded that treating gaps and phylogenetic analysis programs and their effects on accurate interpretation of evolutionary relationship among sequence are complex subjects which are still matters of debates.

Phylogenetic trees are constructed to either show evolutionary relationships among sequences or just the relationships among sequences regardless of their ancestry or the direction of evolutionary changes. In order to investigate the evolutionary relationships among taxa, the sequence of a distant relative of the sequences of interest (outgroup) is needed. Using an outgroup is the most widely used method for constructing rooted trees, and the outgroup could be a virus from another host species or closely related viral species (Hess and DeMoraes Russo, 2007; reviewed by Wohl *et al*., 2016). Therefore, the outgroup is the oldest (last) common ancestor of the taxa included in the tree (in-group) and shows the order of descent from the root toward the tips (McCormack and Clewley, 2002; Williams, 2014).

Proper selection of outgroup is essential as it can affect the topology or the shape of the tree. Long branches in the phylogenetic tree is resulted when the selected outgroup is distantly related to the in-group (Hess and DeMoraes Russo, 2007). Therefore, prior knowledge of the outgroup for the set of sequences being analyzed is necessary and is considered as a limitation of the rooting method. Members of the parvovirus family, including RFAV-Raccoon dog amdo virus (RFAV), GFAV, minute virus of mice (MVM) and mink enteritis virus (MEV) have been used as outgroups in phylogenetic tree construction for AMDV (Table 2.5). The use of two different outgroups in the same tree RFAV and GFAV, PVP and B19 has also been reported.

There are two disadvantages of using outgroups as the rooting method. First, the most important failure for the outgroup method is Long-branch attraction (Sanderson and Shaffer, 2002), where a wrong position of the root is identified due to the attachment of the long-branched outgroup to the other long branches in the tree. Second, error arises due to the differences in nt composition between outgroups and in-groups (Tarrío *et al*., 2000). In this situation, the outgroup is clustered with in-groups based on sequences compositions rather than on their evolutionary relationships. When any of these errors occur, outgroup rooting is not an accurate option for rooting (Holland *et al*., 2003; Tarrío *et al*., 2000). An alternate rooting method is the mid-point rooting (MPR), in which the root is set at the mid-point between the two most divergent sequences in the data set. The main assumption in MPR was that the two most divergent taxa are evolved at the same rate (Nei and Kumar, 2000). Hess and De-Moraes Russo. (2007) conducted an empirical study to evaluate the efficiency of MPR method in retrieving the correct root. In his study, MPR method was carried out for 50 selected data sets from 33 papers by eliminating outgroups. The efficiency of MPR method was then compared in three categories with different root consistencies: (i) single outgroup with correct root, (ii) data sets with multiple outgroup taxa that showed inconsistency in root position, and (iii) data sets with multiple outgroup taxa in which root position was consistent. Phylogenetic analyses were conducted by MEGA 2.1 (Kumar *et al*., 2001) using NJ method. The MPR method displayed a high success rate of retrieving a root in the three data sets in 54%, 82%, and 94% success rates, respectively. It was concluded that the MPR could be used as an alternative rooting method, and it could be adopted by default when outgroup rooting is not straightforward. Stavrinides and Guttman, (2004) performed midpoint rooting phylogenetic analysis to

establish the evolutionary relationship of the Severe Acute Respiratory Syndrome (SARS) virus. There is no report on MPR method for the AMDV. An unrooted tree is used when a proper outgroup cannot be identified or when only the relationships among the taxa relative to one another are required (Williams, 2014; Wohl *et al*., 2016). There are seven reports where unrooted phylogenetic trees were constructed for finding genetic diversity and epidemiology studies (Table 2.5). Software with varying degrees of complexity is available for phylogenetic analyses, which are either distance-based and character-based methods. Examples of the distance-based methods are Unweight Pair Group Method with Arithmetic Mean (UPGMA) and NJ. The main advantage of distance-based methods is their computational efficiency. The neighbor-joining method has widely been used in analyzing large data sets with high levels of sequence homology. The distance-based methods are poorly performing when gaps exist in the sequence alignment. Therefore, the distance-based method is not suitable for analyzing data set with highly distance sequences (Yang and Rannala, 2012).

2.6.3.4. Selection of computer programs and methods used in phylogenetic tree construction

There are three different approaches in the character-based methods, namely maximum likelihood (ML), maximum parsimony (MP), and Bayesian (Drummond *et al*., 2003; McCormack and Clewley, 2002; Yang and Rannala, 2012). The MP is the simplest method and is recommended for tree constructions with well conserved and closely related sequences. The ML analysis has gained popularity due to the increased computational power in recent years and the inclusion of realistic models for nt substitutions.

Consequently, the ML method is more popular than the MP method (McCormack and Clewley, 2002; Yang and Rannala, 2012). The ML analysis has implemented in different software packages including PHYLIP, PAUP, PhyML MOLPHY, MEGA and RAxML (Yang and Rannala, 2012). The Bayesian method was developed after the ML method and implemented in Bayesian evolutionary analysis sampling trees (BEAST), and the MrBays packages. The MEGA software is popular among the biologists for analyzing viral sequences, including five reports on AMDV analysis (Table 2.6), because of its user-friendly environment and accessible to analyze data using both distance-based and character-based methods. The MrBays program is gaining popularity, and five reports have been published on AMDV phylogenetic analysis using this program (Table 2.6).

The reliability of nodes in a phylogenetic tree can be measured in two ways: bootstrap methods and the Bayesian posterior probability distribution method. Discrepancies in the reliability of nodes exist between nonparametric bootstrap proportions and Bayesian posterior probabilities. Erixon *et al*. (2003) investigated the reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetic analysis. Their study indicated that bootstrap frequencies for ML estimates and Bayesian posterior probabilities are not equal measures of confidence and that Bayesian posterior probabilities had higher confidence estimates than corresponding bootstrap frequencies. It was concluded that posterior probabilities and bootstrap supports can be considered as the upper and lower boundaries of confidence limits, respectively, as the measures of node reliability. Alfaro *et al*. (2003) also compared the performance of Bayesian posterior probabilities and bootstrapping in assessing phylogenetic confidence. Large variations of confidence were observed between the two methods in regions of the tree where internodes were short. The

Bayesian posterior probabilities generated higher confidence than bootstrap values, particularly on short internodes. Therefore, it has been suggested that the bootstrap estimation has fewer chances to strongly supporting false node reliability. For that reason, the bootstrap method can be used as a conservative measurement of phylogenetic accuracy. It should be noted that the computational time for software that calculates posterior probability distribution measurements is much shorter than the software that calculates bootstrap values (Murphy *et al*., 2001). Both posterior probabilities and bootstrap frequencies have been used to calculate the reliability of nodes in the phylogenetic analysis of AMDV (Table 2.6).

Table 2.6. List of computer programs and methods used in phylogenetic tree construction of AMDV

Program	Methods of analysis[a]	Method and range of confidence[b]	References
MEGA 3.1	MP	BS 52-99	Knuuttila *et al*., 2009
MEGA 4	Neighbor-joining	BS 42-100	Sang *et al*., 2012
MEGA 5	Neighbor-joining	BS 57-99	Leng *et al*., 2018
MEGA 5	ML	BS 70-100	Knuuttila *et al*., 2015
MEGA 5	Neighbor-joining	BS 74-99	Persson *et al*., 2015
MEGA 6	UPGMA	NR	Jakubczak *et al*., 2016
MEGA 6	ML	BS 73-100	Canuti *et al*., 2016
MEGA 6	Neighbor-joining	BS 97-100	Xi *et al*., 2016
MEGA 6	ML	NR	Jakubczak *et al*., 2017
MrBayes	Bayesian approach	NR	Kowalczyk *et al*., 2018
MrBayes 3.2.3	Bayesian approach	PP 0.51-1	Hagberg *et al*., 2017
MrBayes 3.2.6	Bayesian approach	PP 0.51- 1	Ryt-Hansen *et al*., 2017a
MrBayes 3.2.6	Bayesian approach	PP 0.5-1	Ryt-Hansen *et al*., 2017b
BEAST 1.7.5	Bayesian approach	PP 0.43-1	Leimann *et al*., 2015
Not provided	ML	BS 39-100	Reichert *et al*., 2014
PhyML	ML	BS 70-100	Nituch *et al*., 2012
DNASTAR	Not reported	NR	Li *et al*., 2012
GeneDoc	Not reported	BS 80-100	Christensen *et al*., 2011
PHYLIP 3.5	Jotun Hein	NR	Olofsson *et al*., 1999

[a] MP-Maximum parsimony, ML-Maximum likelihood, UPGMA- Unweight Pair Group Method with Arithmetic Mean,
[b]BS: Bootstrap, PP: Posterior probability. NR: Not reported

2.7. The mutation rate in viruses

A mutation is a change in genetic materials as a result of the replication errors induced by polymerases, DNA editing, or nucleic acid damage (Sanjuan, 2012; Sanjuán & Domingo-Calap, 2016; Sanjuan *et al*., 2010). Mutation rates among the viruses are variable due to the differences in their life cycle, mode of replication, genome size and the secondary structure of their genomes. For DNA and RNA viruses, mutation rates vary inversely with genome size, and average nucleotide substitutions are four times more common than the insertion and deletions (Sanjuan, 2012). The secondary structure of the viral genome has an influence on the mutation rate by pausing the polymerase, which leads

to the increasing chance of strand slippage leading to deletions (reviewed by Duffy *et al*., 2008).

Mutations do not occur randomly across the genome and are influenced by several factors. One factor is the presence of CpG dinucleotide, which are specific targets for methylation of cytosine bases, generating 5- methylcytosine. The 5- methylcytosine has a high tendency to undergo deamination, leading to the transitional mutation of cytosine to the thymine. The extent of methylation of CpG regions of small DNA viruses has been reported (Hoeizer *et al*., 2007; Shackelton *et al*., 2006). Another reason for differences in mutation rate among and within viruses is the base composition. The variation in mutation rate among the closely related viruses and among segments of a genome may be due to differences in the nucleotide sequence of particular regions. Therefore, each viral genome, as well as any segments of each genome, should be viewed differently from those in any other closely related viruses, since genes appear to be evolving at different rates within the virus genomes (Shackelton *et al*., 2006). For example, mutation analysis using global sequences of the porcine parvovirus (PPV) strains revealed that 3×10^{-4} to 4×10^{-4} subs/site/year in capsid genes and 5×10^{-5} subs/site/year for non-structural genes (Streck *et al*., 2011), showing a large difference between the two genes of the same virus.

RNA viruses are the fastest mutating and evolving organism in nature, having high mutation rats between 10^{-2} and 10^{-5} subs/site/year, which is greater than those in DNA viruses (Duffy *et al*., 2008; Sanjuán & Domingo-Calap, 2016). The high mutation rates of the RNA viruses are due to the lack of proofreading mechanism in RNA-dependent RNA polymerase (RdRps) and the reverse transcriptase of the retroviruses. Viruses with dsDNA and ssDNA replicate exclusively by using the host replication machinery, which possesses

high fidelity DNA polymerases (Duffy *et al*., 2008). Among the DNA viruses, the highest mutation rates were reported in ssDNA viruses compared to the dsDNA viruses (Duffy *et al*., 2008). Several recent reports have shown that mutation rates of the ssDNA viruses are as high as those in the RNA viruses, although ssDNA viruses use the host polymerase with high fidelity (Duffy *et al*., 2008). The reasons for the high substitution rates in ssDNA viruses are currently unclear, but it is postulated to be caused by spontaneous deamination while living in the single-stranded state for a long time (Shackelton and Holmes, 2006). Canine parvovirus (CPV) and B19 autonomous parvoviruses, for example, have nucleotide substitution rates of 1×10^{-4} sub/site/year, which is in the range of values for RNA viruses (Shackelton and Holmes, 2006). Estimates of the mutation rates of animal viruses play a significant role in monitoring vaccine efficacy (Streck *et al*., 2011).

2.7.1. Methods of measuring viral mutation rates

Mutation rates remain inconclusive for some viruses due to the inherent difficulty of measuring rare and random events, differences in the methods of measuring mutations and the use of different units of measurements (Sanjuan *et al*., 2010). As an example, mutation rates that are measured per year in natural populations cannot be compared with the mutation rates per replication measured in cell culture. This is because the immune pressure of different hosts or prevailing ecological conditions may have an impact on viral genetic variability, other than the polymerase error rate itself, which is the main cause of mutation in cell culture (Berger, 2015). Also, it is important to distinguish differences between mutation rates and the substitution rates in viruses. The mutation rate is the rate at which genetic errors occur in the genome during replication, which can be expressed as the

number of nt substitutions per nucleotide site per replication. This rate is mainly determined by the polymerases involved in replication and the presence or absence of a post-replicative repair mechanism. Substitution rate is defined as the number of nt substitutions per nucleotide per unit time (days, years or generations) (Drummond *et al*., 2003). Nucleotide substitution rate of a virus depends on the interaction of many factors, such as the host population size, mutation rate, host generation interval and the relative proportion of the mutations that are advantageous, neutral or disadvantageous and thus affecting fitness, therefore, substitution rate is considered as a property of the entire viral population (Drummond *et al*., 2003). For this reason, the mutation rate is calculated as the substitution rate, as it has commonly been reported in the literature (Ogata *et al*., 1991; Willems *et al*., 1993). There are two approaches in measuring mutation rates in viruses, cell culture (*in vitro* assay) or animal model (*in vivo*) (Sanjuan *et al*., 2010).

2.7.1.1. *In vitro* assay measurements

Cell culture systems are used for propagating viruses. Under the controlled laboratory settings, culture tubes are convenient for manipulating growing susceptible cells and obtaining viruses with high titers. Moreover, cell culture systems are more convenient and less expensive than using live animals, and viral proliferation can be examined microscopically (reviewed by legend and Ginocchio, 2007). There is no published information on using cell culture systems for measuring mutation rates in parvoviruses or any DNA viruses. However, the mutation rate of two strains of influenza A and B viruses was measured in the cell culture system using the Madin-Darby Canine Kidney (MDCK) cells by Nobusawa and Satio (2006). In their study, cultured virus plaques were collected

from the influenza A and B viruses after 48 and 72 h pi, respectively. Nucleotides at positions 28 to 860 (833 nt) of the NS gene were analyzed for influenza A and nucleotides at positions 41 to 1072 (1032 nt) of the NS gene were analyzed for influenza B. The mutation rate for the two viruses were estimated as the number of substitutions/site/years by dividing the total number of nt substitutions by the total number of nt sequenced and multiplying by the time from inoculation (8,760/48 and 8,760/72), where 8,760 h were equivalent to the 365 days. Although it is possible to measure the mutation rate of AMDV *in vitro* for the cell culture-adapted AMDV-G (Bloom *et al*., 1994) and SL3 strains (Schuierer *et al*., 1997), no such estimate has been reported.

2.7.1.2. *In vivo* assay measurements

In vivo measurements allow calculating the mutation rate of viruses in nature (Sanjuan *et al*., 2010). Three approaches can be used to measure mutation rates *in vivo*. First, inoculating virus-free animals with a single viral strain, and sequencing the virus in the inoculum and after a specific length of time. This is an accurate but expensive method and has rarely been reported. Williems *et al*. (1993) inoculated one sheep with the retrovirus Bovine Leukemia provirus (BLV) and compared the sequences of two viral genes (envelop glycoprotein (env) and the Long Terminal Repeat (LTR) in the inoculum and the same animal after 1.5 years. A total of 19 clones of the env gene and 16 clones of the LTR genes were sequenced. The *in vivo* rate of mutation was estimated at 0.009% per year for the env gene and 0.034% per year for the LTR region. Second, measuring the changes in viral genome sequences in a single individual over a specific period of time. The problem with this approach is that the time of infection is not known. For example,

nucleotide differences in the H strain of hepatitis C virus were determined using plasma samples collected from a chronically infected human patient in 1977 and then 13 years later, and the mutation rate was estimated at 1.92×10^{-9} subs/site/year (Ogata *et al*., 1991). The third method is to sequencing parts of the genome of a virus sampled from a population of animals and comparing the sequence(s) with published data. This is faster and cheaper than animal inoculation but suffers from the fact that the length of time during which nucleotide variation has occurred cannot be estimated. This method has been used for estimating substitution rates of the canine parvovirus (Shackelton *et al*., 2005) and porcine parvovirus (Strech *et al*., 2011).

2.8. Virus genome sequencing

Viral genome characterization by nucleic acid sequencing is rapidly expanding, and emerging technologies are likely to have a significant impact in the future (Ladner *et al*., 2014; Wohl *et al*., 2016). Ideally, sequencing technology should be fast, accurate, easy to operate, and cheap (Wohl *et al*., 2016). However, each viral family/species has its own challenges in sequencing associated with the genome size, secondary structure and GC content (Ladner *et al*., 2014). Two methodologies, Sanger sequencing and next-generation sequencing (NGS), are currently applied to both fragments and full-length viral genome sequencing.

Direct sequencing of polymerase chain reaction (PCR) products of the viral genome has commonly been used, whereas PCR products were cloned in some cases (Wang *et al*., 2014). The reason for cloning is to obtain better sequences of to identify sequences of individual viral isolates in cases where an animal is infected with more than

one virus, and thus ambiguous nucleotides are observed in the sequence on ssDNA viruses. A limitation of this technique is that the cloning of amplicons into bacteria before sequencing is a tedious process, and only a relatively limited number of clones can be sequenced due to the high cost. Direct PCR amplification of the entire genome is not very feasible, even for viruses with small genomes, and many sets of primers need to be designed and optimized to cover the entire genome of a virus. Because of the high sequencing cost, AMDV sequencing has been focused on small parts of the genome, usually shorter than 500 nt, which can be PCR amplified and sequenced in one run (Olofsson *et al*., 1999; Knuuttila *et al*., 2015). Sequencing short segments of a genome do not accurately measure variability in the entire genome, and longer sequences are required to gain a better view of the diversity and evolution of AMDV isolates. One problem with the Sanger sequencing is the incorporation of incorrect nucleotides by the Taq DNA polymerase during PCR amplification, which could be as high as the 10-fold difference among different polymerase enzymes (reviewed by Eckert and Kunkel, 1991). The rate of incorrect incorporation of nucleotides when using high fidelity enzymes is lower than those with no proofreading ability, yet it is a source of bias in calculating mutation rate.

The use of the Sanger sequencing method has been partially taken over by the NGS in recent years. NGS generates a large number of short (100 nt) sequences with a low cost within a short period of time. NGS has opened new avenues to increase the knowledge of viruses. As rapidly evolving organisms, viruses show a high level of genetic variability. NGS technologies have successfully been applied to characterize the RNA and DNA viruses from clinical samples and the cultured materials (reviewed by Beerenwinkel *et al*., 2012). However, there are limitations on using NGS in viral genome sequencing for

obtaining a high depth of sequence coverage. The first limitation is the need for a large quantity of viral DNA or RNA, and thus NGS is applicable to cases where a large amount of DNA or RNA can be isolated from cell culture or clinical samples, or when the virus genome can be amplified by PCR. Second, the purity of the nucleic acids is critical. Some nucleic acid contaminates should be removed by DNAase or RNAse treatment (Beerenwinkel *et al*., 2012). Third, the AMDV genome contains a homopolymorphic repeat, and NGS techniques have an inherent difficulty with sequencing such repeats. Also, bioinformatics knowledge is required to analyze the sequences generated by NGS.

For the first time, 89.8% of the AMDV-G and Utah1 strains were sequenced by NGS using the Ion Torrent technology after viral genomes were amplified by long-range PCR (Hagberg *et al*., 2016). In another study, Hagberg *et al*. (2017) generated the whole-genome sequences of 48 AMDV samples that originated from 13 Danish farms for the confirmation of inter-farm transmission of AMDV. However, in both studies, a segment of the genome (nt 2470-2520) containing homopolymorphic repeats had very low sequence coverage and was re-sequenced using the Sanger method. Therefore, it is logical to conclude that NGS may not be the preferred method for sequencing of small genomes which contain homopolymorphic repeats, such as AMDV.

CHAPTER 3. RESEARCH HYPOTHESIS AND OBJECTIVES

3.1. Research hypothesis

It is hypothesized that the sequence of AMDV genome will change over time.

3.2. Objectives:

1. To sequence the AMDV isolate circulate in Cape Breton, NS (NS-CB) using the Sanger sequencing method.
2. To characterize the NS-CB.
3. To establish the evolutionary relationship between the NS-CB and the AMDV isolates globally.
4. To determine the mutation rate of AMDV using the sequences of desendents of the NS-CB at 16 and 176 weeks post-inoculation (wpi).

CHAPTER 4. MATERIALS AND METHODS

4.1. The statement of animal care

This study was conducted at the Aleutian Disease Research Centre (ADRC), located in Pictou County, NS. The facility was designed to minimize the chance of viral introduction from outside and escape of the virus to the environment. Animals were always housed in 61.0 cm × 30.5 cm × 20.3 cm cages and had access to feed and water. All protocols were performed according to the standards of the Canadian Council on Animal Care after approval by the institutional Animal Care and Use Committee. Prior to inoculation or blood sampling, animals were anesthetized by intramuscular injection of ketamine hydrochloride (Ketalean, Bimeda-MTC Cambridge, ON, Canada) and xylazine hydrochloride (Rompun 2%, Bayer Health Care) at the dose of 10 mg and 2 mg per kg live weight, respectively. Animals were anesthetized before being euthanized by intracardiac injection of sodium pentobarbital (Euthanyl, Bimedia-MTC) at a dose of 100 mg per kg body weight (Farid *et al*., 2015).

4.2. Preparation of the viral inoculum

Preparation of the viral inoculum has been previously explained (Farid *et al*., 2015). In brief, the source of the virus was a naturally infected mink that tested positive for CIEP, which was randomly selected from a commercial ranch in Cape Breton, NS. The spleen of this mink was harvested aseptically, centrifuged, and a 10% (w/v) suspension was prepared on the ranch (original stock), and three other AMDV-free mink were inoculated at the ADRC with this homogenate. Inoculated animals were euthanized on day 10 (pi), their spleens were harvested, 10% (W/V) homogenates were made, pooled, aliquoted and stored

at -80 °C until use (passage 1). A total of 56 AMDV-free mink were inoculated with this spleen homogenate, and their spleens were harvested and prepared, as explained above (passage 2). This homogenate will be referred to as Nova Scotia-Cape Breton isolate (NS-CB).

4.3. Source of animals and sampling

A total of 122 AMDV-free mink were transferred from Dalhousie University Fur Unit to ADRC and were inoculated intra-nasally with 600 ID50 of the passage 2 in October 2010. These animals were euthanized in January or February of 2011 to 2014, inclusive, and their spleens were harvested and stored at -80 °C in cryovials. Spleen samples from 30 and 21 mink which were euthanized 16 wpi (in 2011) or 176 wpi (in 2014), respectively, were tested for the presence of AMDV DNA by polymerase chain reaction (PCR) and 10 mink from each group which showed strong viral DNA amplifications (explained below) were used in this experiment.

4.4. Laboratory procedures

4.4.1. Tissue preparation and DNA extraction

Tissue homogenates were prepared by cutting approximately 0.25 g of frozen spleen samples from the 20 selected mink into small pieces with sterile scissors, 50 µL of sterile phosphate-buffered saline (PBS, Sigma-Aldrich, St. Louis, MO, USA) was added and homogenized with a battery-operated Kontes grinder (VWR, Mississauga, ON) until it becomes a uniform paste. Another 700 µL of cold PBS was added, and the mixture was briefly vortexed and centrifuged at 16,000×g (Eppendorf 5415C, Hamburg, Germany) for

10 min, and the supernatants were saved for DNA extraction. Viral DNA was extracted from 200 µL of the cell-free spleen homogenate using the Dynabead Saline viral NA kit, according to the manufacturer's instructions (Invitrogen, Burlington, ON). The NS-CB homogenate already prepared was directly used for DNA extraction. Duplicate viral DNA extractions were performed for each of the 21 samples, and the eluted DNA samples were pooled to have a sufficient amount of DNA for PCR amplification.

4.4.2. Primer design and optimization

Originally, eight pairs of overlapping primer were designed on the basis of the sequence of AMDV-G (accession number NC_001662) from GenBank (http://www.ncbi.nlm.nih.gov/genbank) using the Oligo Primer Analysis software, Version 6 (Molecular Biology Insight, Cascade, Co, USA), and were used to amplify the AMDV genome from nucleotide 206 to 4615 (91.8% of the genome), which includes the entire coding region of the AMDV genome and a segment of 3'-UTR. Sequences of some of the primers were changed when sequences of local AMDV isolates became available. Information on the primers is shown in Table 4.1. The sequence of primer sets 60F/R, and 70F/R were already published (Farid, 2013). Primers were synthesized by a commercial company (Sigma, Oakville, Canada) and delivered as a lyophilized powder. Primers were dissolved in Tris-HCl (pH 7.4) to prepare 500 mM stock solutions and stored at -80 °C. The working solutions (10 mM) were prepared using nuclease-free water, and aliquots were stored at -80 ^{o}C.

Table 4.1. Sequences, position and annealing temperature of primer pairs used for the amplification and sequencing of the AMDV genome

Primer name	Primer position[a]	Amplicon size (bp)	Annealing Temp. °C	Primer sequence (5'-3')
140F	206	1213	59.8	ATGGCTCAGGCTCAAATTG
40R	1418			GTCCCATGTCTTTTATAGTTGC
142F	206	724	60	ATGGCTCAGGCTCAAATTGATG
142R[c]	929			CAAYAATSCCACCGTTACC
35F[d]	763	519	58	AGATGGACCTACTAAGCCTTAC
35R	1281			CCGTTTGGTTCTAAAGATAG
45F	1260	718	60	AACTATCTTTAGAACCAAACGG
45R[c]	1977			TACCAAYRGCACTTACCT
49F[d]	1732	714	56.4	TAAAGGCTGTGTGATTGTAA
49R	2145			GCAGTTTTCCGTGTTC
50F	1803	575	60	CTCACGCAGAGCCACTTAAACA
50R	2377			ACCGCAGGGTTAGTTTG
58F	2043	568	58	GACTACTGGACCAAACTCGAGTG
58R	2610			GTACCATTCTAGTAGCGTGACA
52F[d]	2191	759	58	AAGAGACCTCGGCATGA
52R	2949			CAACCTGTAACGACGCAGTT
51F[d]	2195	750	58	GACCTCGGCATGAGTAG
51R	2944			TGTAACGACGCAGTTAAGTCA
55F	2283	632	60	TTAGTTCCTCAGCACTATCCTG
55R	2914			TGCTTGGTAGATGCGTTAC
60F[b]	2771	532	60	GGGTGTATGGATGAGTCCAAA
60R[b]	3302			CCCCAAGCAACGTGTACT
62F	2773	1787	60	GTGTATGGATGAGTCCTAAA
62R[c]	4559			TTAGTATCCTGTATTCMTACAT
65F[c]	3234	556	60	GGCTTRTATGAGTTTAAGAGTA
65R[c]	3789			CTTCTTCCCAYGAGTCT
70F[b]	3646	563	60	ACGAGGTAGACCTATTAGATGG
70R[b]	4208			GCATGTTACTTGGCTTAGTTTG
76F	4028	588	60	GAACAACAACGCTCCATTTGTA
78R[d]	4615			GCATACATTWGGCCATAGT

[a]Positions of the primers are based on AMDV-G. [b] Published sequences. [c] IUPAC ambiguity codes used in degenerative primers are Y = C/T, S = G/C, M = A/C, R= A/G, W = A/T, [d]primer pairs amplified with 2.5% DMSO.

Optimum conditions for PCR amplification were determined using a range of annealing temperatures (8 temperature values from 50 °C to 65 °C) in a gradient thermal cycler (C1000, Bio-Rad, http://www.bio-rad.com/). Dimethyl sulfoxide (DMSO) at 2.5%

and 5.0% (final concentration) were tested as the PCR enhancer (Ralser *et al*., 2006) to improve the amplification of genomic regions which produced faint results. Viral DNA from three known AMDV infected animals was mixed for optimization. PCR products were tested on a 2% agarose gel to assess the amplification success and to check for the presence of secondary bands. DNA size markers were run on the gel to verify the correct amplicon sizes. The temperature that produced a single bright band was used as the proper annealing temperature for the primer set (Table 4.1).

4.4.3. PCR amplification of AMDV DNA

PCR amplification was performed in 15 µL final volumes. The composition of the PCR cocktail is shown in Table 4.2. Takara La Taq DNA polymerase with 3' to 5' exonuclease activity was selected to minimize the miss-incorporation of nts. Extracted DNA was diluted to 25% using Dynabead's elution buffer. Dilution was found to reduce the background generated by the host DNA. Because of the concentration of AMDV DNA was very low and was mixed with mink DNA, its concentration could not be accurately measured by a spectrophotometer (Farid *et al*., 2015). For this reason, three volumes of extracted DNA (1.5, 2.5 and 3.5 µL) were tested and the optimum DNA volume for amplification of each sample was determined using the primer pair 60F/60R. All PCR reactions were performed along with negative control (no DNA) and positive control. The thermal cycler was programmed at 95 °C initial denaturation for 5 minutes followed by 30 cycles of denaturation at 94 °C for 30 sec, primer annealing temperature for 60 sec, extension at 72 °C for 1.0 min for each 1.0 kb amplicon and 30 sec for small amplicons. The final extension was set for 6 min. PCR amplified products were stored at 4 °C for

immediate use and in -20 °C for long-term storage. The amplified products were tested on 1% agarose gels (Agarose 1, Amresco, (www.amresco-inc.com), stained with ethidium bromide and visualized using Molecular Imager Gel Doc TM XR+ gel imaging system (Bio-Rad). Amplified PCR products with single bright bands were selected for sequencing (Figure 4.1).

Table 4.2. Composition of the 15 µl PCR reaction

Reagent	Stock concentration	Final concentration	µl/reaction
Nuclease-free ddH_2o[a]	-	-	7.95
Tween 20[b]	1%	0.1%	1.5
PCR buffer[c]	10X	1X	1.5
dNTPs[d]	2.5 mM	0.2 mM	1.2
Primer 1	10^4 nM	400 nM	0.6
Primer 2	10^4 nM	400 nM	0.6
Taq polymerase[c]	5U/ µl	0.75U	0.15
DNA	-	-	1.5

[a]Qiagen www.qiagen.com
[b]Fisher Scientific (https://www.fishersci.ca/ca/en/home.html)
[c]Takara La Taq DNA polymerase from Clontech (www.clontech.com), PCR buffer contained 25mM $Mgcl_2$ with final concentration of 2.5 mM
[d]Roch (www.rochecanada.com)

4.4.4. Sequencing of PCR products

Sanger sequencing requires 20-100 ng of high-quality PCR products. PCR products were cleaned by the Qiagen PCR clean-up kit (Qiagen, Toronto, Canada), to remove primer-dimer and excess reagents used in PCR reaction. In cases where there were background amplification or multiple bands (Figure 4.2), the correct size amplicon was cut from the low strength (0.8%) agarose gel and amplified DNA was recovered using QIAquick Gel Extraction kit (Qiagen). In cases where the non-specific amplifications were noticed, eight PCR reactions were pooled prior to gel extraction to increase the concentration of the target band. Visualization of ethidium bromide-stained gels under UV

light leads to DNA damage. For this reason, electrophoresed PCR products were stained with methylene blue, and the correct fragments (Figure 4.3) were excised under white light using a scalpel blade. In both cleanup processes, amplified DNA was eluted with nuclease-free water instead of the elution buffer provided by the manufacturer (Qiagen). Purified PCR fragments were re-tested on 1% agarose gel to check the complete removal of primer dimmer (Figure 4.4), and their concentration was measured by the Nanodrop spectrophotometer (ND-1000, Wilmington, DE, USA). The ratio of the absorbance at 260 and 280 nm (A260/280) was used to assess the purity of PCR products. The A260/280 ratio was maintained between 1.8-1.9.

Sequencing was performed by a commercial service provider (Hospital for Sick Kids, Toronto, Canada, (www.tcag.ca/facilities/dnaSequencingSynthesis,html) using the dideoxynucleotide chain termination method in an Applied Biosystem 3730XL DNA analyzer. Sequencing reactions were prepared by adding PCR products and primers according to the specification provided by the commercial service provider prior to shipping samples for sequencing. Primers for sequencing were prepared by further diluting the working solutions by 50% (5 mM) in nuclease-free water. The final concentration of the sequencing reaction was adjusted to 50 ng and 50-100 ng in a 7 µL volume for fragments less than 1.0 kb, and 1.0 to 2.0 kb, respectively. Each amplicon was sequenced bi-directionally. All short amplicons were sequenced using the amplification primers (Table 4.1), and three sequencing primers (Table 4.3) were used for fragments with longer than 750 bps. Cross-contamination of DNA and PCR products were prevented by using the filtered tip in sample handling. DNA extraction, sample preparation, PCR amplification,

and the amplicon testing was performed in four different laboratories to prevent cross-contamination.

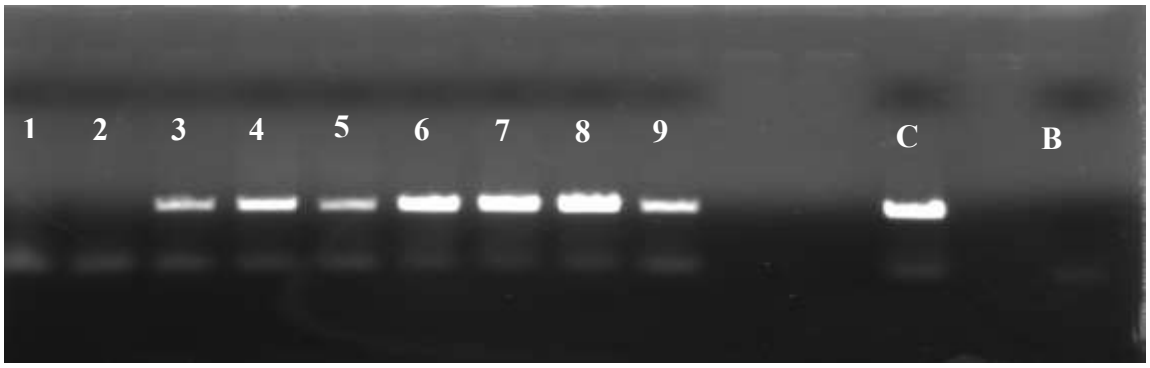

Figure 4.1. Visualization of PCR amplified viral DNA on 1% agarose gel. Lane 1 and 2: amplification failed DNA samples. Lanes 3-9 properly amplified DNA samples, C: Positive control, B: Negative control.

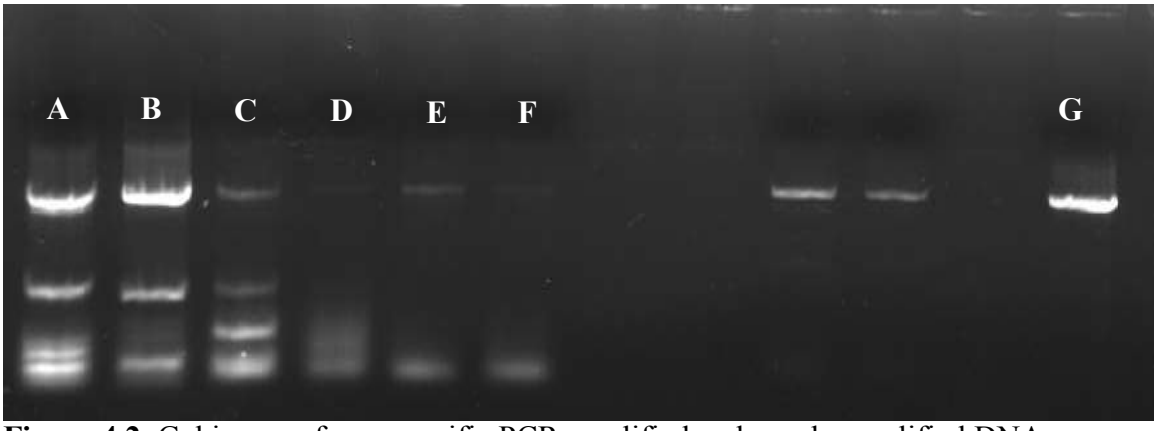

Figure 4.2. Gel image of non-specific PCR amplified and poorly amplified DNA samples. Lane A and B: selected samples for gel purification. Lanes C to F poorly amplified samples, Lane G: maker representing the correct amplicon size.

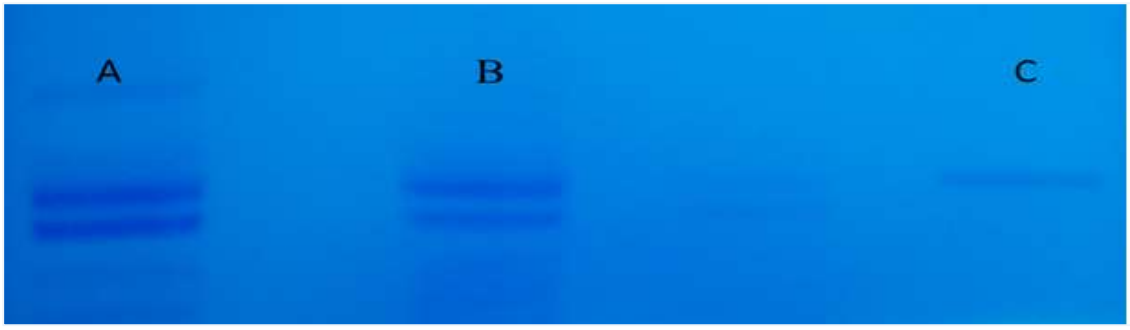

Figure 4.3. Visualization of non-specifically amplified DNA on a methylene blue-stained agarose gel. Lane A and B: Amplified samples with non- specific bands, Lane C: maker

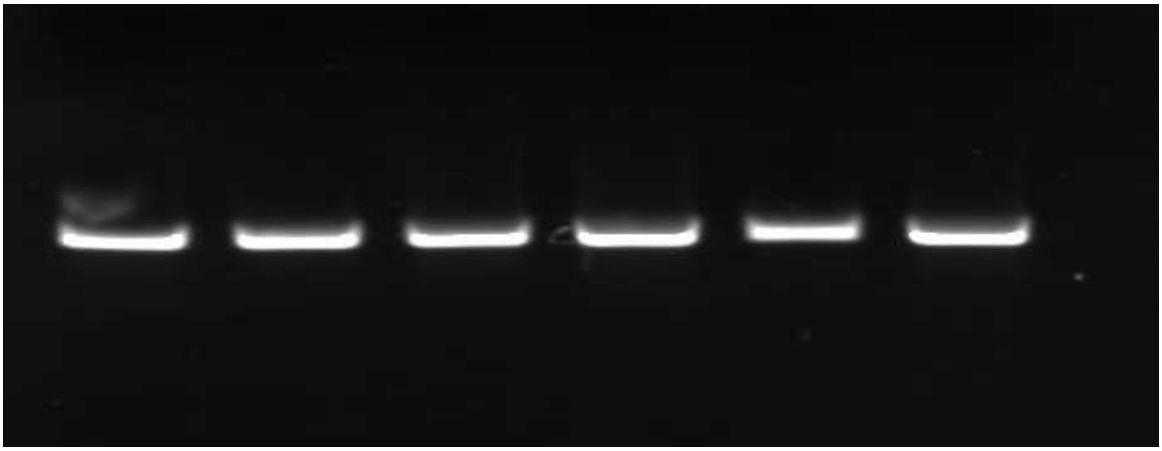

Figure 4.4. Gel image showing complete removal of primer dimer from the amplified amplicons.

Table 4.3. The position and sequence of primers used for sequencing only

Primer	Primer position[a]	Primer sequence (5'-3')
26R	647	AGATGGACCTACTAAGCCTTAC
35F	763	AACCTAAGGATTTTTGAACAT
30R	931	GTCAACAATGCCACCGTTACCAG

[a]Position of primers is based on the AMDV-G sequence.

4.5. Sequence analysis

4.5.1. Sequence editing and assembly

Sequence chromatograms were manually checked using the chromatogram viewing program (Chromas Version 2.6.1, www.technelysium.com.au). Fragments whose forward or reverse sequences were terminated prematurely were re-sequenced. Primer sequences and stretches of sequences with low quality were removed, and miss-called nts (shown as Ns) and low-quality nts (low peaks with high background) were removed. Next, the forward and reverse sequences generated by each primer pair were aligned using the Sequencher software Version 5.4 (http://genecodes.com/). The sequences were manually corrected based on the quality of the electropherograms of the forward and reverse

sequences. Ambiguous positions were identified using the "Call Secondary Picks" option of the Sequencher with 25 % peak height and were retained as the IUPAC codes (http://www.dnabaser.com/articles/IUPAC%20ambiguity%20codes.html) when existed in both directions. The complete genome assembly using overlapping short reads was performed by Sequencher software and consensus sequences were exported as FASTA-formatted files.

4.5.2. Amino acid sequence prediction

Frequencies of the four nucleotides of the NS-CB isolate was calculated by MEGA7 (www.megasoftware.net) (Kumar *et al*., 2016), which ignores the ambiguous nts. The sequence of the NS-CB was aligned with the published AMDV-G sequence using the Sequencher software and the splicing sites, donor and receptor motifs of the three non-structural (NS1, NS2, and NS3) and two structural (VP1 and VP2) genes of the AMDV-G, which have been experimentally predicted (Qiu *et al*., 2006, Table 4.4), were determined. Of the 496 AMDV sequences on GenBank (visited Dec. 2017), 44 which had either near-complete genomes or complete gene sequences were downloaded. The sequences G1, G2 and G3 (acc # KU513985 to KU5139887) which have very similar sequences to AMDV-G, Z18276 which was not annotated, and two patented sequences (acc # JB345270 and JB345272) were not included. The published sequences, along with the corresponding segments of the NS-CB, were divided into eight datasets for further analyses, using the splicing sites shown in Table 4.4. These data sets were (i) the entire coding region (ECR), (ii) the 3'-UTR sequences, corresponding to the 117 nt of the NS-CB, (iii) complete sequence of the NS1 gene, including its intron, (iv) complete sequences of the NS1 CDS,

(v) NS2 CDS, (vi) NS3 CDS, (vii) the unique region of VP1 CDS (VP1u), and (viii) VP2 CDS. The sequences in dataset ii were manually trimmed to have equal sizes.

Nucleotide sequence alignment was performed using MUSCLE (Multiple Sequence Comparison by Log-Expectation) algorithm (Edgar, 2004) implemented in MEGA7. The nt sequences of coding regions were aligned by choosing the "Align Codon" option of the MUSCLE, because it is more appropriate for DNA coding sequences than the direct alignment of nts and avoids frameshift of the sequences due to gap insertion by the software during the alignment (Hall, 2013). All the eight aligned datasets were saved as the MEGA formatted files. The five CDS datasets and the 3'end sequences were opened in the MEGA7, the variable sites were identified (V on the ribbon), and the variable columns were exported to an Excel file (using the Excel icon and selecting variable columns option), and the positions where the NS-CB isolate was different from all other sequences on the Excel file were recorded.

The complete CDS of each of the five genes were manually determined and imported to the ExPASy (Expert Protein Analysis System) software (http://www.expasy.org) to predict aa sequences using the standard genetic codes. The 1st frame (longest) aa sequence of the translated output was selected as the correct one for the input nt sequence. This program assigned an 'X' to aa positions where nts were ambiguous. The possible aa residues at such positions were manually identified and changed to the correct aa when ambiguous nt did not change the aa sequence (synonymous). In cases where the nts at ambiguous positions altered the aa sequence (non-synonymous substitution), the 'X' was not changed. One aa dataset was created for each of the five genes and were aligned using the Muscle using the default settings. The aa residues which

were different between NS-CB and global strains were identified and saved in an Excel file as explained above for the nts.

Table 4.4. Positions and sequences of the exon and intron junctions of AMDV-G

Gene	Nucleotide positions at splicing sites[a]	Sequences at splicing sites
NS1	206……**1961,** **1962**…….**2041** **2042**……2211	ATG……**CTTAG**, **GTAA**……**ACAG,** **AGACT**……AGTAA
NS2	206……**384**, **385**……**2041** **2042**……2207	ATG……**CTAAG,** **GTTA**…… **ACAG**, **AGACT**…… CATGA
NS3	206……**384** **385**……**1736** **1737**……1821	ATG……**CTAAG** **GTTA**……**AAAG** **GCTG**……AAAA
VP1	2204……**2213** **2214**……**2286** **2287**……4349	ATG……**GAGTA** **GTAA**……**TTAG** **TTCC**……ACTAA
VP2	2406……4349	ATG…… ACTAA

[a] Positions are based on the AMDV-G reference sequence (JN040434). The left ORF spans from nt 206 to 2211. The right ORF spans from nt 2204 to 4349. Positions of the beginning and the end of the introns are bolded. Donor positions are marked in blue and acceptor positions are marked in red.

Table 4.5. Published AMDV sequences used for analysis (n=44)

AMDV isolate	Country of isolation	Date posted	Length	GenBank accession #
G	USA	1998	4801	NC_001662
Utah1	USA	1990	4569	Z18276
f-DK-NJ315	Denmark	2017	4369	KU856574
f-DK-NJ321	Denmark	2017	4369	KU856576
f-DK-NJ320	Denmark	2017	4369	KU856573
f-DK-NJ322	Denmark	2017	4369	KU856572
f-DK-NJ37	Denmark	2017	4369	KU856571
f-DK-NJ36	Denmark	2017	4369	KU856570
f-DK-NJ35	Denmark	2017	4369	KU856569
f-DK-NJ343	Denmark	2017	4369	KU856568
f-DK-NJ342	Denmark	2017	4369	KU856567
f-DK-NJ329	Denmark	2017	4369	KU856565
f-DK-NJ328	Denmark	2017	4369	KU856564
f-DK-NJ333	Denmark	2017	4369	KU856566
f-DK-NJ332	Denmark	2017	4369	KU856563
f-DK-NJ327	Denmark	2017	4369	KU856562
f-DK-NJ331	Denmark	2017	4369	KU856561
f-DK-NJ346	Denmark	2017	4369	KU856577
f-DK-NJ345	Denmark	2017	4369	KU856560
f-DK-NJ17	Denmark	2017	4169	KT878961
f-DK-NJ344	Denmark	2017	4369	KU856560
WM25	Canada	2016	4169	KT878961
M228	Canada	2016	4164	KT878960
M195	Canada	2016	4161	KT878959
M173	Canada	2016	4200	KT878958
SL-3	Germany	2016	4718	X97629
LM	China	2017	4809	KJ680280
Beijing	China	2016	4802	KT878963
LN1	China	2012	4543	GU183264
LN2	China	2012	4566	GU183265
LN3	China	2012	4566	GU269892
K	Denmark	2016	2086	X77084
United	Denmark	2016	2086	X77085
MC42.2.1	Canada	2016	2165	KT878963
Pullman	USA	1996	1905	U39014
TR	USA	1996	1917	U39013
FIN05/C8	Finland	2009	1944	GQ336866
Rus09	Russia	2015	1944	KJ174162
Rus11	Russia	2015	1944	KJ174158
Rus14	Russia	2015	1944	KJ174160
Rus17	Russia	2015	1944	KJ174164
Rus19	Russia	2015	1944	KJ174159
Bel1	Russia	2015	1944	KJ174163
Bel2	Russia	2015	1944	KJ174161

Positions are based on the AMDV-G, NC_001662.1.

4.5.3. Identification of hypervariable regions of the nucleotides and major proteins

In order to identify the hypervariable regions in various parts of the nt and aas, first, the nt and aa MSA files were created by the MEGA7, saved in MEGA format and imported to the Entropy-One software (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html) for calculation of the entropy values. The software calculates Shannon's entropy values, defined as $H=-\Sigma^{n}_{j=1} P(s_i)\log_e P(s_i)$, where $P(s_i)$ is the probability of a given nt or aa appearing at position (i), j and n are number of variables and number of sequences respectively (Korber et al., 1994). The indels and mixed aas denoted as (-) and (X), respectively, in the MSA files were considered variable sites in entropy calculation. The datasets which were used included the left and right ORF nt and NS1 and VP2 aa sequences.

The entropy values generated by the Entropy-One software were exported to Microsoft Excel and were used to plot the entropy values. Although Entropy-One generates the same plot, the detailed position of nts or aas is not reported. Entropy values at each nt and aa positions were manually inspected, and variable regions were identified using the following criteria (a) should have at least four consecutive variable nt positions, and (b) should not have more than three positions with zero entropy. Among these variable regions, those which had at least one position with entropy value greater than 0.8 were selected as HVR.

4.5.4. Recombination analysis

4.5.4.1. Detection of recombination breakpoints

Identification of the approximate locations of breakpoints and recombinant sequences were conducted using the RDP 4.97 (Recombination Detection Program) (http://web.cbio.uct.ac.za/~darren/rdp.html) (Martin *et al*., 2015). Recombination analysis was conducted using the ECR data set containing 32 AMDV sequences (Table 4.5), as aligned by MEGA7 and saved as a MEGA formatted file and was imported to RDP 4.97. Next, in the Backup Distribution Plot window, the RDP setting for establishing a preliminary recombination hypothesis under general recombination options were set up as follows: (i) the sequences were considered as linear instead of circular, (ii) statistical significance level was tested at α=0.05 and (iii) Bonferroni correction factor was selected for adjustment of p values for false-positive detection of recombination signals. The default setting in RDP was selected for the data processing options, with the window size to detect the recombination event at 200 bp (Martin *et al*., 2015). Seven recombination detection methods available in RDP4 (RDP, Gene Conversion (GeneConv), BootScan, Maximum chi-square test (MaxChi), Maximum mismatch chi-square (Chimaera), Sister-scanning (SiScan), and 3Seq) were selected in this window for recombination events prediction. In the Run window, the 'Do a Full Explanatory Recombination Scan Using All Methods' selected in the General Option Table' was selected.

4.5.4.2. The accuracy of breakpoints and recombinant sequence identification

In a new overview window, **t**he seven recombination detection methods, with varying degrees of accuracy in identifying the recombination breakpoint positions

appeared. The accuracy of the identified breakpoints was manually performed one event at a time. Once the correct locations of the breakpoints were verified for each method, the recombination events were accepted by pressing the accept button on the RDP4 interface. The recombination events were accepted for further analysis only when the breakpoints were significant by at least three of the seven programs. The entire data set was rescanned for new recombination events, after the preliminary acceptance and or rejecting the recombination events. The accuracy of the identified recombinant sequence was performed side by side phylogenetic tree comparisons. Two NJ phylogenies were constructed for recombination event using (i) the portion of the alignment between the identified breakpoints (the recombinant), and (ii) the remainder of the alignment, using 100 bootstrap replications. The sequence that is contributing to the smaller fraction of the recombinant was considered as a minor parent, while the sequence that is contributing to the larger fraction of the recombinant was considered as a major parent. If the recombinant sequence has moved from the clade with a major parent in one phylogeny to the clade with a minor parent in the second phylogeny, the recombination event was considered as true (Martin *et al*., 2015).

4.5.5. Phylogenetic analysis

The impact of recombination on the topology of the phylogenetic tree was assessed by comparing the phylogenetic trees. The first tree was constructed using the alignment of 32 AMDV sequences, which had ECR (Table 4.5). The second tree was constructed using 24 non-recombinant sequences identified in section 4.5.4.1. These phylogenies were constructed with MEGA7 using the ML method, which requires a nt substitution model to

assess the accuracy of the particular substitution. The Bayesian Information Criterion (BIC) was used to determine the nt substitution model that best fitted for each sub-alignment. Finding the best-fit nt substitution model for two alignments was performed by "Find Best DNA/ protein Models (ML)" option of MEGA7 using the default settings. The model that had the lowest BIC value for each alignment was selected for the phylogenetic tree construction. The lowest BIC values were considered as the true model based on the assumptions made during software stimulation. Mid-point rooted phylogenies were constructed using the "Test/ Maximum Likelihood Tree" option of MEGA7. The site coverage cutoff value was set at 96%, i.e. 4% alignment gaps, missing data, and ambiguous bases were allowed at any position. The inferred tree reliability was tested using the bootstrap method. One thousand bootstrap replicates were performed to support the nodes. Bootstrap values which exceeded 70% were considered as significant (Felsenstein, 1985).

4.5.6. Sequencing of the NS-CB decendents over time

DNA was extracted from 10 spleen samples collected at each of 16 and 176 wpi, sequenced (NS-CB descendants), and sequences were edited and assembled as described in Section 4.5.1. Three nt data sets were constructed. The first and second data sets were constructed for each of the sampling dates (16 and 176 pi). The third data set was constructed for the nearly complete genome sequences generated from both 16 and 176 wpi and was used to identify sequence variation in GT repeat region. A total number of nt substitutions, number and % of ambiguous positions, transitional and transversional substitutions were recorded, and the mutation rate of the NS-CB descendants was determined using the first and the second data sets. The mutation rate was estimated as the

number of substitutions/site/years by dividing the total number of nt substitutions by the total number of nt sequenced and multiplying by the time from inoculation (52/16 and 52/176) as described Nobusawa and Satio, (2006). The substitution rate for the NS-CB was estimated as the number of substitution positions /site/year by dividing the total number of nt substitutions positions by the total number of nt sequenced of NS-CB at the time of inoculation and multiplying by the time from inoculation. The mutation rate for the 20 NS-CB descendants was estimated and t-test was performed using SAS 9.44 for windows. Nucleotide deletions were not counted as substitution and were excluded from total nts, but ambiguous codes were counted as substitutions. In a case where one or two nts changed in an ambiguous position, the number of substitutions was considered as one or two, respectively.

CHAPTER 5. RESULTS

5.1. Genome organization of the NS-CB

The nearly complete genome of NS-CB which was sequenced is 4397 bp long with the nt composition of 24.7 T%, 18.8% C, 37.7% A and 18.8% G, excluding one ambiguous position (G/T) at nt 2696 (nt 2901 of the AMDV-G). This sequence represented 91.6 % of the AMDV genome and covered the entire coding region (NS1, NS2, NS3, VP1, and VP2 genes) as well as 254 bp of the 3'-UTR. The L-ORF encoded the three non-structural proteins; NS1, NS2, and NS3 with 641, 114 aa, and 87 aas, and the R-ORF spanned between nt 1999 and 4144 of the NS-CB, encoding VP1 and VP2 proteins containing 690 and 647 aa residues. The positions of exon and intron boundaries of the five genes of the NS-CB are shown in Table 5.1.

5.2. Nucleotide and amino acid variability

The number and percentage of nt variants in different genome regions of the 26 global AMDV with and without the NS-CB isolate is shown in Table 5.2. The percentage of nt variation in the entire coding region of the 26 global isolates was 14.11%, and percentage variations in other segments of the structural genes were close to each other, ranging between 18.87% in the NS3 gene to 15.90% in the NS3 CDS. Percentage variations in the structural genes were almost twice as large as that of the VP2 gene, and three times greater than that of the VP1u nt. Variability of the 3'-UTR fragments were comparable with those of the structural genes (8.47%). Adding the NS-CB isolate slightly increased nt variability, from 0.10% for the NS1 CDS to 1.69% for the 3'-UTR.

Table 5.1. Nucleotide positions for the exon and intron junction for the NS-CB

	Nucleotide positions of exon and intron junctions[a]				
Protein	NS1	NS2	NS3	VP1	VP2
AMDV strain					
NS-CB	Join (1 ....1756, 1837…2006)	Join (1..179, 1837….2002)	Join (1…179, 1532…1616)	Join (1999…2008, 2082…4144)	(2201…4144)
AMDV-G	Join (206…1961, 2042…2211)	Join (206…384, 2042....2207)	Join (206….384, 1737….1821)	Join (2204….2213, 2287….4399)	(2406…4349)

[a] Positions are based on the AMDV-G, Reference Sequence: JN040434). The left ORF spans from nt 206 to 2211. The right ORF spans from nt 2204 to 4349. Nucleotides at the beginning of the exons are identified in red. The nucleotides located at the end of the exons are identified in purple. The nt gap between 2 exons were introns.

There were 16 ambiguous positions in the coding region, of which the one at position 214 was present in three isolates from Newfoundland (M228, M195, and M173), 12 others were in the M195 strain (nt, 564, 573, 583, 990, 1427, 1433, 1646, 1654, 1653, 2343, 2414, 2416), two in strain M228 (1359, 1360), and one in NS-CB (nt 2901). There were also two insertions in these sequences, one 'A' at position 2004 in the intron of the NS1 gene of three strains from Newfoundland (WM25, M228, M173) and in the LN1 sequence from China, and one 'T' at position 4357 (3'-UTR) of the LN1 strain. Theses insertion increased the length of the aligned sequences, which was the basis for calculating percentage variable positions. Four isolates (WM25, M228, M173, and LN1) had one nt deletion at position 1977 in the NS1 gene. The glycine track of the VP2 genes was shorter by three nt in one strain from Newfoundland (M173), by 27 nt in a Chinese strain (LN1) and by 39 nts in three strains from Newfoundland (WM25, M228, and M195) (Table 5.3). The ambiguous nts and indels were included as a position in the alignment but were not counted as variable sites by the MEGA7.

The number and percentage of aa variants in different genes of the 29 to 42 global AMDV with and without the NS-CB isolate are shown in Table 5.4. Similar to the nt variations, aa compositions of the non-structural genes (21.83% for NS3 to 26.98% for NS1) were much greater than those of the structural genes (4.65% and 10.66%). Variations in aa composition of the three non-structural genes were between 4.9% for NS2 and 9.07% for NS1 greater than the corresponding nt variations, whereas variations in aa composition and nts of the structural genes were small (-3.87% for VP1u and -0.53% for VP2). The addition of the NS-CB isolate increased aa variability of the structural genes from 0.88% for NS2 gene to 3.45% for NS3 genes, but there was no change for VP1u and a small

increase (0.46%) for VP2 gene. The ambiguous codes in the global AMDV isolates were reported as an 'X' (mixed aa) on GenBank, which were shown as '?' in the MEGA7. There were eight mixed aa positions in the NS1 protein, of which seven (positions 120, 123, 262, 408, 410, 4801 and 483) were located on the M195 strain. The M228 strain had one position at aa 385 of the alignment. The NS2 and NS3 proteins did not contain mixed aa. The M195 strain had two mixed aa on the third and fourth positions of the VP2 protein, and the NS-CB isolate had one mixed aa (L/R) at position 116 of the VP2. The nt insertions were not in the coding regions, thus they did not affect aa compositions, but one, 9 and 13 glycine residues were missing from the glycine track of the VP2 protein in five isolates (Table 5.3).

Table 5.2. Number and percentage of nucleotide variants in different genome regions of 26 global AMDV with and without the NS-CB isolate[a]

Genome region	Alignment length, bp	The number and the % variable nt positions	
		Global AMDV isolates	Global AMDV and NS-CB
Coding region	4145[b]	585 (14.11%)	620 (14.95%)
NS1 CDS	1926	381 (19.78%)	383 (19.88%)
NS1 gene	2007	372 (18.53%)	394 (19.63%)
NS2 CDS	345	61 (17.68%)	63 (18.26%)
NS2 gene	2003	371 (18.52%)	393 (19.62%)
NS3 CDS	264	42 (15.90%)	46 (17.42%)
NS3 gene	1616	305 (18.87%)	323 (19.98%)
VP1u nt[b]	129	10 (7.75%)	11 (8.52%)
VP2 gene	1944	195 (10.02%)	206 (10.59%)
3'-UTR[c]	118	10 (8.47%)	12 (10.16)

[a]The list of the 26 global AMDV isolates is shown in Table 4.5.
[b] VP1u: Unique region of the VP1 gene.
[c]Four isolates (WM25, M228, M173, and M195) had short sequences and were excluded from this analysis

Table 5.3. Nucleotide deletions in the TG repeat region of AMDV isolates

AMDV isolate	Nucleotide position[a]													
	2469	2472	2475	2478	2481	2484	2487	2490	2493	2496	2499	2502	2505	2508
G	GGT	GGT	GGG	GGG	GGG	GGG	GGT	GGT	GGG	GGT	GGT	GGT	GGT	GGT
M173	. . .[c]	. . .	. . **A**	. . .	. . .	. . .	. . .	. . .	-[b]	. . .	. . .	. . .	. . .	. . .
LN1	-	-	-	-	-	-	-	-	-	. . .	. . .	. . .	. . .	. . .
WM25	-	-	-	-	-	-	-	-	-	-	-	-	-	. . .
M228	-	-	-	-	-	-	-	-	-	-	-	-	-	. . .
M195	-	-	-	-	-	-	-	-	-	-	-	-	-	. . .

[a] Positions are based on the AMDV-G sequence (GenBank accession number NC_001662).
[b] A dash indicates the deletion of the nucleotides.
[c] A dot represents nucleotides the same to those in the AMDV-G sequence.

Table 5.4. Number and percentage of amino acid variants in different genes of 26 AMDV with and without the NS-CB isolates[a]

Genome region	No of amino acids aligned	The number and the % variable aa positions	
		Global AMDV isolates	Global AMDV & NS-CB
NS1	641	173 (26.98%)	184 (28.70%)
NS2	114	27 (23.68%)	28 (24.56%)
NS3	87	19 (21.83%)	22 (25.28%)
VP1u	43	2 (4.65%)	2 (4.65%)
VP2	647	69 (10.66%)	72 (11.12%)

[a]The list of the 26 global AMDV isolates are shown in Table 4.5.

5.3. The unique nucleotide and amino acid substitutions

Nucleotide sequence alignment of the NS1 gene (206 and 2211) of 33 AMDV isolates (Table 4.5) and the NS-CB showed one nt deletion at position 1977 of seven isolates (WM25, M228, M173, LN1, LN2, LN3, and LM). One nt insertion was also observed in the non-coding region of five isolates from Newfoundland (WM25, MC42.1.1, M228, M173) and three isolates from China (LN2, LN1, and LM). Eighteen unique nt substitution positions (nts which appeared only in NS-CB at each position in the alignment) were identified in the coding region of the NS1 gene of the NS-CB (Table 5.5). The heights number of nt substitutions (7 of 18) were observed at the 2nd position of the codon and two unique nt substitutions were observed in one codon (348). The majority of the unique substitutions (11 of 18) were those where all 33 global sequences had the same nucleotide, of which seven were transitional substitutions (i.e., pyrimidine to pyrimidine or purine to purine). Seven unique substitutions were in the variable positions (433, 574, 879, 882, 1192, 1249 and 2180). Sixteen of non- synonymous substitutions, led to 14 unique aa substitutions (the aa which appeared only on NS-CB at each position in the alignment) in the NS1 of NS-CB (Table 5.6). Three nt substitutions at codon 76, 123, and 67 were not unique to the NS-CB due to the degeneracy of the codons. Nine of the 14 unique aa residues

were located at positions where all global isolates had the same aa. The CDSs of NS2 and NS3 of the NS-CB isolate were aligned with the corresponding segments of 29 global isolates (Table 4.5), and unique nts and aas substitutions were identified. Because 342 bp long CDS of the NS2 and 261bp long CDS of NS3 share their first 179 nts with the NS1 CDS, the unique nts and aas of the segments that were not shared were identified. Two and three nts unique to the NS1 and NS2 of NS-CB, respectively, were detected in these regions (Tables 5.6 and 5.7). In the NS2 CDS, the A to G substitution at position 1784 of the NS-CB isolate was a non-synonymous, but the nt substitution at position 1795 of the NS-CB resulted in a unique aa residue at position 106 of the NS2 aa sequence (Table 5.7). All the three nt substitutions of NS3 CDS resulted in unique aa substitutions in the NS3 aa sequence (Table 5.8)

Table 5.5. The unique nucleotide variants and their codon positions in the NS1 gene of the NS-CB

AMDV Isolates	Codon number and the nucleotide position[a]																							
	(42)[b]			(74)			(76)			(123)			(167)			(220)			(225)			(226)		
	329[c]	330	331	425	426	427	431	432	433[d]	572	573	574	704	705	706	863	864	865	878	879	880	881	882	883
NS-CB	A	.[e]	.	.	C	.	.	.	T	.	.	T	.	.	T	.	G	.	C	G	.	.	G	.
G	G	A	C	A	A	T	C	A	C	C	A	A	G	A	C	A	A	C	A	A	T	A	C	A
WM25	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
Utah1	.	.	.	.	.	.	.	.	G	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.
United	.	.	.	.	.	.	T	G	G	.	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
MC42.1.1	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	C	.	G	T	.
M228	.	.	.	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
M195	.	.	.	.	.	.	A	.	.	.	R[g]	C	.	.	.	.	.	.	.	C	.	G	T	.
M173	.	.	.	.	.	.	T	G	.	.	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
LN3	.	.	.	.	.	.	T	G	G	.	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
LN2	.	.	.	.	.	.	T	G	G	A	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
LN1	.	.	.	.	.	.	T	G	G	.	.	C	.	.	.	.	.	.	**C**	.	.	.	A	G
LM	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	.	.	T	A	.	A	.
K	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	.	.	C	.	C	A	G
Danish-1[f]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.
Danish-2[f]	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	A	.
BJ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	C	A	G

[a] Positions are based on the AMDV-G sequence (GenBank accession number NC_001662).
[b] Codon number in parenthesis.
[c] The unique nt substitutions of the NS-CB in the conserved region are highlighted.
[d] The unique nt substitutions of the NS-CB in the variable region are marked in red.
[e] A dot represents similar nucleotide to the AMDV-G sequence.
[f] Of the 19 Danish strains, 18 are in Danish-1, and one (fDKNJ34) in Danish-2 (Table 4.5.)
[g] R represents an ambiguous position (A/G)

Table 5.5. Continued

AMDV isolates	Codon number and the nucleotide position[a]																											
	(246)[b]			(329)			(348)			(496)			(526)			(581)			(582)			(628)			(632)		
	941	942	943[c]	1190	1191	1192[d]	1247	1248	1249	1691	1692	1693	1781	1782	1783	1946	1947	1948	1950	1951	1952	2168	2169	2170	2180	2181	2182
NS-CB	G	.	G	.	.	A	C	.	G	A	.	.	.	A	.	T	.	.	.	C	.	.	A	.	A	.	.
G	A	A	C	G	A	T	G	A	C	G	A	T	T	G	T	C	C	T	A	A	C	A	G	C	G	C	A
WM25	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
Utah1	.[e]	.	.	.	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	C	.	G
United	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
MC42.1.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
M228	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
M195	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
M173	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
LN3	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
LN2	.	.	.	.	.	C	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	G
LN1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
LM	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
Danish strains[f]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.
BJ	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.

[a] Positions are based on the AMDV-G sequence (GenBank accession number NC_001662).
[b] Codon number in parenthesis.
[c] The unique nt substitutions of the NS-CB in the conserved region are highlighted.
[d] The unique nt substitution of the NS-CB in the variable regions are marked in red.
[e] A dot represents similar nucleotide to the AMDV-G sequence.
[f] The list of Danish strains used for the comparison are shown in Table 4.5.

Table 5.6. Unique amino acid substitutions of the NS1 protein of the NS-CB

AMDV isolats	Amino acid positions in NS1[a]													
	42[d]	74	220	225	226	246	329	348	496	526	581	582	628	632
NS-CB	N	T	S	R	R	E	E	Q	N	I	S	D	N	T
G	D	N	N	N	T	N	D	D	D	V	P	E	S	A
Utah1	.[b]	.	.	.	.	.	.	.	.	.	.	.	.	.
WM25	.	.	.	H	K	D	.	.	.	.	.	.	.	P
United	.	.	.	H	K	.	.	.	.	.	.	.	.	P
M228	.	.	.	H	K	.	.	.	.	.	.	.	.	P
M195	.	.	.	T	V	.	.	.	.	.	.	.	.	P
M173	.	.	.	H	K	.	.	.	.	.	.	.	.	P
LN3	.	.	.	H	K	.	.	E	.	.	.	.	.	P
LN2	.	.	.	H	K	.	.	E	.	.	.	.	.	P
LN1	.	.	.	H	K	.	.	.	.	.	.	.	.	P
Beijing	.	.	.	T	Q	T	.	.	.	.	.	.	.	P
LM	.	.	.	I	K	.	.	.	.	.	.	.	.	P
K	.	.	.	T	Q	.	.	.	.	.	.	.	.	P
MC42.1.1	.	.	.	T	V	.	.	.	.	.	.	.	.	P
Danish Strains	.	.	.	.	K	.	.	.	.	.	.	.	.	.

[a] Positions of amino acids are based on the AMDV-G sequence (GenBank accession number NC_001662).

[b] A dot represents similar amino acids to the AMDV- G strain.

[c] The list of Danish strains used for the comparison are shown in Table 4.5.

[d] Unique aa residues of the NS-CB are highlighted.

A- Alanine (Ala), D-Aspartic acid (Asp), E-Glutamic acid (Glu), H-Histidine (His), I-Isoleucine (Ile), K-Lysine (Lys), N-Asparagine (Asn), P-Proline (Pro), Q-Glutamine (Gln), R-Arginine (Arg), S-Serine (Ser), T-Threonine (Thr), V-Valine (Val)

Table 5.7. The unique nucleotide variants, codon positions and amino acid substitutions in the unique region of the NS2 CDS of the NS-CB isolate

AMDV isolates	Codon number and the Nucleotide positions[a]						Amino acid[f] position[a]
	(102)[c]			(106)			
	1782	1783	1784[b]	1794	1795	1796	106
NS-CB	.[d]	.	A	.	A	.	**Y**[g]
G	G	A	G	T	G	C	C
Danish strains[e]	.	.	.	.	.	.	.
K	.	.	.	.	C	.	S
LN1	.	.	.	.	C	.	S
M173	.	.	.	.	C	.	S
M195	.	.	.	.	C	.	S
M228	.	.	.	.	C	.	S
MC42.1.1	.	.	.	.	C	.	S
United	.	.	.	.	C	.	S
Utah1	.	.	.	.	C	.	.
WM25	.	.	.	.	C	.	S

[a] Positions of nts and aa are based on the AMDV-G sequence.
[b] Unique nt substitution positions are highlighted.
[c] Codon positions are in parenthesis.
[d] A dot represents similar nucleotide and the amino acid to the AMDV-G strain.
[e] Danish strains are listed in Table 4.5.
[f] C- Cysteine, S-Serine, Y-Tyrosine.
[g] Unique amino acid substitution of the NS-CB is bolded.

Table 5.8. The unique nucleotides positions, codons and the unique aa of the NS3 of the NS-CB

AMDV isolates	Nucleotide positions[a]									Amino acid[f] positions[a]		
	(63)[c]			(75)			(86)					
	1744[b]	1745	1746	1780	1781	1782	1813	1814	1815	63	75	86
NS-CB	A	A	T	T	A	T	A	C	C	**N**[g]	**Y**	**T**
G	G	.[d]	.	.	G	.	G	.	.	D	C	A
WM25	.	.	.	.	.	.	.	.	.	.	.	.
Utah1	.	.	.	.	.	.	.	.	.	.	.	.
United	.	.	.	.	.	.	.	.	.	.	.	.
MC42.1.1.	.	.	.	.	.	.	.	.	.	.	.	.
M228	.	.	.	C	.	.	.	.	.	.	R	.
M195	.	.	.	.	.	.	.	.	.	.	.	.
M173	.	.	.	C	.	.	.	.	.	.	R	.
LN1	.	.	.	.	.	.	.	.	.	.	.	.
Danish strains[e]	.	.	.	.	.	.	.	.	.	.	.	.

[a] Positions of nts are based on the AMDV-G sequence
[b] The unique nt substitutions and positions of the NS-CB are highlighted.
[c] Codon positions are in parenthesis. [d]A dot represents similar nucleotide and the amino acid to the AMDV- G strain. [e] The Danish strains are listed in Table 4.5. [f] A- Alanine, C- Cysteine, D- Aspartic acid, R- Arginine. [g] Unique amino acids of the NS-CB are bolded.

Alignment of 129 nt long sequences of VP1u of the NS-CB with 30 global AMDV sequences (Table 4.5) revealed two unique nt sequences at positions 2357 and 2581 (Table 5.9), and both nt substitutions were non-synonymous. Three nt insertions at positions 2291, 2292 and 2293, and three deletions at positions 2233, 2334 and 2335 were identified in this region of the LM strain from China. When compared with the 42 global isolates, nine unique nt substitutions were identified in seven codons of the 1944 bp long sequence of the VP2 region of the NS-CB (Table 5.10). There were multiple nt substitutions at codons 166 and 373, and seven of the substitutions were unique to NS-CB. Five of the nine substitutions (2422, 2902, 2946, 2996 and 3524) were transitional (Table 5.10). The 'T' of the ambiguous position nt (T/G) at position 2901 of the NS-CB was unique to this isolate,

whereas the other nt was the same as that in AMDV-G. The nucleotide at positions 2996 and 3524 on codon 197 and 373 were synonymous substitution, resulted in five unique aa substitutions in the NS-CB (Table 5.11).

Table 5.9. The unique nucleotide substitutions and codon positions in VP1u of the NS-CB

AMDV isolates	Nucleotide positions[a]					
	(27)[b]			(35)		
	2283	2284	2357[c]	2307	2308	2381
NS-CB	.[d]	.	A	.	.	T
G	A	A	G	G	T	C
WM25	.	.	.	.	.	G
Utah1	.	.	.	.	.	G
SL3	.	.	.	.	.	.
MC42.2.1	.	.	.	.	.	.
M228	.	.	.	.	.	.
M195	.	.	.	.	.	G
M173	.	.	.	.	.	.
LM	.	.	.	.	.	G
Danish strains[e]	.	.	.	.	.	G
Beijing	.	.	.	.	.	.

VP1u: Unique region of the VP1spands from nt 2204 to 2405 of the AMDV-G.
[a] Positions of nts are based on the AMDV-G sequence [b] Codon positions are in parenthesis. [c] Unique nt positions in the NS-CB are highlighted. [d] A dot represents similar nucleotide to the AMDV- G strain.
[e] Danish strains are listed in Table 4.5.

Table 5.10. Codon positions and the unique nucleotide substitutions in the NS-CB of theVP2

AMDV Isolates	Nucleotide positions[a] of VP2																				
	(6)[b]			(93)			(166)			(181)			(197)			(373)			(592)		
	2421	2422[c]	2423	2682	2683	2684	2901	2902	2903	2946	2947	2948	2994	2995	2996	3523	3523	3524	4179	4180	4181
NS-CB	.[d]	T	.	A	T	.	G/T	T	.	A	.	.	.	.	A	.	.	C	A	.	.
G	G	C	T	C	A	G	G	C	A	G	T	T	T	T	G	A	G	T	C	T	A
LM	C	.	.	.	.	A	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Group 1[e]	C	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
FIN05-C8	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Group 2[f]	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Rus11	.	.	.	.	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Utah- 1	C	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

[a]VP2 gene spans nt 2406 to 4349 of the AMDV-G. [b] Codon positions are in parenthesis. [c] The unique substitution positions of the NS-CB are highlighted. d A dot represents the same nucleotide to the AMDV- G strain.
[e] RUS17, TR, Bel2, WM25, MC42.2.1 Beijing, M173, M228, M195, LN1, LN2, LN3, and Pullman. [f] Danish strains listed in Table 4.5, Bel1, Rus09, Rus14, Rus19, and SL3.

Table 5.11. The unique amino acid substitution in VP2 of the NS-CB

AMDV isolates	Positions[a] of the amino acids[b]				
	6	93	166	181	592
NS-CB	**V**[c]	**M**	**V/L**	**I**	**I**
G	A	Q	A	V	L
Rus11	.[d]	R	.	.	.
Group 1[e]	P	.	.	.	.
LM	P	.	T	.	.
FIN05/C8	.	.	.	.	.
Group 2[f]	.	.	.	.	.

[a] Positions of aa are based on the AMDV-G sequence. [b] A- Alanine, I- Isoleucine, L- Leucine, M-Methionine, P-Proline, Q-Glutamine, R-Arginine, T-Threonine, V-Valine. [c] The unique substitutions for the NS-CB are bolded. [d] A dot represents similar aa to the AMDV- G strain. [e] Beijing, LN1, LN2, LN3, WM25, M228, M195, M173, MC42.2.1, TR, Bel2, Rus 17 and Pullman. [f] Danish strains are listed in Table 4.5, Utah, SL3, Rus09, Rus14, Rus 19, and Bel1.

5.4. Amino acid sequence motifs

There are two caspase recognition sites in the NS1 protein. The first one is from aa 224 to 228, with the cleavage site at aa 227 (INTD↓S in AMDV-G), and the second one is from aa 282 to 286 with the cleavage site at aa 285 (DQTD↓S in AMDV-G). These two sequences of the NS-CB were aligned with 33 global AMDV sequences. The aa isoleucine (I) at the first position of the first caspase recognition site (224) was conserved in all AMDV isolates, but aa residues at other positions were variable among the AMDV isolates. The aa residues at positions 225, 226 and 228 of the first caspase recognition site were all unique to the NS-CB isolate, but the aa residue at the cleavage site of the first caspase (position 227) of the NS-CB isolate (D) was shared with the AMDV-G, Utah1, SL3 and 19 Danish strains. All the five aa residues at the second caspase recognize site was conserved in the NS-CB and the global isolates (Table 5.12). There is an additional caspase recognition site from aa 417 to 421 of the VP2 protein, with the cleavage site at aa 417

(DLLD↓G in the AMDV-G). This region of the NS-CB isolate was aligned with that of 42 global AMDV isolates, and all five aa residues were conserved among all AMDV isolates, including three US, five Chinese, five Canadian, seven Russian, and 19 Danish isolates (listed in Table 4.5).

Table 5.12. The amino acid residues at the two caspase recognition sites of the NS1 protein of the NS-CB and 33 global AMDV isolates

AMDV isolates	First caspase recognition site[a]					Second caspase recognition site[a]				
	224[b]	225	226	227[b]	228	282	283	284	285[b]	286
NS-CB	.[c]	R	R	.	D	.	.	.	.	.
G	I	N	T	D	S	D	Q	T	D	S
Group-1[d]	.	H	K	E	G	.	.	.	.	.
SL3/ Utah1	.	.	.	.	.	.	.	.	.	.
MC42.1.1	.	T	V	E	G	.	.	.	.	.
M195	.	T	V	E	G	.	.	.	.	.
LN1	.	H	K	N	G	.	.	.	.	.
LM	.	I	K	E	G	.	.	.	.	.
K	.	T	Q	E	G	.	.	.	.	.
Danish strains[e]	.	.	K	.	G	.	.	.	.	.
Beijing	.	T	Q	E	G	.	.	.	.	.

[a] E-Glutamic acid, G- Glycine, H-Histidine, I-Isoleucine, K-Lysine, N-Asparagine, Q-Glutamine, R-Arginine, S-Serine, T-Threonine, V-Valine. [b] Amino acids located at the caspase cleavage sites are highlighted. [c]A dot indicates the same aa with the AMDV- G. [d]WM25, United, M228, M173, LN2, and LN3. [e]19 Danish strains (Table 4.5.)

The five aa residues of the VP2 protein at positions 352, 395, 434, 491 and 534 which have been reported to be associated with the pathogenicity were compared between NS-CB and 42 global AMDV isolates (Table 4.5). Two of the aa residues at positions 395 and 534 have also been reported to be associated with the host range. The glutamine (Q) at position 395 and the aspartic acid (D) at position 534 were the same in the NS-CB and all other isolates, except in the AMDV-G at position 395, which is an H, and AMDV-G and

RUS11, 14, 19 which have and H at position 534 (Table 5.13). The aa valine (V) at position 352 and histidine (H) at position 434 in the NS-CB were similar to those in all other isolates, except in AMDV-G and SL3 at position 352 and AMDV-G, TR and SL3 at position 434. Comparison of the five aa residues in the NS-CB with those in strains with known pathogenicity (G, Utah1, TR, Pullman, SL3), showed that all five aa residues in the NS-CB were different from those in the non-pathogenic AMDV-G, but were all the same as those in the highly pathogenic Utah1 strain, except at position 491. The NS-CB shared four of the five aa residues with the pathogenic TR strain and mildly pathogenic Pullman strains but shared the same aa residues at only two positions (395 and 534) with the mild pathogenic SL3 strain.

Table 5.13. The amino acids associated with the pathogenicity and the host range in the NS-CB and 42 global AMDV isolates

AMDV isolates	Positions of the amino acids[a]				
	352	395	434	491	534
NS-CB	V	Q	H	D	D
G	I	H	N	N	H
WM25	V	Q	H	E	D
Utah1	V	Q	H	E	D
TR	V	Q	.	D	D
SL3	.	Q	.	.	D
Rus11, 14, 19	V	Q	H	D	.
Rus17	V	Q	H	E	D
Pullman	V	Q	H	E	D
MC42.2.1	V	Q	H	D	D
M228	V	Q	H	E	D
M195	V	Q	H	E	D
M173	V	Q	H	D	D
Group-1[b]	V	Q	H	E	D
FIN05/C8	V	Q	H	.	D
Danish-1c	V	Q	H	.	D
Danish-2[d]	V	Q	H	D	D
Bel1	V	Q	H	D	D

[a] D-Aspartic acid, E-Glutamic acid, H-Histidine, I- Isoleucine, N-Asparagine, Q-Glutamine, V-Valine. [b]Five Chinese isolates, Bel2. [c] Of the 19 Danish strains, 15 are in Danish-1, and four (fDKNJ320, 321, 322, 327) are in Danish-2 (Table 4.5.)

The six aa residues (G, T, G, K, T, L), from aa 435 to 440 of the NS1 protein, have been identified as the ATP-binding pocket for the AMDV. These aas in the NS-CB were compared with those in 33 global AMDV isolates (Table 4.5), and all six were conserved among all isolates. Nineteen aa residues of the VP2 protein, from aa 428 to 446, have been reported to be associated with the immune-complex formation and AMDV pathogenesis. These aa residues in the NS-CB were compared with 42 global AMDV isolates (Table 4.5). Fifteen of the 19 aa residues were conserved in all 42 isolates, but one histidine (H) at position 434 was present in every isolate, except in AMDV-G, which was an asparagine (N) (Table 5.14). The aspartic acid (D) at position 433 was the same in NS-CB, AMDV-G, and several other isolates, but there were asparagine (N) and glutamic acid (E) in a few isolates. All isolates had isoleucine (I) at position 436, except two Chinese isolates (LM & Beijing) which had a methionine (M) at this position. Similarly, the glutamic acid at position 437 was occupied by lysine (K) at position 437 in the isolate Bell from Russia. (Table 5.14).

Table 5.14. The amino acid residues at the immune-reactive region of the VP2 protein of the NS-CB and the 42 global AMDV isolates

AMDV isolates	Positions of the amino acids[a]																		
	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446
NS-CB	.[b]	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
G	S	N	Y	Y	S	D	N	E	I	E	Q	H	T	A	K	Q	P	K	L
WM25	.	.	.	.	.	N	H	.	.	.	.	.	.	.	.	.	.	.	.
Utah1	.	.	.	.	.	N	H	.	.	.	.	.	.	.	.	.	.	.	.
TR	.	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.
SL3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Rus 09, 11 14 &19	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
Rus17	.	.	.	.	.	N	H	.	.	.	.	.	.	.	.	.	.	.	.
Pullman	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
MC42.2.1	.	.	.	.	.	N	H	.	.	.	.	.	.	.	.	.	.	.	.
M228	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
M195	.	.	.	.	.	N	H	.	.	.	.	.	.	.	.	.	.	.	.
M173	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
LN1, 2 & 3	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
LM & Beijing	.	.	.	.	.	E	H	.	M	.	.	.	.	.	.	.	.	.	.
Bel2	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
Bel1	.	.	.	.	.	.	H	.	.	K	.	.	.	.	.	.	.	.	.
FIN05/C8	.	.	.	.	.	E	H	.	.	.	.	.	.	.	.	.	.	.	.
Danish 1[c]	.	.	.	.	.	.	H	.	.	.	.	.	.	.	.	.	.	.	.
Danish 2[c]	.	.	.	.	.	N	H	.	.	.	.	.	.	.	.	.	.	.	.

[a] A- Alanine, D- Aspartic acid, E- Glutamic acid, H-Histidine, I- Isoleucine, K- Lysine, L-Leucine, M-Methionine N- Asparagine, P-Proline, Q-Glutamine, S- Serine, T-Threonine, Y-Tyrosine. [b] A dot represents similar nucleotide to the AMDV- G strain. [c] Of the 19 Danish strains, 12 are in Danish-1, and seven (fDKNJ17, 35, 36, 37, 315, 342, 343) are in Danish-2 (Table 4.5.)

5.5. Entropy and hypervariable regions

5.5.1. Entropy and hypervariable regions of the left ORF

The entropy profile of the nts of the left ORF (from nt 206 to 2211) of 34 AMDV isolates, including the NS-CB, are shown in Figure 5.1. A total of 546 nts were non-zero positions in this 2007 nt long left ORF (27.2%) of which 22 were variable regions, according to the criteria defined in section 4.5.5, i.e. having at least four consecutive variable nt positions and fewer than three positions with zero entropy (Table 5.15). The length of the variable regions varied between four and 17 nts, and the number of consecutive variables nt positions varied between four and seven. Five of the 22 regions (570-575, 610-613, 1282-1286, 2114- 2118 and 2137- 2141) contained no zero entropy. The entropy values for the variable regions varied between 0.13 (one variable nt) and 1.070 (14 variable nts). Nine regions had entropy values greater than 0.8, of which the values in the two regions were greater than 1.0 and were located between nts 845 and 853 and 1427 and 1433. The average entropy values of the variable regions varied between 0.163 and 0.421. Eight of the variable regions were selected as being HVR (HVR1F to HVR8F). These regions were seven to 12 nts in length and had at least one position with maximum entropy value greater than 0.8 (Table 5.15 and Figure 5.1). Five of the eight HVRs were located at the middle of the left ORF, between nts 730 and 1174, and two were located towards the 3' end of the left ORF. The HVR2, located between nt 730 and 735, had the highest average entropy (0.421).

5.5.2. Entropy and hypervariable regions of the right ORF

The entropy profile of the nts of the right ORF (from nt 2204 to 4349) of 42 AMDV isolates, including NS-CB, are shown in Figure 5.2. There were 293 non-zero entropy values in this 2146 nt long segment (13.6%). The entropy values of the right ORF varied between 0.15 (one variable nt) and 1.05 (11 variable nts). Only four variable regions satisfied the criteria presented in section 4.5.5. One region (nt 2472 to 2510) was 39 nt long with no zero entropy, had the greatest maximum entropy (0.876) and the highest mean entropy (0.473) among the four regions (Table 5.16). Two of the variable regions were selected as being HVRs (Table 5.16 and Figure 5.2). These regions spanned nt 2472 to 2510 (HVR1R) and 3101 to 3122 (HVR2R), with one and two entropy values greater than 0.8, respectively, and the greatest mean entropy values.

Table 5.15. Variable regions of the left ORF of the AMDV isolates based on the entropy of nucleotides

Variable region	Nucleotide positions[a]	Length	Number of consecutive variable positions	Number of zero entropy positions	Number of nt positions with selected entropy values		Minimum entropy	Maximum entropy	Average entropy
					>1	>0.8			
1	252-268	17	6	2	0	0	0.130	0.560	0.210
2	**272-280[b]**	9	5	2	0	1	0.219	0.890	0.337
3	426-436	11	5	2	0	0	0.130	0.760	0.296
4	485-491	7	5	3	0	0	0.130	0.458	0.200
5	570-575	5	5	0	0	0	0.130	0.718	0.291
6	610-613	4	4	0	0	0	0.130	0.567	0.317
7	706-714	10	4	3	0	0	0.130	0.673	0.260
8	730-735	6	4	1	0	1	0.130	0.826	0.421
9	**827-838**	12	5	2	0	1	0.130	0.890	0.346
10	**845-853**	9	4	2	1	2	0.130	1.070	0.416
11	**878-888**	11	7	1	0	1	0.130	0.925	0.384
12	**1166-1174**	9	6	2	0	1	0.130	0.870	0.273
13	1198-1206	9	4	3	0	0	0.219	0.620	0.293
14	1239-1247	9	6	2	0	0	0.130	0.560	0.290
15	1282-1286	4	4	0	0	0	0.130	0.637	0.301
16	1323-1328	6	4	1	0	0	0.219	0.673	0.375
17	1331-1342	12	7	1	0	0	0.130	0.772	0.333
18	**1345-1354**	10	5	3	0	1	0.130	0.916	0.329
19	**1393-1399**	7	5	1	0	1	0.130	0.355	0.163
20	**1427-1433**	7	5	1	1	1	0.130	1.032	0.375
21	2114-2118	5	5	0	0	0	0.355	0.458	0.386
22	2137-2141	5	5	0	0	0	0.130	0.622	0.319

[a] Positions are based on AMDV-G sequence.
[b]Hypervariable regions are bolded.

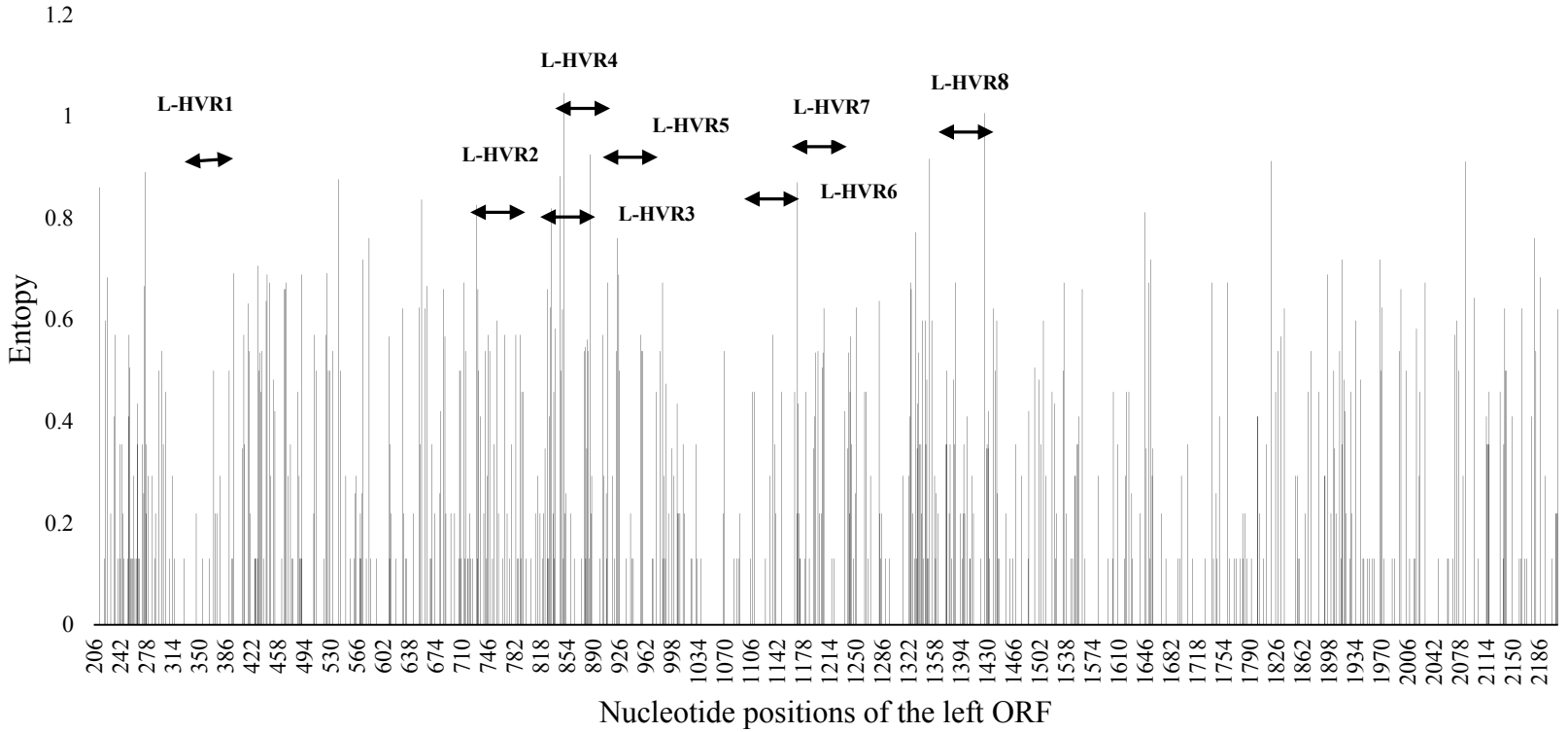

Figure 5.1. Variation of entropy values in left ORF of the AMDV isolates using Entropy-ONE Web tool. Approximate positions of hypervariable positions are identified by arrows

Table 5.16. Variable regions of the right ORF nucleotides of 42 AMDV isolates based on entropy values

Variable region	Nucleotide positions	Length	Number of consecutive variable positions	Number of zero entropy positions	Number of nt positions with selected entropy values		Minimum entropy	Maximum entropy	Average entropy
					>1	>0.8			
1	**2472-2510**[a]	39	39	0	0	1	0.401	0.876	0.473
2	2682-2690	9	6	1	0	0	0.150	0.642	0.319
3	**3101-3122**[a]	22	15	2	0	2	0.150	0.837	0.433
4	3126-3133	8	4	2	0	0	0.510	0.603	0.320

[a] Hypervariable regions are bolded.

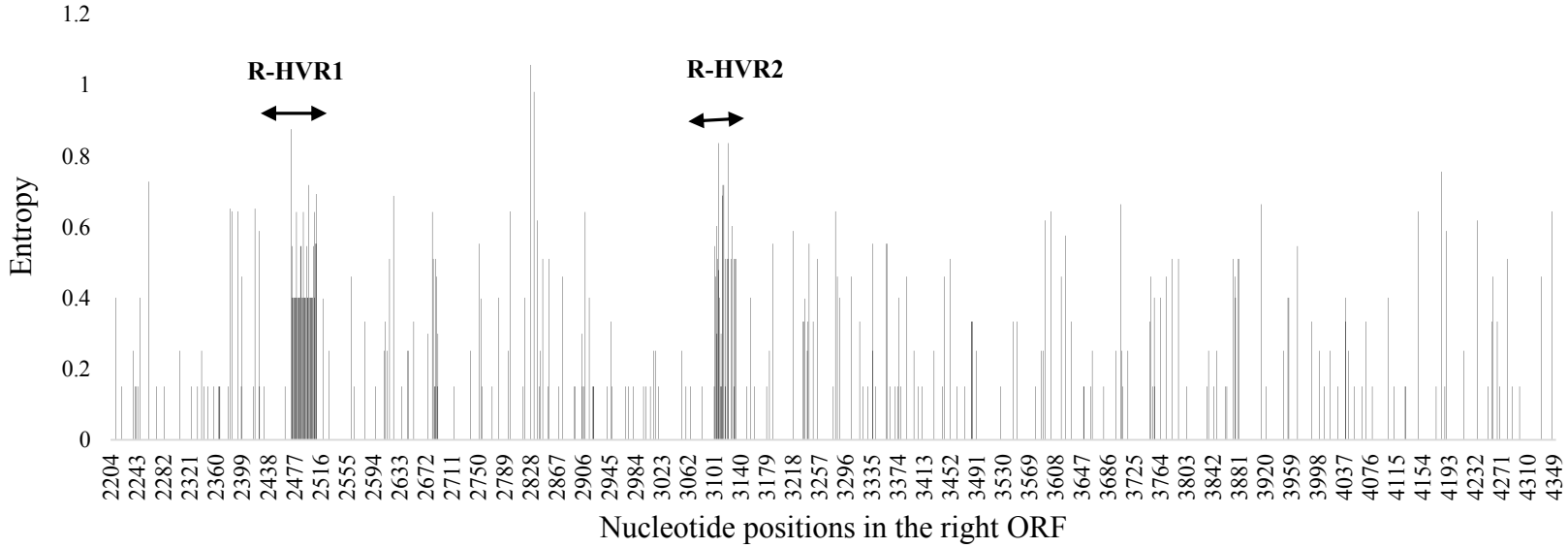

Figure 5.2. Variation of entropy values in the right ORF of the AMDV isolates using Entropy-ONE Web tool. Approximate positions of hypervariable positions are identified by arrows.

5.5.3. Entropy and hypervariable regions of the major protein NS1 and VP2

The entropy profile for the 641 aa of the NS1 protein of 34 AMDV isolates, including NS-CB, is shown in Figure 5.3. A total of 233 aa positions (36.6%) were non-zero. Entropy values varied between 0.133 (one variable aa) and 1.283 (6 different aa), and 58 of the 233 non-zero positions (24.8%) had the lowest entropy value of 0.133. Ten variable regions were identified according to the criteria suggested in section 4.5.5, i.e., having at least four consecutive variable aa positions, and fewer than four positions with zero entropy. The length of the variable regions varied between eight and 16 aa (Table 5.17). The number of consecutive variable aa for the variable regions varied between 30.7% (variable region 9) and 70.0% (variable region 7). The highest entropy value (1.283) was at variable region one (aa positions 10 to 25). The average entropy of the variable regions varied between 0.0247 and 0.470. Six HVRs (NS1-HVR1 to NS1-HVR6) were identified, ranging from 11 to 12 aa in length and had a minimum of one position with maximum entropy value greater than 0.8 (Table 5.17 and Figure 5.3). The highest variability was in 13 aa long NS1-HVR3 regions, which includes four aa positions with entropy values greater than 0.8 and the average entropy of 0.470. Amino acid positions with entropy values higher than 0.8 were towards the C-terminal region of the NS1 protein. These four positions with entropy values greater than 0.8 were not, however, within the identified HVRs (Figure 5.3).

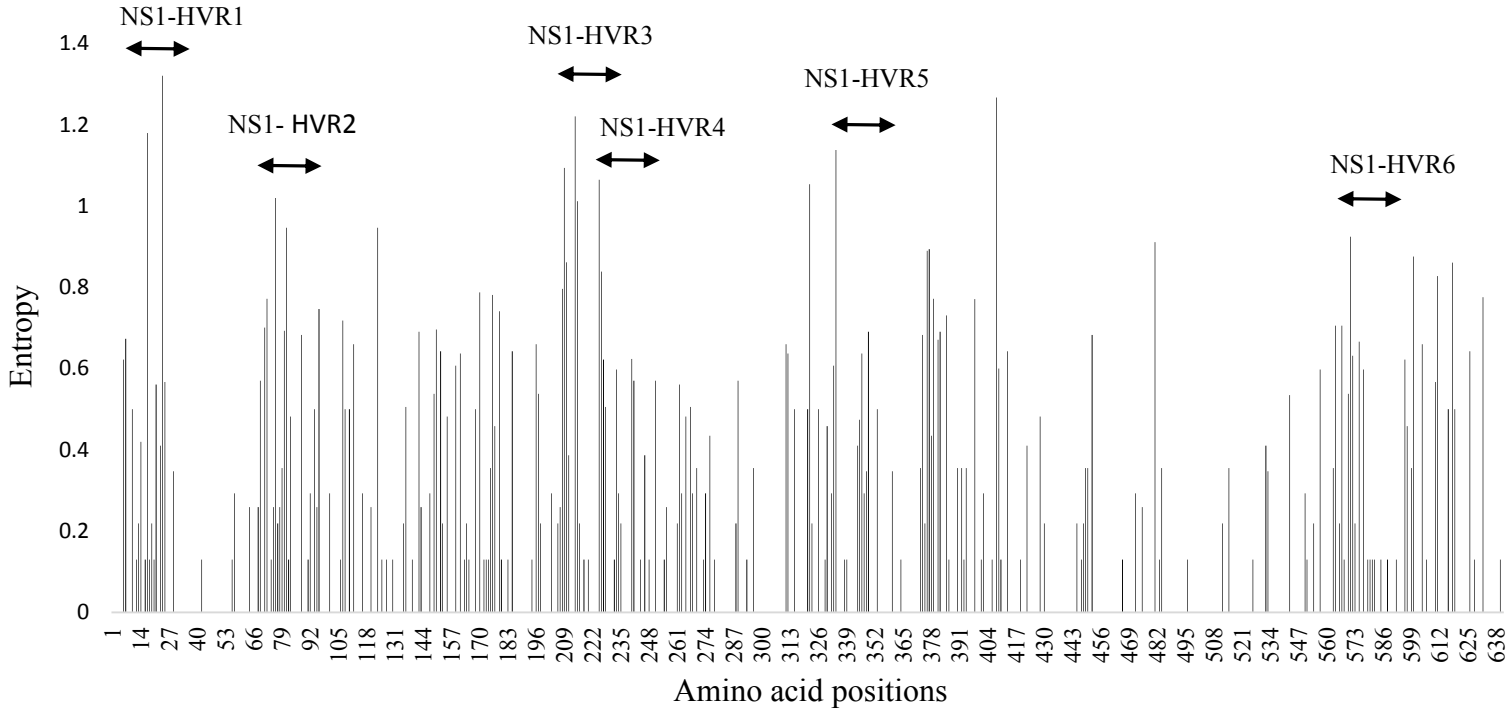

Figure 5.3. Variation of entropy values in NS1protein of the AMDV isolates using Entropy-ONE Web tool. Approximate positions of hypervariable positions are identified by arrows.

Table 5.17. Variable regions of the NS1 protein of 34 AMDV isolates based on the entropy of amino acids

Variable region	Amino acid position	Length	Number of consecutive variable positions	Number of zero entropy positions	Number of positions with selected entropy values		Minimum entropy	Maximum entropy	Average entropy
					>1	>0.8			
1	**10-25**[a]	16	6	3	2	2	0.133	1.283	0.366
2	**68-83**[a]	16	10	2	1	2	0.133	1.038	0.428
3	168-180	13	6	2	0	0	0.133	0.794	0.322
4	**206-218**[a]	13	6	3	3	4	0.219	1.208	0.470
5	**225-235**[a]	11	4	3	1	2	0.219	1.082	0.382
6	344-354	12	6	3	0	0	0.293	0.670	0.330
7	**373-386**[a]	11	7	3	0	2	0.133	0.913	0.452
8	445-452	8	4	2	0	0	0.133	0.677	0.247
9	**563-575**[a]	13	4	4	0	1	0.133	0.908	0.390
10	577-585	9	4	3	0	0	0.133	0.578	0.138

[a] Hypervariable region are bolded.

The entropy profile of the 647 aa of the VP2 protein of 42 AMDV isolates, including the NS-CB, is shown in Figure 5.4. The entropy values varied between 0.110 (one variable aa) and 1.056 (five different aa residues) for 124 positions of the 647 aas (19.1%) of the VP2. Five variable regions matched the criteria used in finding variable regions for the VP2 and their lengths varied between six and 13 aa. The variable region 1 (aa 22-34) contained no zero entropy positions. The variable region one, two and five were identified as HVRs (VP2- HVR1 to VP2-HVR3), and comprised of 13, 11 and six aa residues, respectively (Table 5.18 and Figure 5.4). Also, two aa positions with the entropy value greater than 0.8 were identified in the regions that did not meet the criteria for being variable (Figure 5.4).

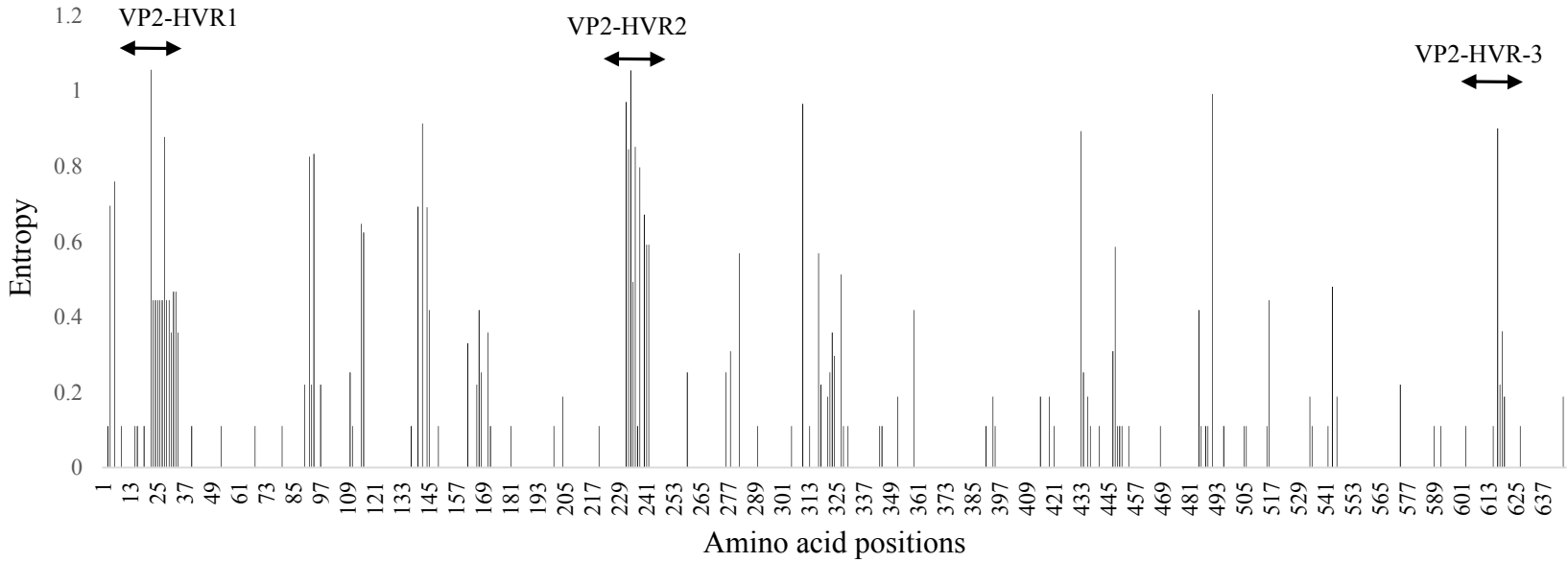

Figure 5.4. Variation of entropy values in VP2 of the AMDV isolates using Entropy-ONE Web tool. Approximate positions of hypervariable positions are identified by arrows.

Table 5.18. Variable regions of the VP2 protein of 42 AMDV isolates based on the entropy of amino acids

Variable region	Amino acid position	Length	Number of consecutive variable position	Number of zero entropy positions	Number of positions with selected entropy		Minimum entropy	Maximum entropy	Average entropy
					>1	>0.8			
1	**22-34**[a]	13	13	0	1	2	0.359	1.056	0.514
2	**232-242**[a]	11	7	1	1	4	0.110	1.055	0.634
3	321-330	10	4	3	0	0	0.110	0.513	0.183
4	447-454	8	5	2	0	0	0.110	0.586	0.11
5	**615-620**[a]	6	4	1	0	1	0.110	0.901	0.296

[a] Hypervariable region are bolded.

The entropy profile analysis of the left ORF also showed that the NS-CB genome had 12 unique nt positions outside of the identified eight HVR regions, which caused the lowest entropy value of 0.13 (Table 5.19). There were 14 positions within the left ORF HVRs, two of them were unique nts, and 12 nt existed in global isolates, and all lowered the entropy values between 0.011 and 0.118. The two unique nts caused the highest entropy differences (0.117 and 0.118) between the global AMDV isolates and NS-CB at positions 879 and 882 of the L-HVR5 (Table 5.20). There were ten unique nt positions in the NS-CB genome outside of the identified HVRs in the right ORF (Table 5.19), which resulted in the lowest entropy value (0.15) in the entropy profile (Table 5.19). There were two nt positions (2472 and 3105) in the R-HVR-1 and R-HVR-2 in the NS-CB genome which reduced the entropy values by 0.020 and 0.145, respectively, and the nt at position 3105 was a unique nt substitution (Table 5.19).

There were five unique aa positions in the NS1 protein and two in the VP2 protein of the NS- CB outside of the identified HVRs, which caused minimum entropy values of 0.133 and 0.110 in the NS1 and VP2, respectively (Table 5.20). Within the NS1-HVR2, NS1-HVR3 and NS1-HVR4, one, two, and one aa substitutions, respectively, were unique in the NS-CB sequence (Table 5.21). There was no unique aa substitution in the VP2-HVRs. The highest entropy difference of 0.075 was observed at position 232 of the VP2-HVR2, where the NS-CB shared the same aa with one of the global AMDV isolates.

Table 5.19. Position, entropy value, and the unique nucleotides in the left and right ORF of the NS-CB isolate

Positions	Entropy value	Nucleotide in left and right ORF	
		Global AMDV isolates	NS-CB
329	0.130[b]	G	A
426	0.130	A	C
706	0.130	C	T
864	0.130	A	G
943	0.130	C	G
1090	0.130	G	A
1247	0.130	G	C
1691	0.130	G	A
1781	0.130	G	A
1946	0.130	C	T
1951	0.130	A	C
2169	0.130	G	A
2272[a]	0.150[c]	G	A
2357	0.150	G	A
2425	0.150	C	T
2686	0.150	A	T
2905	0.150	C	T
2942	0.150	G	A
2949	0.150	G	A
2999	0.150	G	T
3329	0.150	G	A
3527	0.150	C	C

[a] The right ORF start at nt 2204
[b] The lowest entropy in the left ORF
[c] The lowest entropy in the right ORF

Table 5.20. Position, entropy value, and nucleotide composition of the NS-CB inside HVRs of the left and right ORFs

HVR region	Position	Entropy			Nucleotide variation and number	
		Global AMDV & NS-CB	Global AMDV	Entropy difference	Global AMDV	NS-CB
L-HVR1	276	0.890	0.875	0.015	T (22)[a], C (8), A (2), G (1)	C
L-HVR3	827	0.660	0.649	0.011	A (20), G (13)	G
,,	833	0.819	0.750	0.069	A (21), T (11), C (1)	C
,,	835	0.219	0.133	0.086	G (31), A (2)	A
L-HVR4	845	0.946	0.929	0.017	T (20), G (7), A (6)	G
,,	850	1.070	1.052	0.018	T (20), G (5), A (6), C (2)	A
L-HVR5	878	0.538	0.508	0.030	A (26), C (7)	C
,,	**879[b]**	0.546	0.428	**0.118**	A (29), C (3), T (1)	**G**
,,	**882[b]**	0.561	0.444	**0.117**	A (29), C (2), T (2)	**G**
,,	886	0.925	0.875	0.050	T (22), A (8), C (2), G (1)	C
,,	888	0.293	0.224	0.069	G (31), A (2)	A
L-HVR6	1172	0.219	0.133	0.086	T (32), A (1)	A
L-HVR7	1351	0.916	0.894	0.022	A (21), T (6), G (6)	T
,,	1354	0.598	0.578	0.020	C (23), T (10)	C
R-HVR1	2472	0.912	0.892	0.020	A (18), G (6), gap (4)	G
R-HVR2	**3105[b]**	0.299	0.154	**0.145**	27 (G), 1 (C)	**T**

[a] number of different nucleotides at each position are in parenthesis.

[b]The unique nucleotide substitution positions, the highest entropy differences, and the unique nucleotide substitutions are bolded.

Table 5.21. Position, entropy value of the unique amino acids of the NS-CB isolate outside of the hypervariable regions of the NS1 and VP2

Position[a]	Entropy	Amino acid[b]	
		Global isolates	NS-CB
NS1[c]-42	0.133	D	N
NS1-74	0.133	N	T
NS1-220	0.133	N	S
NS1-496	0.133	D	N
NS1-526	0.133	V	I
VP2[d]-181	0.110	V	I
VP2-592	0.110	I	L

[a] Positions started from one for each protein.
[b] D-Aspartate, I-Isoleucine, L- Leucine, N-Asparagine, S-Serine, T- Threonine, V-Valine.
[c] NS1-non-structural protein 1.
[d] Virion protein 2

Table 5.22. Positions, entropy values, and amino acid composition of the NS1 and VP2 HVRs of the NS-CB

HVR	Position	Entropy			Amino acid variation and number	
		Global AMDV & NS-CB	Global AMDV	Entropy difference	Global AMDV	NS-CB
NS1-HVR1	71	0.715	0.658	0.057	T (27)[a], K (3), R (2), I (1)	R
,,	81	0.942	0.902	0.040	Q (20), E (10), D (3)	D
NS1-HVR2	208	0.794	0.791	0.003	I (19), V (13), T (1)	V
,,	210	0.859	0.856	0.003	M (19), L (12), I (2)	L
,,	**214**[b]	1.208	1.150	0.058	S (22), T (5), V (5), I (1)	**A**
NS1-HVR3	**225**	1.082	0.978	0.104	N (21), H (7), T (4), I (1)	**R**
,,	**226**	0.787	0.674	0.113	K (28), Q (2), V (2), T (1)	**R**
,,	228	0.444	0.363	0.081	G (30), S (2), D (1)	D
,,	234	0.298	0.229	0.069	A (31), T (2)	T
NS1-HVR4	**377**	0.913	0.803	0.110	M (28), I (2), L (2), Q (1)	**T**
,,	379	0.753	0.730	0.023	N (24), T (7), S (2)	T
NS1-HVR5	571	0.908	0.886	0.022	K (23), R (6), G (3), A (1)	R
VP2-HVR1	28	0.878	0.751	0.127	G (31), S (2), W (1), gap (7)	G
VP2-HVR2	232	0.971	0.942	0.029	V (29), T (4), M (3), S (3), A (2)	V
,,	233	0.845	0.770	0.075	A (28), G (11), P (1), S (1)	S
,,	236	0.852	0.846	0.006	T (29), P (1), Q (8), gap (3)	T
,,	238	0.797	0.795	0.002	T (27), E (12), N (1), I (1)	T
,,	240	0.672	0.663	0.009	D (29), T (11), S (1)	D
VP2-HVR2	241	0.592	0.582	0.010	A (29), G (12)	A
VP2-HVR2	242	0.592	0.582	0.010	V (29), T (12)	V

[a] Number of different amino acid in each position are in parenthesis.
[b] The unique aa substitutions and their positions in HVRs are bolded.

5.6. Recombination analysis

5.6.1. Recombination breakpoints and recombinant sequences

Eleven recombination events were detected in the 32 ECR of AMDV isolates (Table 5.23). Five of the recombination events were supported by all seven detection methods, and six events were supported by at least four of the seven detection methods. The breakpoints of the seven of the 11 recombination events were located in L-ORF, two were in the R-ORF, and two were on both ORFs (Table 5.22. and Figure 5.5). The identification of the breakpoint end for event number 10 by the software is opposite to others and cannot be explained. The eleven recombination events involved eight isolates (Utah1, M195, M173, M228, LN3, LN2, LN1, and WM25) which carried segments of sequences from major and minor parents. The Utah1, M195, LN3, and WM25 showed two recombination events each, and the other 24 isolates were not recombinant, which included NS-CB, 19 Danish strains, G, SL3 and two of the Chinese strains (LM and Beijing). The NS-CB was the major parent for three recombination events (3, 7, and 11) and minor parents for another 3 (1, 4, 9). The 19 Danish strains were not equally contributed to the recombination. The LN3 isolate associated with the event three and M195 isolate associated with event 8 showed trace evidence for the particular event, i.e. lower p-values compared with others.

Table 5.23. The summary of the significant recombination events in the entire coding regions of 32 AMDV sequences

Event No	Breakpoint[a] Begin	Breakpoint[a] End	Location of ORF	Recombinant isolate	Parent Major	Parent Minor	Recombination detection method[c]	P-value
1	18	2124	L ORF	Utah1	Unknown	G, SL3, NS-CB, Danish[d]	R, G, B, M, C, S, 3	2.09E-03 to 2.53E-35
2	37	2100	L ORF	M195	WM25	Unknown	R, G, B, M, C, S, 3	4.50E-02 to 4.41E-24
3	14	2125	L ORF	M173/M228	G, SL3, NS-CB, Danish[e]	WM25	R, B, M, C, 3	8.76E-03 to 5.16E-17
4	920	1840	L ORF	LN3	LN1	G, SL3, NS-CB, Danish[f]	R, G, B, M, C, S, 3	1.53E-05 to 2.63E-14
5	70	919	L ORF	LN2	SL3, G, Danish[g]	LN3, LN1	B, M, S, 3	2.64E-04 to 5.98E-45
6	231	1458	L ORF	LN1/[LN3][b]	LM	WM25	R, G, B, M, C, S, 3	1.97E-02 to 6.28E-13
7	1750	3394	L/R ORF	LN2	G, SL3, NS-CB, Danish[g]	LM	R, M, C, 3	1.44E-03 to 1.98E-07
8	2894	4099	R ORF	WM25/[M195][b]	LN3, LN1	Unknown	R, B, M, C, S, 3	1.51E-04 to 2.13E-09
9	2976	3460	R ORF	Utah1	WM25 M195	G, SL3, NS-CB, Danish	R, G, B, M, C,S, 3	3.60E-03 to 2.44E-06
10	1808	47	L ORF	M195	G, SL3, Danish	M173	G, M, C, 3	1.62E-02 to 1.09E-09
11	1660	2919	L/R ORF	WM25	G, NS-CB	Beijing	M, C, S, 3	2.30E-02 to 3.14E-03

[a] The first and the last nt of the recombination fragment. [r] Recombinants with trace evidence for recombination are in square brackets. [c] R=RDP, G= GENECONV, B= Bootscan, M=Maxchi, C=Chimaera, S=SiSscan, 3= 3Seq , [d] 18 Danish sequences, [e] 19 Danish sequences, [f] 1 Danish sequence, [g] 4 Danish sequences,

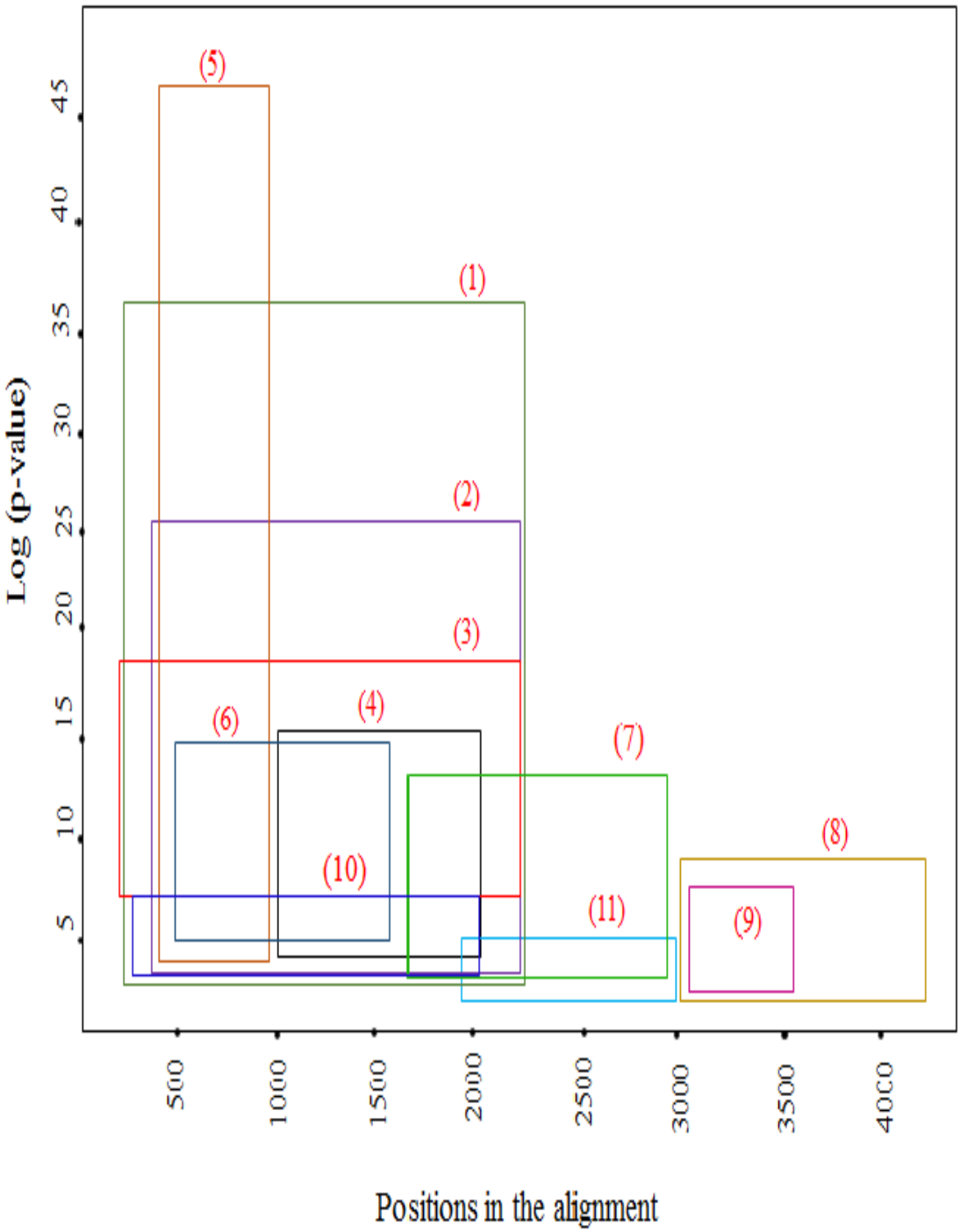

Figure 5.5. Approximate positions of 11 recombination events of the entire coding region of 32 AMDV sequences. Numbers in parenthesis above each rectangle are the order of the recombination event depicted in Table 5.23.

5.6.2. Impact of recombination on the phylogenetic incongruity

The phylogenetic trees constructed using the ECR of the NS-CB along with that of the 23 non-recombinant sequences and the 31 global AMDV sequences are shown in Figure 5.6 and Figure 5.7, respectively. These trees were constructed using the best substitution models with the lowest Bayesian Information Criterion (BIC). The model that gave the lowest BIC value for each set of sequences was considered as the true model. The best substitutions for the two phylogenies are shown in Table 5.24.

Table 5.24. The best substitution model for the two phylogenies

Phylogeny	Best substitution Model
Recombinant and non- recombinant sequences	GTR+G+I[a]
Non recombinant sequences	HKY+G[b]

[a]General Time Reversible + Gamma+ Invariant Sites
[b]Hasegawa-kishino-Yano+G

5.7. Phylogenetic analysis

The phylogenetic trees constructed using ECR of NS-CB and the 23 non-recombinant sequences and the 31 global AMDV sequences are shown in Figures 5.6 and 5.7, respectively. In both phylogenies, the sequences clustered into two clades. The clade (i) contained most isolates in both phylogenies, and clade (ii) contained two non-recombinant sequences from China (LM and Beijing). The true evolutionary relationships of the sequences are represented by the phylogeny constructed using the non-recombinant sequences (Figure 5.6). The relationship of the NS-CB to the 19 Danish strains was supported by 74% bootstrap value. The NS-CB isolate showed a closer relationship to Danish isolates compared to the LM and Beijing strains from China (Figure 5.6). The relationships of SL3 to AMDV-G and the LM to Beijing strains were supported by 100%

bootstrap values (Figures 5.6 and 5.7). The seven recombinant strains clustered in the clade (i) in between the non-recombinant strains (Figure 5.7). The three recombinant strains from China (LN1, LN2 and LN3) showed a close relationship to the Newfoundland strains M228, M173, and WM25 (Figure 5.7). The phylogenetic tree constructed with the addition of recombined sequences did not change the overall topology of the tree constructed using non-recombined sequences. However, the changes in bootstrap values were observed using recombinant and non-recombinant sequences. For example, the NS-CB had 100% bootstrap support for the Danish strains on recombinant sequences, but the NS-CB had lower bootstrap support (74%) on the non-recombinant tree. Several unresolved nodes (node with more than two descendants) were associated with the Danish strains in both phylogenies.

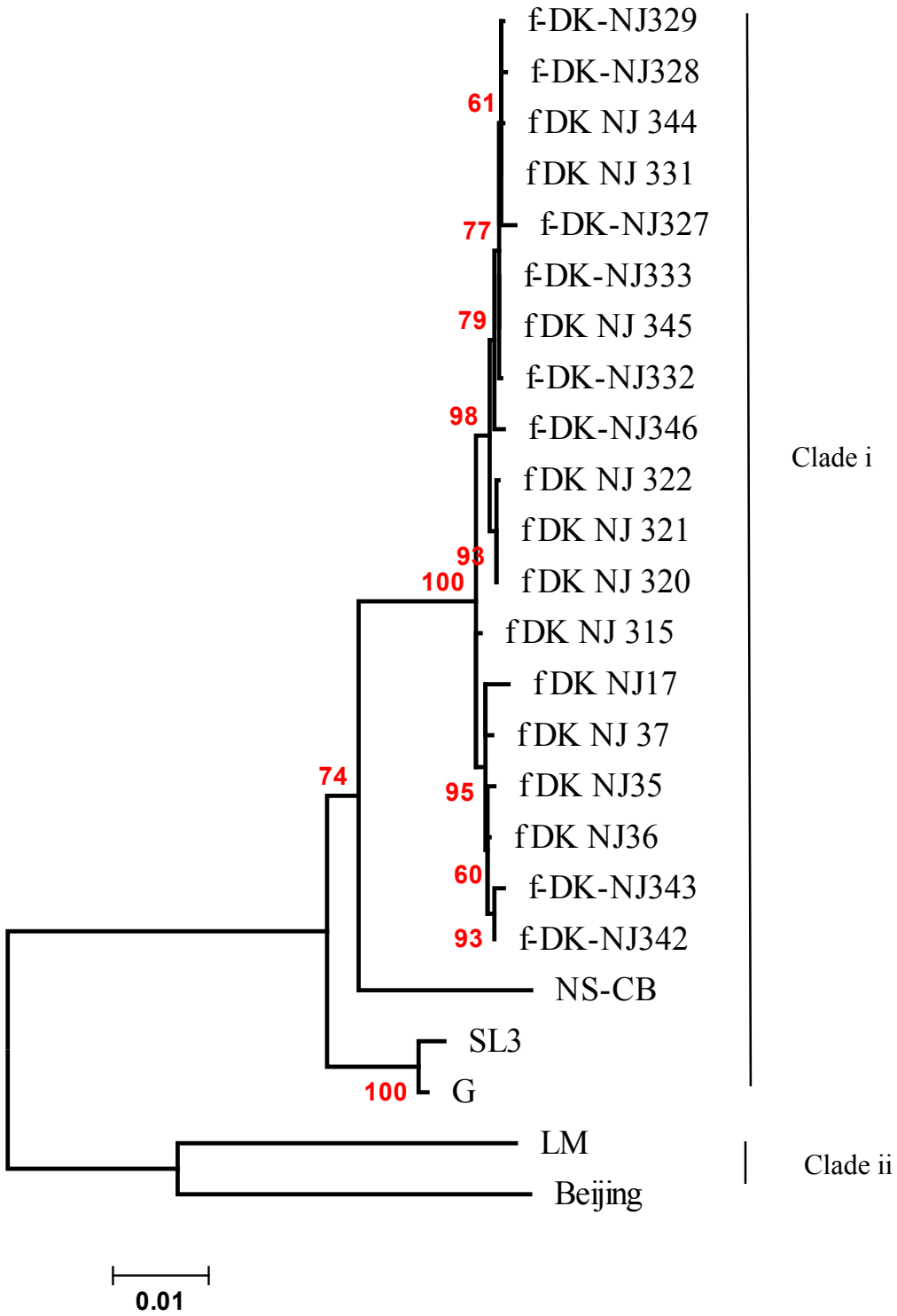

Figure 5.6. ML mid-point phylogenetic tree constructed from the entire coding region of 23 non- recombinant global AMDV sequences and the NS-CB. The analysis was performed using 1,000 bootstrap replicates. The number of nt substitutions per site is indicated in the scale bar.

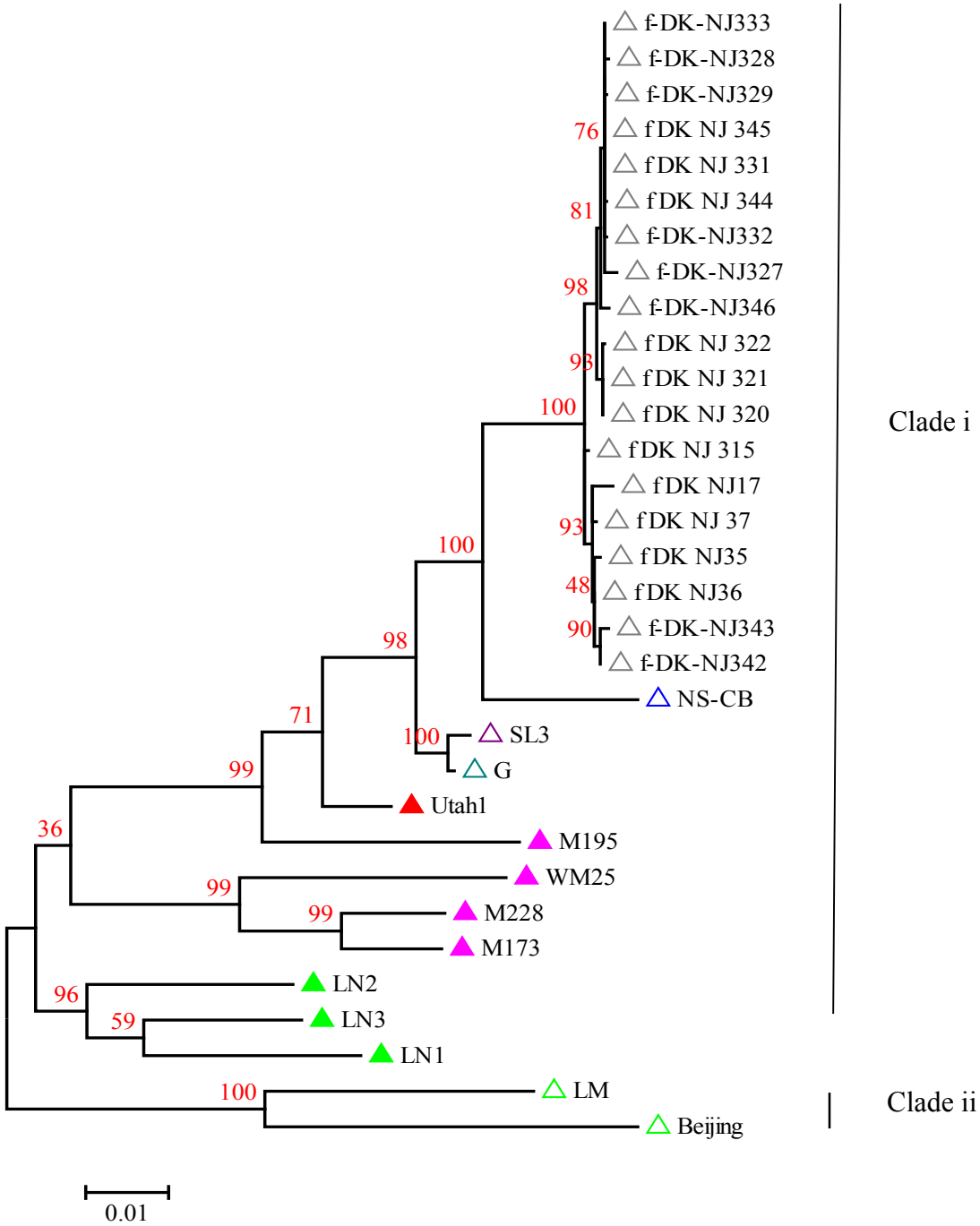

Figure 5.7. ML mid-point rooted phylogenetic tree represents the entire cording region of 31 global AMDV sequences and the NS-CB. The analysis was performed using 1,000 bootstrap replicates. Recombinant and non- recombinant sequences are represented in solid and open triangles, respectively. The number of nt substitutions per site is indicated in the scale bar.

5.8. Sequences of the NS-CB descendants over time

Positions and lengths of the sequences of the 20 NS-CB descendants on 16 and 176 wpi are shown in Table 5.25. Nine NS-CB descendants had nearly complete genome sequences as one fragment, including complete ECR and the partial 3' terminal sequences. The descendant # 318 was sequenced as one fragment, but the sequence of the 3' end of ECR was not obtained. Seven of the nine nearly complete genome sequences of the NS-CB descendants were the samples collected on 16 wpi. Ten of the NS-CB descendants had genome sequences as two fragments, because of the failure of the sequencing of the GT repeat region and the 3'end of the ECR. Sequences of ECR in these samples ranged between 71.0% (descendant #422) and 96.1% (descendant #11) (Table 5.25), with an average of 83.48%.

5.8.1. Changes in the sequence of the NS-CB over time and mutation rate

The sequence of the GT repeat region (nts 2496 to 2516) was available for 10 NS-CB descendants and sequencing of this region in the remaining samples were not successful. Multiple sequence alignment generated at 16 and 176 wpi showed that NS-CB descendant number 315 had 27 nt deletions between nts 2,490 and 2,516 in the GT repeat region, in addition to four other nt substitutions (Table 5.26). In addition, only two descendants (137 and 496) had the same sequence as the NS-CB in this region, six descendant had an ambiguous code (G/C) at position 2495, and one variant (313) had a T instead of a G at this position

Table 5.25. Positions and lengths of the genome sequences of the NS-CB descendants

Descendant ID	Weeks pi and year	Fragment 1	Fragment 2	Total sequenced length (nt)	Percentage of ECR[b] sequenced length
		Positions[a]	Positions[a]		
42	16 weeks (2011)	206-4600	_	4395	100%
44		206-4561	_	4356	100%
313		206-4604	_	4399	100%
315[c]		206-4604	_	4372	100%
437		206-4554	_	4349	100%
495		206-4602	_	4397	100%
496		206-4602	_	4397	100%
318		206-3333	_	3127	75.47%
348		206-2523	2684-4197	3832	92.49%
494		206-2136	2784-4198	3346	80.76%
137	176 weeks (2014)	206-4604	_	4399	100%
656		206-4602	_	4397	100%
11		206-2366	2783-4604	3983	96.13%
139		206-2138	2783-4560	3711	89.57%
162		206-2146	2783-4560	3719	89.76%
320		206-2143	2783-4349	3504	84.57%
378		206-1743	2773-4546	3107	74.99%
422		206-1735	2785-4196	2942	71.01%
424		206-2136	2784-4200	3348	80.81%
445		206-2135	2850-4350	3431	82.81%

[a] Positions are based on the AMDV-G.
[b] ECR spans from nt 206-4349.
[c] Descendant with 27 nt deletions in the sequence.

. There were 21 and 69 nt substitution positions at 16 and 176 wpi, respectively (Tables 5.27 and 5.28). The number of substitutions varied between one and seven in different descendants after 16 wpi, for a total of 36, which increased to three to 21 substitutions in 176 wpi, for a total of 113. The number of ambiguous positions ranged between 0 (descendant #315) and 4 (descendant #437) after 16 weeks, and between 0 (descendant 139, 152, 378) and 15 (descendant #11) after 176 weeks, for a total of 18 and 37, respectively. A rare ambiguous code D (A/G/T) was observed in the sequence of the descendant #44, which was counted as two substitutions.

The total number of substitution positions, nt substitutions, ambiguous positions are shown in Table 5.29. These values were expressed as percentages of the total number of nts sequence at each of 16 (40,970 bp) and 176 (36,541 bp) wpi, and as percentages of the NS-CB sequence (4397 bp). Sequences at 176 wpi had greater percentages of substitution positions, total nt substitutions and ambiguous positions compared to that at16 wpi based on the total number of nts sequenced and based on the length of the NS-CB. The number of transitional substitutions at 16 wpi were much smaller than those at 176 wpi (13 vs59), and the former constituted a smaller percentage of all substitutions (36,11%) compared with the latter (52.21%). The number of transversional substitutions was smaller at 16 wpi (23) compared with that at 176 wpi (54), but the former constituted a higher percentage of all substitutions (63.88%) compared with the latter (47.78%).

The mutation rate (substitution rate) was computed based on the total number of sequences of the 10 the NS-CB descendants at each sampling occasion and based on the length of the NS-CB which was sequenced and expressed per year. The estimates based on all nucleotides were 2.85×10^{-3} and 9.13×10^{-4} substitutions/site/year for 16 and 176 wpi, respectively and the t-test showed that these two estimates were significantly different at ($p< .001$). The estimates based on the length of NS-CB were 1.5×10^{-2} and 4.6×10^{-3} substitutions/site/year for 16 and 176 wpi, respectively.

Table 5. 26. Nucleotide deletions in the TG repeat region of 10 NS-CB descendants

Descendant ID	Nucleotide positions[a]															
	2469	2472	2475	2478	2481	2484	2487	2490	2493	2496	2499	2502	2505	2508	2511	2514
NS-CB	GGT	GGT	GGG	GGG	GGT	GGG	GGT	GGT	GGG	GGT	GGT	GGT	GGT	GGT	GGT	GGG
Desce[d].	. . .[b]	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . K	. . .	. . .	. . .	. . .	. . .	. . .	. . .
137, 496	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
313	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . T	. . .	. . .	. . .	. . .	. . .	. . .	. . .
315	. . .	. . .	. . T	. . T	. . .	. . T	. . G	-[c]	-	-	-	-	-	-	-	-

[a] Positions are based on the AMDV-G sequence.
[b] A dot represents same nucleotide to the NS-CB sequence.
[c] A dash indicates the deletion of three nucleotides.
[d] Descendant: 42, 44, 318, 437, 495 and 656.

Table 5.27. Nucleotide substitution of the NS-CB descendant after 16 wpi

Substitute positions[a]	NS-CB	Descendant ID										No of ambiguous substitutions	No of substitutions[d]
		42	44	313	315	437	495	496	318	348	496		
224	G	**R**	.	.	.	.	.	.	**A**[b]	.	.[c]	1	2
255	C	.	.	.	**T**	.	.	.	.	.	.	0	1
374	T	.	.	.	.	.	.	**W**	.	.	.	1	1
674	C	.	.	.	**T**	.	.	.	.	.	.	0	1
746	C	.	.	**Y**	.	.	.	.	.	.	.	1	1
845	G	**R**[e]	.	.	.	.	.	.	.	.	.	1	1
848	A	**W**	.	.	**C**	.	.	.	.	.	.	1	2
864	G	.	.	.	.	.	.	**R**	.	.	.	1	1
1323	A	.	.	**R**	.	.	.	.	.	.	.	1	1
1347	T	.	.	**A**	.	.	**W**	.	.	**A**	.	1	3
1429	A	.	.	.	**C**	.	.	.	.	.	.	0	1
1553	G	**C**	.	**C**	.	.	.	.	**Y**	.	**S**	2	5
1951	C	**G**	.	.	**G**	.	.	.	.	.	.	0	2
2495	G	.	.	.	.	**K**	-[f]	**K**	-	**K**	.	3	3
2751	T	**A**	**D**	.	.	**K**	-	.	-	.	.	2	4
2835	C	**G**	.	.	.	.	-	.	-	**G**	.	0	2
2902	T	.	.	.	.	.	.	.	.	**C**	.	0	1
3110	A	.	.	.	.	.	.	**M**	.	.	.	1	1
3123	G	.	.	**R**	.	.	.	.	.	.	.	1	1
3333	G	.	.	.	.	.	.	.	.	**A**	.	0	1
4561	A	.	**G**	.	.	-	.	.	.	.	.	0	1

[a] Nucleotide substitution positions are based on AMDV-G.

[b] Nucleotide substitutions are bolded and are identified compared to the NS-CB.

[c] A dot represents the same nucleotide to the NS-CB isolate.

[d] Number of substitution positions is the sum of transitional and transversional substitutions compared to the NS-CB at each substitution position including the ambiguous positions.

[e]Ambiguous codes D- A/G/T, K= G/T, M = A/C, R = A/G, S = G/C, W = A/T and Y =C/T

[f] Positions with no available sequences.

Table 5.28. Nucleotide sequence variation of the NS-CB descendants after 176 wpi

Substitute positions[a]	NS-CB	Descendant ID										No of ambiguous positions	No of substitutions[b]
		11	137	139	162	320	378	422	424	445	656		
221	A	.	.	.	.	.	.	**G**[c]	.	.	.	0	1
224	G	.	.	**A**	.	.	.	**A**	.	.	.	0	2
239	C	.	.	.	.	.	.	.	.	.	**S**	1	1
244	G	.[d]	.	.	.	.	.	.	.		**R**	1	1
256	G	.	.	.	.	.	.	.	.	.	**R**	1	1
329	A	.	.	.	**G**	.	.	.	**R**	.	**G**	1	3
419	A	.	.	.	**G**	.	.	.	.	.	.	0	1
469	C	.	.	.	.	**Y**	.	.	.	.	.	1	1
479	A	**R**[e]	.	.	.	.	.	.	.	.	.	1	1
485	C	.	.	**A**	**G**	**G**	.	**A**	**M**	.	**M**	2	6
633	A	.	.	.	.	**T**	.	.	.	.	.	0	1
637	A	.	.	.	.	**T**	.	.	.	.	.	0	1
659	A	.	.	.	.	**G**	.	.	.	.	.	0	1
666	A	.	.	.	.	**T**	.	.	.	.	.	0	1
814	C	.	**Y**	.	.	.	.	.	.	.	.	1	1
845	G	.	**A**	**A**	.	.	.	**W**	.	.	.	1	4
864	G	.	.	.	.	.	.	.	**A**	.	.	0	1
848	A	.	**G**	.	.	.	.	.	.	.	.	0	1
879	G	.	**A**	.	.	.	.	.	.	.	.	0	1
988	T	.	.	.	.	**Y**	.	.	.	.	.	1	1
1076	A	.	**T**	.	.	.	.	.	.	.	.	0	1
1251	G	.	.	.	**A**	**A**	.	.	.	.	.	0	2
1282	A	.	**G**	.	.	.	**G**	.	**G**	.	**G**	0	4
1323	A	**G**	.	.	.	.	.	.	.	.	.	0	1
1347	T	.	.	.	.	.	**A**	**A**	.	.	.	0	2
1360	A	**G**	.	.	.	.	.	.	.	.	.	0	1
1446	G	**T**	.	.	.	.	.	.	.	.	.	0	1
1553	G	**C**	.	**C**	**C**	.	**C**	**C**	**S**	**C**	**C**	1	8
1744	A	.	.	.	.	.	**G**	-	.	.	.	0	1
1748	A	.	.	.	.	.	**G**	-	.	.	.	0	1
1800	A	.	**G**	.	.	.	-	-	.	.	.	0	1
1905	G	.	.	**A**	.	**A**	-[f]	-	.	.	.	0	2
1916	A	.	.	.	.	.	-	-	**G**	.	.	0	1
1951	C	**S**	**M**	**G**	**A**	.	-	-	.	.	.	2	4
2061	G	.	.	.	.	**R**	-	-	.	.	.	1	1
2076	C	.	**Y**	.	.	.	-	-	.	.	.	1	1
2077	A	.	**R**	.	.	.	-	-	**C**	**C**	**C**	1	4
2081	A	.	**M**	.	.	.	-	-	.	.	.	1	1

Table 5. 28. Continued

Substitute positions[a]	NS-CB	Descendant ID										No of ambiguous positions	No of substitutions[b]
		11	137	139	162	320	378	422	424	445	656		
2113	A	**R**	.	.	.	.	-	-	.	.	.	1	1
2115	G	**R**	.	.	.	.	-	-	.	.	.	1	1
2122	G	**R**	.	.	.	.	-	-	.	.	.	1	1
2146	G	.	.	-	**C**	-	-	-	-	-	.	0	1
2181	A	**R**	.	-	-	-	-	-	-	-	.	1	1
2187	G	**R**	.	-	-	-	-	-	-	-	.	1	1
2294	G	**S**	.	-	-	-	-	-	-	-	.	1	1
2441	G	-	**A**	-	-	-	-	-	-	-	.	0	1
2495	G	-	.	.	.	.	.	.	.	.	**K**	1	1
2901	K	**T**	**T**	**G**	**T**	**T**	**T**	**T**	**T**	**M**	**T**	1	2
2969	G	.	.	**A**	.	.	.	.	.	.	.	0	1
3102	T	.	.	.	.	.	.	.	**S**	.	**S**	2	4
3105	A	**W**	.	.	.	.	.	.	**W**	.	**T**	2	3
3118	C	**M**	.	.	.	.	.	.	.	.	.	1	1
3170	C	**Y**	.	.	.	.	.	.	.	.	.	1	1
3251	C	**M**	.	.	.	.	.	.	**A**	.	**A**	1	3
3302	G	.	.	.	**A**	.	.	.	.	.	.	0	1
3333	G	**R**	.	.	.	.	.	.	.	.	.	1	1
3334	C	.	**T**	.	.	.	.	.	.	.	.	0	1
3337	C	**Y**	.	.	.	.	.	.	.	.	.	1	1
3384	G	**T**	.	.	.	.	.	.	**S**	.	**C**	1	3
3688	G	.	.	.	**A**	.	.	.	.	.	.	0	1
3706	A	.	.	**G**	.	.	.	.	.	.	.	0	1
3710	G	.	.	.	**A**	.	.	.	.	.	.	0	1
3747	A	.	.	.	.	.	.	.	**T**	.	**T**	0	2
3858	A	.	.	.	.	.	.	.	**G**	.	**G**	0	2
3870	G	.	.	.	**A**	.	.	.	.	.	.	0	1
3876	G	**R**	.	.	.	.	.	.	.	.	.	1	1
3901	C	.	**G**	.	.	.	.	.	.	.	.	0	1
4359	C	.	.	**T**	.	-	**T**	-	-	-	-	0	2
4560	C	.	.	**A**	**A**	-	-	-	-	-	-	0	2

[a] Nucleotide substitution positions are based on AMDV-G.
[b] Number of substitution positions are the sum of transitional and transversional substitutions compared to the NS-CB at each substitution position including the ambiguous positions. [c] Nucleotide substitutions are bolded and were identified compared to the NS-CB. [d] Ambiguous codes D=A/G/T, K= G/T, M = A/C, R = A/G, S = G/C, W = A/T and Y = C/T. [e] A dot represents same nucleotide to the NS-CB isolate.
[f] Positions with no available sequences.

Table 5.29. Summary of sequence variation of the NS-CB descendants

Variations	Based on the total number of nucleotides sequenced		Based on the length of the NS-CB sequenced	
	16 weeks	176 weeks	16 weeks	176 weeks
Number of nucleotides sequenced[a]	40970	36541	4397	4397
Number and % of substitution positions[d]	21[b](0.05%)	69[c](0.18%)	21(0.02%)	69(1.56%)
Number and % of nt substitutions[d]	36[b](0.08%)	113[c](0.30%)	—	—
Number and % of ambiguous substitutions[d]	18[b](0.04%)	37[c](0.10%)	18(0.40%)	37(0.84%)
Number and % of transitional substitutions[e]	13(36.11%)	59(52.21%)	13(36.11%)	59(52.20%)
Number and % of transversional substitution[e]	23(63.88%)	54(47.78%)	23(63.88%)	54(47.78%)
Mutation rate (number of substitutions/site/year)	2.85×10^{-3}[f]	9.13×10^{-4}[f]	---	---
Substitution rate for NS-CB (number of substitution positions /site/year)	—	—	1.5×10^{-2}[g]	4.6×10^{-3}[g]

[a] See Table 5.24, [b] see Table 5.2.6, [c] see Table 5.27 for details.

[d] Percentages were calculated based on the number of nucleotides sequenced of the 10 NS-CB descendants after 16 and 176 wpi. and the number of nucleotides sequenced of the NS-CB at the time of inoculation.

[e] Percentages were calculated based on the number of substitutions that occurred over 16 and 176 wpi.

[f] Mutation rate was estimated as the number of substitutions/site/years by dividing the total number of nt substitutions by the total number of nt sequenced and multiplying by the time from inoculation.

[g] Substitution rate for the NS-CB was estimated as the number of substitution positions /site/year by dividing the total number of nt substitutions positions by the total number of nt sequenced of NS-CB at the time of inoculation and multiplying by the time from inoculation.

CHAPTER 6. DISCUSSION

6.1. Genome organization, gene and protein characteristics of the NS-CB

The first objective of this study was to sequence the strain of AMDV (NS-CB) using the Sanger Sequencing method. In this study, 91.6% of the NS-CB genome was sequenced covering the ECR and segment of the 3'-UTR. Sequencing of 5' and 3' end was not attempted in this study because of the difficulties in amplifying palindromic structures at the 3' end and 25 bp A+T rich repeat region at the 5' end (Bloom *et al*., 1994). The occurrence of multiple ambiguous positions is reported in AMDV sequences previously (Canuti *et al*., 2016 ; Jakubczak *et al*., 2017). Thus, the observation of an ambiguous position at position 2901 in NS-CB is not unusual. Sequence analysis of NS-CB showed that the existence of previously reported (Haung *et al*. 2014), 67 nt long untranslated regions in right ORF, which provide evidence of conservation of the number of aa in VP1 sequence. This study reports the existence of 641, 114 and 87 aa NS1, NS2 and NS3 proteins respectively and VP1 and VP2 of the NS-CB, as a result of the conservation of splicing sites identified by Haung *et al*., (2014). In agreement with Bloom *et al*., (1994) and Haung *et al*., (2014), there was 43 unique aa at the N-terminal region of the VP1 of NS-CB. The number of aa in VP2 varies with strains of the AMDV (Alexanderson *et al*., 1986), and 647 aa in the VP2 of the NS-CB was similar to the AMDV-G sequence with 16 glycine residues (Haung *et al*., 2014) and Utha1 (Bloom *et al*., 1994).

6.2. Nucleotide and amino acid variations of the NS-CB

The findings of this study show that nucleotide variations in the three non-structural CDSs and genes of 26 global AMDV isolates and the NS-CB were comparable (17.4% to 18.9%) and were higher than those in the VP2 gene (10.6%), VP1u gene (8.5%) and the 3'-UTR segment (10.2%) (Table 5.2). Amino acid variabilities in the same regions were more pronounced, ranging between 24.6% and 28.7% in the three non-structural proteins, 11.1% in the VP2 protein and 4.7% in VP1u protein (Table 5.4). The aa variability of the NS1 protein was 9.1% greater than the nt variability of the NS1 gene (28.7% vs 19.6%) in the present study, which is close to 10% difference (26% and 16%) in the corresponding region in a previous study (Olofsson *et al*., 1999). Gottschalck *et al*. (1994) reported a high degree of nt variability in the non-structural genes of AMDV-G, Utah1, K and United strains when compared with each other. Canuti *et al*. (2016) evaluated aa compositions of the NS1 and VP2 in isolate site by site selection pressure of NS1 and VP2. They concluded that both NS1 and VP2 aa residues were under positive selection, and the selection pressure on NS1 was higher than that on the VP2 (1.4% vs 0.6% of aas). Canuti *et al*. (2016) concluded that the degree of variability in NS1 was higher than VP2, which is similar with findings in the current study.

It was stated that the reason for the high degree of variability of the non-structural genes is that limited coding sequences are needed to be conserved to maintain the functional properties of the NS genes (Olofsson *et al*., 1999). It should be noted that although sequences of a large number of AMDV isolates are available on GenBank, only 26 global isolates with the same lengths as the NS-CB were compared in the current study for the differences in nt and aa variability among different regions to be meaningful. Five

AMDV strains with nearly complete genomes (SL3, LM, Beijing, LN2, and LN3) available on GenBank were not used in this study because there were no annotation.

6.3. Unique nucleotide and amino acid variations and sequence motifs of the NS-CB

The observation that there are 34 nts and 18 aa residues in the NS-CB which are not present in any of the other 26 to 42 isolates to which it was compared, i.e. unique variants (Tables 5.5 to 5.11), suggests that NS-CB was different from those isolates. The presence of these numbers of unique variants does not necessarily mean higher genetic variability of the NS-CB than the other isolates, because the number of such variants depends on the number of isolates involved in the comparisons. The presence of aa variations between the NS-CB and global isolates at the functional motifs of the NS1 and VP2 proteins could have biological significance.

Amino acid composition of the first caspase recognition site in the NS1 (224INTDS228) showed marked differences among the AMDV isolates which were compared (Table 5.12). The aa I224 is conserved among the 25 global isolates as well as the NS-CB, which may suggest that this aa has an important function. The other four aas at this site were variable among global isolates, and interestingly NS-CB had unique aas at positions 225 (R vs N, H, T, I), 226 (R vs K, T, V, Q) and 228 (D vs G, S), whose effects on caspase activity is not clear. The aa at the caspase cleavage site (D227) in the current study was the same in AMDV-G, SL3, Utah1, and the 19 Danish strains. This position is occupied by glutamic acid (E) in several isolates, which is shown to block the cleavage of the NS1 at this position (Best *et al*., 2003). It may be concluded that (D227) caspase cleavage site is not essential for AMDV replication. The N terminal portion of the NS-CB contained four unique aas at positions 42, 74, 76 and 220 (Table 5.6), making the

N-terminal fragment of NS1 (aas 1 to 226) unique, and could affect activity of the cleavage site of each AMDV isolate, including NS-CB (Wang *et al*., 2008). The five aas at the right caspase recognition site (282DQTDS286) were conserved in all 34 isolates compared in this study (Table 5.12), which is in agreement with previous reports (Best *et al*., 2003; Leng *et al*., 2018), and implies that caspase activity and NS1 cleavage are essential in the replication of all AMDV strains (Best *et al*., 2003). When D227 and D285 were replaced by aa E at both positions, viral replication was reduced by approximately 1,000- to 10,000-fold compared to that of the wild-type virus (Best *et al*., 2003). The caspase cleavage site in the VP2 protein (417DLLD↓D) was conserved in all isolates compared, suggesting the importance of this motif for virial production which is controlled by caspase activation (Cheng *et al*., 2010). The six aa residues (G, T, G, K, T, L) from aa 435 to 440 which are associated with the ATP binding pocket of the NS1 protein were completely conserved in NS-CB and all 42 AMDV isolates compared, which has been previously reported (Gottschalck *et al*., 1994), and suggests the importance of this region in viral replication (Canuti *et al*., 2015).

Perhaps the most practical conclusion that can be deduced from comparing different viral sequences is estimating the pathogenicity of the NS-CB isolate. Five aas at positions 352, 395, 434, 491 and 534 of the VP2 protein have been associated with the AMDV pathogenicity (Bloom *et al*., 1994; Fox *et al*., 1999; Hadlow *et al*., 1983, 1984; Oie *et al*., 1996), and all are different from those in non-pathogenic AMDV-G. These five aas in AMDV-G are different from those in the NS-CB and most of the 42 other global isolates used in the current study (Table 5.13). Because AMDV-G can propagate in cell culture, it

may be concluded that the NS-CB isolate cannot likely be propagated in cell culture. Of the 42 global isolates used in the current study, seven have the same five aas as in the NS-CB (MC42.2.1 and M173 from Newfoundland, Canada, BelI from Russia and four Danish strains: fDKNJ-320, -321, -322, and -327), but pathogenicity of these isolates have not been reported, and thus cannot be used to assess pathogenicity of the NS-CB. The strains whose pathogenicity have been studied include the highly pathogenic Utah1 (Bloom *et al*., 1998) and TR (Oie *et al*., 1996), and moderately pathogenic Pullman (Hadlow *et al*., 1983) and SL3 (Haas *et al*., 1991). The NS-CB isolate does not have the same five aas as in any of these strains but has the same aas at positions 395 (Q) and 534 (D) as in these pathogenic strains (Table 5.13). The NS-CB has different aas at positions 491 compared with Utha-1 and Pullman (D vs E), at positions 434 compared with TR (H vs N) and at three positions with SL3: 352 (V vs I), 434 (H vs N) and 491 (D vs N). These differences could suggest that NS-CB is not highly pathogenic.

Nineteen aa residues of the VP2 protein, from aa 428 to 446, have been reported to be associated with the immune-complex formation and AMDV pathogenesis (Bloom *et al*., 2001). Fifteen of these aas were conserved among the 42 isolates which were compared, and the other five had the same aas as in most other isolates. It can be concluded that NS-CB aas at this region has the potential to bind to the AMDV antibodies as efficient as other AMDV isolates, resulting in the antibody bounded virus entering into macrophages (Bloom *et al*., 1994) and maintain persistent infection (Bloom *et al*., 1994).

6.4. Hypervariable regions

The numbers of regions of the AMDV genome with high degrees of variability (HVR) which were identified in the current study were eight in the left ORF (Table 5.15 and Figure 5.1) and two in the right ORF (Table 5.16 and Figure 5.2), which are greater than the one HVR identified in the right ORF (VP2) by comparing the AMDV-G and Utah1 strains (Bloom *et al*., 1988), and the four HVR regions in the left ORF (NS1) identified by comparing 39 isolates (Olofsson *et al*.,1999). The numbers of HVR in the NS1 and VP2 proteins in the current study (6 and 3, respectively) (Table 5.17, Figure 5.3 and Table 5.18, Figure 5.4) were also greater than those previously identified, namely four in the NS1 by Olofsson *et al*. (1999) and two in the VP2 by Oie *et al*. (1966). The larger number of HVR identified in the current study is comparable with that in the previous reports. In previous studies (Bloom *et al*., 1988; Oie *et al*., 1996; Olofsson *et al*., 1999) the number of variable nts and aas were manually counted along the sections of AMDV DNA or protein sequences and regions with larger number of variable positions, compared with the variable positions in the adjacent regions, which were identified. In the current study, however, Shannon's entropy values, defined as the probability of a given nt or aa appearing at a position, were employed. In addition to using different approaches, there is some degree of ambiguity in both methods, because the boundaries of each HVR is set somewhat arbitrarily, as well as the entropy criteria used in the current study, i.e. regions with seven to 12 nts in length with at least one position with maximum entropy value greater than 0.8. There is no published report on the number of HVR in the AMDV genome using entropy to compare the results with.

Four of the eight HVR in the left ORF (L-HVR2 to L-HVR5) were at the same positions as those previously reported (Olofsson *et al*., 1999), therefore L-HVR1, L-HVR6, L-HVR7, and L-HVR8 were novel. Of the six HVR in the NS1 protein, two (NS1-HVR3 and HVR4) were previously reported (Olofsson et al., 1999) and four (NS1-HVR1, HVR2, HVR5, and HVR6) were novel. Only one of the two HVR in the right ORF (R-HVR2) was previously identified (Bloom *et al*., 1988; Oie *et al*., 1996), and one of the three HVR in the VP2 protein (VP2-HVR2) was previously reported (Bloom *et al*., 1988; Oie *et al*., 1996). The number of repeats in the GT-repeat region located between nts 2469 and 2514 of AMDV genome, coding for glycine residues, are variable among AMDV isolates (Canuti *et al*., 2016; Li *et al*., 2012). This variable region was identified in the current study as HVR in the right ORF (R-HVR1) and in the VP2 protein (VP2-HVR1) and suggests that the GT-repeat region has little effect on the virus function. Obviously, the reason for finding novel HVR in both nts and aa residues in the current study is the method of detection employed.

The finding that a larger number of HVR was present in the left- than in the right - ORF based on nts (8 vs 2) (Table 5.15 and Table 5.16) and aa residues (6 vs 3) (Table 5.17 and 5.18), suggests a greater degree of variability of the left ORF, which agrees with the reports of Canuti *et al*. (2016) and Hagberg *et al*. (2016). The biological significance of low levels of variability in the structural protein has not been fully studied. It may be hypothesized that a high degree of aa homology is required in the structural protein which controls host range (Bloom *et al*.,1993,1998; Fox *et al*., 1999) and antibody recognition (Bloom *et al*., 1994, 2001). None of the five aa residues (352, 395, 434, 491, and 534) that have been suggested to modulate pathogenicity fell into the three HVR (Table 5.16) of the

VP2 in current study, which is in concordance with the published reports (Bloom *et al*., 1993; Oie *et al*., 1996) indicating that aa residues of the VP2-HVR2 were not related to AMDV pathogenicity.

The larger degree of variability of the left ORF, particularly HVRs, have been used in studying virus classification, identifying the source of outbreaks and AMDV strain identification (Christensen *et al*., 2011; Jensen *et al*., 2011; Knuuttila *et al*., 2009; Leiman *et al*., 2015; Nituch *et al*., 2012; Persson *et al*., 2015; Ryt-Hansen *et al*., 2017). Most of these studies used nt 563-952, which includes HVR1-4. The right ORF has been used less frequently for these purposes (Bloom *et al*.,1988; Oie *et al*., 1996) use the region of nt 3094-3131 for differentiation of different AMDV strains, which is located in the HVR2.

6.5. Recombination analysis

One of the major objectives of this study was to find the evolutionary relationship of the NS-CB with global AMDV isolates by phylogenetic analysis. Recombination analysis was performed prior to the phylogenetic analysis because recombination leads to the incongruent topologies in inferred phylogenetic trees due to the differences in evolutionary histories of the exchanged segments of the sequences (Martin *et al*., 2015; Shacklton *et al*., 2007). In this study, the ECR region was selected for detecting all possible recombination events which had occurred in the NS-CB, which previously were performed in a few studies (Canuti *et al*., 2016; Hagberg *et al*., 2017). Of the 32 AMDV isolates studied, only 11 recombinants events were detected which involved eight isolates (Table 5.22 and Figure 5.5), and 24 isolates did not have any significant recombination. The finding that a higher number of recombination events occurred in the left ORF compared

to the right ORF agrees with the previous reports (Canuti *et al*., 2016; Hagberg *et al*., 2017). It has been suggested that low degree of recombination in the right ORF could be an adaptation to prevent deleterious effect of recombination by avoiding disruption of capsid protein folding, which could prevent the assembly of protein particles and formation of the 3D structure of the capsid (reviewed by Martin *et al*., 2011).

The VP1 and VP2 overlap region is the recombination hot spot common to the parvoviruses and the dependoviruses (Lefebvre *et al*., 2009), and the only two breakpoints in the R-ORF in the current study occurred in this region (nt. 2894 and 2976). These two breakpoints are located between nts 2553 and 3084, which is the region that determines the replication capacity of AMDV in cell culture (Bloom *et al*., 1993). It is possible to conclude that new characteristics, such as extended host-tropism or strains with different pathogenicity, may be originated in AMDV by exchanging segments of the genomes in this region by recombination events, as has been reported in other viruses, such as influenza virus (Lam *et al*., 2010).

Multiple infections is common in mink farms (Gottschalck *et al*., 1991; Canuti *et al*., 2016), particularly if farms are located in areas with a high concentration and the co-infection of mink with different strains of AMDV provide opportunities for recombination (Canuti *et al*., 2016). The absence of recombination in the NS-CB could be because this isolate originated from an isolated mink farm in NS. Another reason could be that the seven methods that were used for detecting the recombination did not have enough statistical power to detect recombination in the NS-CB due to the low level of genetic diversity. A minimum of 5% nt diversity in the data set was suggested for achieving high statistical power for detecting recombination (Posada and Crandall, 2001_b). Surprisingly, no

recombination was detected among 19 AMDV isolates collected from different farms in Denmark (Hagber *et al*., 2017). Based on the current recombination analysis, there was no region of all the 32 AMDV genomes without recombination (Figure 5.7), and thus the ECR of these isolates cannot be used for accurate phylogenetic analysis. Therefore, it is appropriate to estimate the evolutionary relationship of the NS-CB by comparing it with the AMDV sequences without recombination, i.e 19 Danish isolates, SL3, G, LM, and Beijing to avoid inconsistencies in phylogenies as suggested by Martin *et al*. (2015).

6.6. Phylogenetic analysis

The evolutionary relationship of the NS-CB was calculated in this study using 4143 bp of the sequence of the ECR, which is the first phylogenetic analysis performed for the AMDV isolates circulating in NS. Many of the phylogenetic analyses (16 of 21) were performed using shorter than 500 bp sequences of the genome (Table 2.5), which can be generated from one sequencing run by the conventional Sanger sequencing method with low cost. The segments of the NS1 gene have been more frequently used in the phylogenetic analysis of AMDV (nine analyses, Table 2.5) than the VP2 region (six analyses, Table 2.5) because of the higher genetic variability of the NS1 compared with VP2 region. Although the use of longer sequences in tree building in AMDV improves the tree reliability (Hagberg *et al*., 2017; Leimann *et al*., 2015), there is only one report on the use of longer than ECR fragment (Hagberg *et al*., 2017). The use of a 4143 bp segment of the genome makes the current analysis more accurate than most previous reports.

More than 50 computer programs are available for phylogenetic tree constructions, and each has advantages and disadvantages for the particular data set. The phylogenetic

analysis for the NS-CB was performed in this study using the ML method implanted in MEGA7, which was used more than any other software for phylogenetic analysis of AMDV (Table 2.6). The reason is likely the user-friendly settings in MEGA for the biologists. In addition, software that uses the Bayesian approach gaining popularity for the AMDV tree building (Table 2.6).

The bootstrap and posterior probability are the methods for measuring tree reliability and are specific to the computer program that is used to build the tree (Table 2.6). Phylogenetic analyses performed in the current study, i.e. using ML implemented in MEGA7, showed that lowering of bootstrap support for the NS-CB from 100% to 74% after removing the eight sequence with recombination. However, the bootstrap value for the NS-CB remained at an acceptable confidence level above 70% (Felsenstein, 1985). One of the reasons for the reduced bootstrap support for the NS-CB after removing the recombinant strains could be the smaller number (24 vs. 32) suggested by Erixon *et al*. (2003). The reduced bootstrap support for the NS-CB could also be the result of increased sequence similarity in the dataset after the removal of recombinant sequences. Even though not explicitly stated, the effect of high sequence similarity of the dataset on lower bootstrap value for the phylogenetic tree construction was documented (Leimann *et al*., 2015).

The rooted phylogenetic tree provides the evolutionary relationships of the sequence of interest. The outgroup method is needed in rooting the phylogenetic trees, and seven of the 16 reports which used phylogenetic analysis of AMDV isolates used distinct sequences as outgroups, but mid-point rooting has not been previously used (Table 2.5). No explanation has previously been provided on the use of specific sequences as outgroups for AMDV analysis. It has been suggested that outgroup rooting may not be accurate for

viruses because of their heterogeneous evolutionary rates and the lack of a priory phylogenetic information on them (Hess and De Moraes Russo, 2007). Heterogeneous evolutionary rates were documented for AMDV (Canuti *et al*., 2016), which makes the use of outgroups somewhat questionable and favors the used of the midpoint rooting for AMDV. Midpoint rooting has been used to establish the evolutionary relationship of severe acute respiratory syndrome (SARS) virus (Stavrindes and Guttman, 2004).

The phylogenetic analysis in this study revealed that the NS-CB was closely related to 19 geographically distant Danish strains. This observation revealed a phenomenon of the transmission of AMDV via trading of breeding stock from North America to Europe, including Finland (Knuuttila *et al*., 2009), which took place in the 1930s and 1940s. No American isolate without recombination was used in phylogenetic analyses (Figure 5.6), but those with recombination were more closely associated with the NS-CB (Utah1, M195, WM25, M228, M173) than the Chinese isolates (NL1, LN2, LN3, LM, Beijing) (Figure 5.7). The absence of clustering according to the geographical origin has not been observed in previous studies (Knuuttila *et al*., 2009; Leimann *et al*., 2015; Wang *et al*., 2014), whereas some clustering according to the country or region of origin have been documented (Canuti *et al*., 2016; Christensen *et al*., 2011; Jensen *et al*., 2012; Nituch *et al*., 2012; Sang *et al*., 2012). The pathogenicity of the NS-CB and Danish isolates are not known, and it is not possible to comment on the relationships between NS-CB and Danish strains based on their pathogenicity. In previous studies, AMDV isolates have not been clustered according to the pathogenicity (Knuuttila *et al*., 2009, 2015; Nituch *et al*., 2012; Olofsson *et al*., 1999; Schuierer *et al*., 1997). Phylogenetic analysis showed that the

NS-CB is not closely related to the LM and Beijing strains from China. The reason could be that although mink were imported to China from Europe almost 69 years ago, many isolates were not closely related to current European isolates (Wang *et al*., 2014). The phylogenetic analysis also showed a close relationship between the AMDV-G and the SL3 with a high confidence level, which is in agreement with previous reports using the near-complete genome sequences (Hagberg *et al*., 2017), partial NS1 sequences (Canuti *et al*., 2016; Knuuttttila *et al*., 2015; Li et al., 2012; Xi *et al*., 2016) and VP2 sequence (Wang *et al*., 2014).

6.7. Sequence variation over time

Spleen samples were used for AMDV DNA extraction in this study because it is one of the primary sites of virus replication and sequestration (Bloom *et al*., 1994), and because of the high success rate of PCR amplification of the viral genome collected from this organ (Farid *et al*., 2015; Jensen *et al*., 2014; Knuuttila *et al*., 2009). Blood samples are easy to collect, but viral titer may be too low in the circulation during the chronic stage of the AMDV infection (Farid *et al*., 2018), resulting in PCR amplification failure.

Nearly complete sequences of nine and partial sequences of 11 NS-CB descendants were determined in this study (Table 5.25), for a total of 40,970 nt at 16 wpi and 36541 nt at 176 wpi. There were problems of obtaining complete sequences of two regions of the NS-CB descendants, namely from nts 1743 to 2783 (the GT repeat region), and from nts 4197 to 4602 (3' end of the sequences) (Table 5.25). A total of 80 PCR tests were performed for the region covering the nt 1743 to 2783 of the 11 incomplete sequences, using three different primer pairs and DMSO as PCR enhancer, of which 20% failed,

18.75% generated a single faint band, and 61.25% generated multiple bands. The products of the 16 reactions with faint bands were used as the targets for the second round of PCR amplification, of which only five samples generated short (<350 bp) sequences. Similarly, a total of 51 PCR reactions were performed on the 11 samples for the 4197 to 4602 region. The incidence of PCR failure, faint bands, and multiple bands were 52.95%, 25.49%, and 21.56%, respectively. The second round of amplification of faint bands of nine samples by recovering correct size amplicons from the agarose gels followed by sequencing resulted in the production of one short sequence. Incomplete sequences in this study were likely the result of the miss-match of the target and the primer sequences in some regions of the genomes. PCR failure, insufficient PCR products, and the incorrect target amplification are common features of the miss-match between the target and primer sequences (Whiley and Sloots, 2005). Premature termination of sequencing in the GT-repeat region was another reason for failure to obtain complete sequences. Attempts to amplify long stretches of viral DNA, which is required for obtaining the high concentration of clean viral DNA for next-generation sequencing (Hagberg *et al*., 2016, 2017) were not successful.

One of the main objectives of this study was to estimate the mutation rate (substitution rate) of the NS-CB isolate by sequencing its descendant's DNA collected at 16 and 176 wpi. The observation that the number of nucleotide substitutions at 176 wpi was almost 3.1 times greater than that at 16 wpi (113 vs 36) is logical, and clearly shows the accumulation of nt substitution over time. The estimates of mutation rate at 16 wpi (2.85×10^{-3}) is 3.1 times greater than that at 176 wpi (9.13×10^{-4}), a completely reverse relationship. Similar to the mutation rate based on the total number of nucleotides sequenced, the estimates of nt substitution positions/site/year for the NS-CB at 16 wpi

(1.5×10^{-2}) was smaller than that at 176 wpi (4.6×10^{-3}). If the rate of substitution per week, which was 36/16=2.25 per week, continued linearly, there would be 396 substitutions at 176 wpi rather than 113. Reduced rate mutation over time was reported by Nobusawa and Sato (2006) for influenza virus A in an *in vitro* study, who showed 2.6×10^{-3} and 0.5×10^{-3} substitutions/site/year after 48 and 71 h, respectively. The results imply that substitution rate overtime was not linear and the rate decreased at later times after inoculation. It can be speculated that such change in the rate of substitutions over time could have been because of the reduced viral replication in chronically infected mink where viruses were sequestrated in organs (Bloom *et al*., 1994; Best and Bloom, 2006), or purifying selection where harmful substitutions are eliminated (Lam *et al*., 2010). There is only one report on the nucleotide changes of AMDV over time, where no sequence variation was observed for 328 bp long fragment of the NS1 gene (from nt 578 to 951) for 12 animals inoculated with a Danish strain (Saeby/DEN/799.1/05) after 24 wpi (Jensen *et al*., 2014), which is contradictory to the result of the current study.

It is possible that the higher mutation rate in 16 wpi than 176 wpi is due to the accumulation of enough adaptation to the environment, and because during the late stages of infection beneficial mutations are not required. AMDV was considered as a different member of autonomous parvovirus because of the unusual replication and the pathogenicity (Alexandersen *et al*., 1986; Alexandersen and Bloom, 1987). DNA replication, RNA transcription, protein expression, and production of infectious progeny occurring at low levels during the chronic phase of the AD at the macrophages (Bloom *et al*., 1994). Therefore, it can be speculated that the low mutation rate at 176 wpi compared to the 16 wpi could be due to the low level of DNA replication. Alternatively, AMDV does

not develop strategies for persistence, such as establishing nonlytic phase of replication, or avoid the immune response, because of the unusual pathogenicity with elevated levels of antibodies as the immune response. The antibodies help the virus to entre into macrophages through the Fc receptors (Kanno *et al*., 1993). The mechanism that led to the high substitution rates in ssDNA viruses is not clear since the ssDNA viruses use the same host cellular replication machinery with high fidelity for the replication (Duffy *et al*., 2008; Shackelton *et al*., 2005). The possibilities of the use of only a part of the DNA replication machinery by the ssDNA viruses, or alternation of proofreading capabilities of the host cellular replication machinery and miss-match repair mechanisms in infected cells has been suggested (Shackelton *et al*., 2005).

The method of inferring the mutation rates for the NS-CB in the current study was different from the method of estimating mutation rates for the other parvoviruses. For instance, the analysis of 33 sequences derived from the field samples and GenBank, Streck *et al*. (2011) computed mutation rate (number of substitutions per site per year) for the capsid gene (VP1) and NS1 gene of the PPV at 3×10^{-4} for 1970-2010 data and 5×10^{-5} for 1980 to 2010 data. Shackelton *et al*. (2005) computed mutation rate of the VP2 gene of CPV using 91 sequences, and that of NS1 gene of this virus using 35 sequences, derived from clinical samples and GenBank dated from 1978 to 2004, and estimated mutation rate (number of substitutions per site per year) at 1.7×10^{-4} and 7.9×10^{-4}, respectively. The main differences in inferring the mutation rate in NS-CB and the PPV and CPV are two-fold. The first difference is the genomic region used for estimating the mutation rate. In this study, the mutation rate was established using a nearly complete genome and the partial genome sequence data. The mutation rates of PPV and CPV were established using the

individual genes. The second reason is the unknown time between the infection and the data collection in PPV and CPV in the above studies. The mutation rates were inferred for the PPV and CPV using time differences between the oldest and the most recent years of the data collection. Therefore, mutation rates established in this study for the NS-CB were more accurate than the mutation rate established for the PPV and CPV.

All 20 sequences of the NS-CB descendants were different from each other and different from the sequence of the NS-CB. It may be hypothesized that the 20 NS-CB descendants could have different antigenic properties compared to the NS-CB due to the transversions substitutions occurred over time. This hypothesis is supported by the finding of the association of genetic variation and the emergence of new antigenic variants of PPV, which resulted in the occurrence of reproductive failure in the swine industry after the vaccination due to the surface aa substitutions in field samples (Streck *et al*., 2011).

Twenty-seven nt deletions were detected in the GT repeat region of the NS-CB descendant # 315 (Table 5.26), which is similar to 27 deletions in the LN1 isolate from China (Li *et al*., 2012). There were 39 nt deletion in the WM25, M228, M195, and 3 nt deletion in M173 isolates from Newfoundland (Canuti *et al*., 2016). The one possible way of explaining for the deletions would be variation infidelity of the host polymerase on the template of the virus by falling off and rejoining in random places in the GT repeat region during the rolling hairpin replication (RHR) (Canceill *et al*., 1999). The other possible explanation would be based on the degree of conservation of the glycine-rich region. The glycine-rich region of the capsid protein is highly conserved among the parvoviruses (Chapman and Rossmann, 1993). The X-ray crystallography showed conformational changes in both of the empty capsid and the virion during the different stages in the life

cycle of the MVM (Agbandje-McKenna *et al*., 1998), particularly in the N-terminal regions of some capsid subunits and the aa residues with the small sizes and the absence of a side chain (glycine residues) provide the flexibility and functionality to the N-terminal regions of some capsid subunits (Wu and Rossmann, 1993). The study by Castellanos *et al*. (2013) revealed that mutations leading to the aa substitutions with aa with side chains in the glycine-rich region at positions 31 and 33 of the capsid gene did not affect the capsid assembly or stability but had an effect on the infectivity of MVM. It was concluded that changes in the molecular weight of the capsid with bulkier residues within the glycine-rich tract had no effect on capsid assembly and stability but impaired the infectivity of MVM. Therefore, it is possible to conclude that observation of aa deletion in glycine track of one of the NS-CB descendants in this study and previous studies could not have any effect on capsid assembly or stability. Therefore, nucleotide deletions observed at the GT repeat region indicate that AMDV could be a constrained variable number of glycine residues that coding from GT repeat region for the encapsulation in the capsid.

In this study, a higher percentage of transversional substitution was observed compared to transitional substitutions after 16 wpi (63.88% vs 36.11%) (Table 5.29), but the differences were smaller at 176 wpi (52.21% vs 47.78%). If the mutation happened in NS-CB randomly, transversional substitutions (purine-pyrimidine changes) should be observed twice as frequently as the transitional substitution (purine to purine or pyrimidine to pyrimidine changes) (Lyons and Lauring, 2017). However, in this study, non-random nt substitution bias proposed by Fitch (1967) was observed at 176 wpi, suggesting that transversions substitutions are likely to be lethal compared to the transitional substitutions as in RNA viruses described by Lyons and Lauring (2017), because transitions are

favorable to conserve the important biochemical nature of the aa (Zhang, 2000). The observed mutation rate was not the part of the miss incorporation of nt by Taq polymerase during the PCR, because high fidelity polymerase was used in this study. Besides, the bidirectional sequencing was employed in the study to minimize the incorrect identification of nt substitution, especially at ambiguous positions.

CHAPTER 7. CONCLUSION

Aleutian mink disease virus (AMDV) possesses a threat to wild and captive carnivore species, particularly the American mink. This study is the first to report the near-full length sequence of AMDV from Cape Breton, NS (NS-CB), which included the entire coding region (4143 bp) and 254 bp of the 3'end, for a total of 4397 bp, or 91.6% of the viral genome. The sequence obtained in this study is the first valuable tool for identifying sources of infection on mink farms as well as in feral mink in Cape Breton. Comparison of the NS-CB sequence with 34 to 42 sequences with the same size from GenBank revealed the presence of 10 hypervariable regions at the nt level and nine at the aa level, of which five and three, respectively, have not been previously reported. Comparisons of five aa that were known to influence the pathogenicity of AMDV suggested that NS-CB is not possibly highly pathogenic. Sequence comparisons led to the identification of 34 novel nucleotide substitutions and 18 novel aa in the NS-CB, suggesting that NS-CB diverged somewhat from global strains used in this comparison. A total of 11 recombination events were detected between NS-CB and 31 global strains. Recombination is caused by infection of mink with multiple viral strains and increases the risk of distinct AMDV strains arising in farmed mink and expands AMDV diversity. Genetic diversity is a key determinant of the ability of viruses to escape host immunity. Phylogenetic analyses using ECR of NS-CB and the 23 global isolates with no recombination provided an accurate picture of the evolutionary relationship among the isolates with a high degree of resolution. The NS-CB was closely associated with 19 Danish strains.

The second objective of the study was to obtain, for the first time, an accurate estimate of the mutation rate of AMDV. Of the mink which were inoculated with the

NS-CB, 10 were sampled after each of 16 and 176 wpi, and their near-full length sequences were analyzed. The number of nucleotide substitutions at 176 wpi was almost 3.1 times greater than that at 16 wpi (113 vs 36) whereas the estimates of mutation rate at 176 wpi was 3.1 times lower than that at 176 wpi (2.85×10^{-3} vs 9.13×10^{-4} substitutions/ site/ year), showing a decreasing trend in the mutation rate per unit of time. The rate of spontaneous mutation is a major determinant of viral diversity and evolution and influences the risk of disease emergence at the epidemiological level. Increases in the number of nucleotide substitutions over time and subsequent divergence make it more difficult to accurately trace back a virus isolate to its origin when several years elapsed between the two samplings. It should be noted that the rate of spontaneous mutations *in vivo* is complicated by several factors, namely the number of viral generations (i.e., infection cycles), the removal of deleterious mutations by the selection, and genetic drift.

BIBLIOGRAPHY

Aasted, B., Alexandersen, S., & Christensen, J. (1998). Vaccination with Aleutian mink disease parvovirus (AMDV) capsid proteins enhances disease, while vaccination with the major non-structural AMDV protein causes partial protection from disease. *Vaccine*, *16*(11-12), 1158–1165.

Agbandje-McKenna, M., Llamas-Saiz, A. L., Wang, F., Tattersall, P. & Rossmann, M. G. (1998). Functional implications of the structure of the murine parvovirus, minute virus of mice. *Structure*, *6*(11), 1369–1381.

Alexandersen, S., & Bloom, M. E. (1987). Studies on the sequential development of acute interstitial pneumonia caused by Aleutian disease virus in mink kits. *Journal of virology*, *61*(1), 81-86.

Alexandersen, S., Uttentha-Jensen, A., & Aasted, B. (1986). Demonstration of non-degraded Aleutian disease virus (ADV) proteins in lung tissue from experimentally infected mink kits. *Archives of virology*, *87*(1-2), 127-133.

Alexandersen, S., Bloom, M. E., & Wolfinbarger, J. B. (1988). Evidence of restricted viral replication in adult mink infected with Aleutian disease of mink parvovirus. *Journal of virology*, *62*(5), 1495-1507.

Alfaro, M. E., Zoller, S., & Lutzoni, F. (2003). Bayes or Bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular biology and evolution, 20*(2), 255-266.

Batista, M. V. A., Ferreira, T. A. E., Freitas, A. C., & Balbino, V. Q. (2011). An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. *Infection, genetics and evolution*, *11*(8), 2026-2033.

Beerenwinkel, N., Gunthard, H. F., Roth, V., Metzner K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next -generation sequencing data. *Frontiers in microbioloy*, *3*, 329. doi: 10.3389/fmicb.2012.00329

Berger, G. (2015). Escape of pathogens from the host immune response by mutations and mimicry. Possible means to improve vaccine performance. *Medical hypothesis*, *85*(5), 664-669.

Bergeron, J., Hébert, B. & Tijssen, P. (1996). Genome organization of the Kresse strain of porcine parvovirus: identification of the allotropic determinant and comparison with those of NADL-2 and field isolates. *Journal of virology, 70*(4), 2508-2515.

Best, S. M., & Bloom, M. E. (2006). Aleutian mink disease parvovirus. In J. Kerr, S. Cotmore, & M. E. Bloom (Eds), *Parvoviruses* (457–471). London, UK: Hodder Arnold Publication.

Best, S. M., Shelton, J. F., Pompay, J. M., Wolfinbarger, J. B., & Bloom, M. E. (2003). Caspase cleavage of the non- structural protein NS1 mediates the replication of Aleutian mink disease parvovirus. *Journal of virology*, *77*(9), 5305-5312.

Bloom, M. E., Alexanderson, S., Perryman, S., Lechner, D., & Wolfinbarger, J. B. (1988). Nucleotide sequence and genomic organization of Aleutian mink disease parvovirus (ADV): sequence comparison between a nonpathogenic and a pathogenic stain of ADV. *Journal of virology*, *62*(8), 2903-2915.

Bloom, M. E., Berry, B. D., Wei, W., Perryman, S., & Wolfinbarger, J. M. (1993). Characterization of chimeric full-length molecular clones of Aleutian mink disease parvovirus (ADV): Identification of a determinant governing replication of ADV in cell culture. *Journal of virology, 67*(10), 5976-5988.

Bloom, M. E., Best, S. M., Hayes, S. F., Wells, R. D., Wolfinbarger, J. B., McKenna, R., & MaKenna, M. A. (2001). Identification of Aleutian mink disease parvovirus capsid sequences mediating Antibody-dependent enhancement of infection, virus neutralization and immune complex formation. *Journal of virology*, *75*(22), 11116-11127.

Bloom, M. E., Fox, J. M., Berry, D., Oie, K. L., & Wolfinbarger, J. B. (1998). Construction of pathogenic molecular clones of Aleutian mink disease parvovirus that replicate on both *in vivo* and *in vitro*. *Virology*, *251*(2), 288-296.

Bloom, M. E., Kanno, H., Mori, S., & Wolfinbarger, J. B. (1994). Aleutian mink disease: puzzles and paradigms. *Infectious agents and diseases*, *3*(6), 279-301.

Bodewes, R., Ruiz-Gonzalez, A., Schapendonk, C. M., Van den Brand, J. M., Osterhaus, A. D., & Smits, S. L. (2014). Viral metagenomic analysis of feces of wild small carnivores. *Journal of Virology*, *11*, 89. doi:10.1186/1743-422X-11-89

Canceill, D., Viguera, E., & Ehrlich, S. D. (1999). Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. *The journal of biological chemistry*, *274*(39), 27481-27490.

Canuti, M., Doyle, H. E., Britton, A. P., & Lang, A. S. (2017). Full genetic characterization and epidemiology of a novel amdoparvovirus in striped skunk (*Mephitis mephitis)*. *Emerging microbes and infections*, *6*(5), 1-8.

Canuti, M., O̓ Leary, K. E., Hunter, B. D., Spearman, G., Ojkic, D., Whitney, H. G., Lang, A. S., (2016). Driving forces behind the evolution of the Aleutian mink disease parvovirus in the context of intensive farming. *Virus evolution*, *2*(1), vew004. doi:10.1093/ve/vew004

Canuti, M., Whitney, A. H., & Lang, A. S. (2015). *Amdoparvoviruses* in small mammals: expanding our understanding of parvovirus diversity, distribution, and pathology. *Frontiers in microbiology*, *6*, 1119. doi:10.3389/fmicb.2015.01119

Castelruiz, Y., Blixenkrone-Møller, M., & Aasted, B. (2005). DNA vaccination with the Aleutian mink disease virus NS1 gene confers partial protection against disease. *Vaccine*, *23*(10), 1225-1231.

Castellanos, M., Pérez, R., Rodríguez-Huete, A., Grueso, E., Almendral, J. M & Mateu, M. G. (2013). A slender tract of glycine residues is required for translocation of the VP2 protein N-terminal domain through the parvovirus MVM capsid channel to initiate infection. *Biochemical journal*, *455*, 87-94.

Chang, S. F., Sgro, J. Y., & Parish, C. R. (1992). Multiple amino acids in the capsid structure of canine parvovirus coordinately determine the canine host range and specific antergenic and hemagglutination properties. *Journal of virology*, *66*(12), 6858-6867.

Chapman, M. S.*, &* Rossman, M. G. (1993). Structure, sequence, and function correlations among parvovirus. *Virology, 194*(2), 491-508.

Cheng, F., Chen, A. Y., Best, S. M., Bloom, M. E., Pinter, D., & Qiu, J. (2010). The capsid protein of Aleutian Mink Disease virus activate caspases and are specifically cleaved during infection. *Journal of virology*, *84*(6), 2687-2696.

Cho, H. J., & Ingram, D. G. (1972). Antigen and antibody in Aleutian disease in mink: Precipitation reaction by agar- gel electrophoresis. *Journal of immunology*, *108*(2), 555-557.

Christensen, L. S., Gram-Hansen, L., Chriél, M., & Jensen, T. H. (2011). Diversity and stability of Aleutian mink disease virus during bottleneck transitions resulting from eradication in domestic mink in Denmark. *Veterinary microbiology, 149*(1-2), 64-71.

Cotmore, S. F., Agbandje-McKenna, M., & Chiorini, J. A. et al. (2014). The family *Parvoviridae*, *159*(5), 1239-1247.

Dam-Tuxen, R., Dahl, J., Jensen, T. H., Dam-Tuxen, T., & Struve, T. (2014). Diagnosing Aleutian mink disease infection by a new fully automated ELISA or by counter current immunoelectrophoresis: a comparison of sensitivity and specificity. *Journal of virological methods*, *199*, 53-60.

Drummond, A. J., Pybus, O. G., Rambaut, A., Frosberg, R., & Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in ecology and evolution*, *18*(9), 481-488.

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rate of evolutionary changes in viruses: patterns and determinant. *Nature review genetics*, *9*, 267-276.

Dwivedi, B., & Gadagkar, S. R. (2009). Phylogenetic inference under varying proportions of indel-induced alignment gaps. *Evolutionary biology*, *9*, 211. doi:10.1186/1471-2148-9-211

Eckert, K. A., & Kunkel, T. A. (1991). DNA polymerase fidelity and the polymerase chain reaction. *Genome research*, *1*, 17-24.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792–1797.

Erixon, P., Svennblad, B., Britton, T., & Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic biology, 52*(5), 665-73.

Farid, A. H. (2013). Aleutian mink disease in furbearing mammals in Nova Scotia, Canada. *Acta veterinaria. scandinavica*, *55*, 10. doi: org/10.1186/1751-0147-55-10

Farid, A. H., Daftarian, P. M., & Fatehi, J. (2018). Transmission dynamics of Aleutian mink disease virus on farm under test and removal scheme. *Journal of veterinary science and medical diagnosis, 7*, 2. doi: 10.4172/2325-9590.1000253

Farid, A. H., & Ferns, L. E. (2011). Aleutian mink disease virus infection may cause hair depigmentation. *Scientifur, 35*(4), 55-59.

Farid, A. H., Hussain, I., & Arju, I. (2015). Detection of Aleutian mink disease virus DNA and antiviral antibodies in American mink *(Neovison vision)* 10 days post inoculation. *Journal of veterinary diagnostic investigation*, *27*(3), 287-294.

Farid, A. H., & Segervall, J. (2014). A Comparison between ELISA and CIEP for measuring antibody titers against Aleutian mink disease virus. *Virology and Mycology*, *3*(3), 137.

Farid, A. H., Zillig, M. L., Finley, G. G., & Smith, G. C. (2012). Prevalence of the Aleutian mink disease virus infection in Nova Scotia. *Preventive veterinary Medicine, 106*(3-4), 332-338.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*(6), 368-376.

Fox, J. M., Stevenson, M. A. M., & Bloom., M. E. (1999). Replication of Aleutian mink disease parvovirus in vivo is influenced by residues in VP2 protein. *Journal of virology*, *73*(10), 8713-8719.

Fitch, W. M. (1967). Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *Journal of molecular biology*, *26*(3), 499–507.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, *162*(3), 705-708.

Gottschalck, E., Alexandersen, S., Cohn, A., & Poulsen, L. A., Bloom, M. E., & Aasted, B. (1991). Nucleotide sequence analysis of Aleutian mink disease parvovirus shows that multiple virus types are present in infected mink. *Journal of virology*, *65*(8), 4378-4386.

Gottschalck, E., Alexanderson, S., Storgaard, T., Bloom, M. E., & Aasted, B. (1994). Sequence comparison of the non- structural genes of four different types of Aleutian Mink Disease parvovirus indicates an unusual degree of variability. *Archives of virology, 138*(3-4), 213-231.

Gunnarsson, E. (2001). Documenting freedom from disease and re-establishing a free status after a breakdown Aleutian disease (plasmacytosis) in farmed mink in Iceland. *Acta veterinaria Scandinavica*, *42*, 1. doi:org/10.1186/1751-0147-42-S1-S87

Hadlow, W. J., Race, R. E., & Kennedy, R. C. (1983). Comparative pathogenicity of four strains of Aleutian disease virus for pastel and sapphire mink. *Infection and immunity 41*(3), 1016-1023.

Hadlow, W. J., Race, R. E., & Kennedy, R. C. (1984). Royal pastel mink respond variously to inoculation with Aleutian disease virus of low virulence. *Journal of virology*, *50*(1), 38-41.

Hagberg, E. E., Krarup, A., Fahnoe, U., Larsen, L. E., Dam-Tuxen, R., & Pedersen, A. G. (2016). A fast and robust method for whole genome sequence of the Aleutian mink disease virus (AMDV) genome. *Journal of virological methods, 234*, 43-51.

Hagberg, E. E., Pedersen, A. G., Larsen, L. E., & Krarup, A. (2017). Evolutionary analysis of whole-genome sequences confirms inter-farm transitions of Aleutian mink disease virus. *Journal of general virology, 98*(6), 1360-1371.

Hall, B. G. (2013). Building phylogenetic trees from molecular data with MEGA. *Molecular biology and evolution*, *30*(5), 1229-1235.

Hall, B. G., & Barlow, M. (2006). Phylogenetic analysis as a tool of molecular epidemiology of infectious diseases. *Annals of epidemiology*, *16*(3), 157-169.

Hall, B. G., Pikis, A., & Thompson, J. (2009). Evolution and biochemistry of family 4 glycosidases: implications for assigning enzyme function in sequence annotations. *Molecular biology and evolution*, *26*(11), 2487–2497.

Hass, L., Wohlsein, P., Trautwein, G., Stolze, B., & Kaaden, O. R. (1991). Violet mink develop an acute disease after experimental infection with Aleutian disease virus (ADV) isolate ADV-SL-3. *Journal of veterinary medicine*, *37*(2), 106-117.

Hess, P. N., & De Moraes Russo, C. A. (2007). An empirical test of the midpoint rooting method. *Biological journal of the Linnean society*, *92*(4), 669-674.

Hoeizer, K., Shackelton, L. A., & Parish, C. R., (2008). Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic acids research 31*, 3026-3037.

Holland, B. R., Penny, D., & Hendy, M. D. (2003). Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - a simulation study. *Systematic biology*, *52*(2), 229-238.

Huang, Q., Luo, Y., Cheng, F., Best, S. M., Bloom, M. E., & Qiu, J. (2014). Molecular characterization of the small nonstructural proteins of parvovirus Aleutian mink disease virus (AMDV) during infection. *Virology*, *452-453*, 23-31.

Hussain, I., Price, G. W., & Farid, A. H. (2014). Inactivation of Aleutian mink disease virus through high temperature exposure *in vitro* and under field-based composting conditions. *Veterinary microbiology*, *173*(1-2), 50-58.

Jahns, H., Daly, P., McElroy, M. C., Sammin, D. J., Bassett, H. F., & Callanan, J. J. (2010). Neuropathological features of Aleutian disease in farmed mink in Ireland and molecular characterization of Aleutian mink disease virus detected in brain tissues. *Journal of veterinary diagnostics and investigation*, *22*(1), 101-105.

Jakubczak, A., Kowalczyk, M., Kostro, K., Horecka, B., & Jezewska-witkowska, G., (2016). High molecular polymorphism of the hypervariable region in the VP2 gene of Aleutian mink disease virus. *Acta virologica*, *60*(54), 354–360.

Jakubczak, A., Kowalczyk, M., Kostro, K., & Jezewska-witkowska, G. (2017). Comparative molecular analysis of strains of the Aleutian disease virus isolated from farm and wild mink. *Annals of agricultural and environmental medicine*, *24*(3), 366-371.

Jensen, T. H., Christensen, L. S., Chriél, M., Harslund, J., Charlotte, M., & Salomonsen, C. M. (2012). High prevalence of Aleutian mink disease virus in free-ranging mink on a remote Danish island. *Journal of wildlife diseases, 48*(2), 497-502.

Jensen, T. H, Hammer, A. S., & Chriél, M. (2014). Monitoring chronic infection with a field strain of Aleutian mink disease virus. *Veterinary microbiology*, *168*, 420-427.

Kanno, H., Wolfinbarger, J. B., & Bloom, M. E. (1993). Aleutian mink disease parvovirus infection of mink peritoneal macrophages and human macrophages cell lines. *Journal of virology*, *67*(4), 2075-2082.

Kato, N., Sekiya, H., Ootsuyama, Y., Nakazawa, T., Hijikata, M., Ohkoshi, S., & Shimotohno, K. (1993). Humoral immune response to hypervariable region 1 of the putative envelope glycoprotein (gp70) of hepatitis C virus. *Journal of virology*, *67*(7), 3923–3930.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis Version 7.0 for bigger datasets. *Molecular biology and evolution*, *33*(7), 1870-1874.

Kumar, S., Tamura, K., Jakobsen, I. B., & Nei, M. (2001). MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics*, *17*(12), 1244 -1245.

Knuuttila, A., Aronen, P., Eerola, M., Gardner, I. A., Virtala, A. M. K. & Vapalahti, O. (2014). Validation of an automated ELISA system for detection of antibodies to Aleutian mink disease virus using blood samples collected in filter paper strips. *Journal of virology*, *11*, 141. doi:10.1186/1743-422X-11-141

Knuuttila, A., Aaltonen, K., Virtala, A. M., Henttonen, H., Isomursu, M., Leimann, A., Maran, T., Sarma, U., Timonen, P., Vapalathi, O., & Sironen, T. (2015). Aleutian mink disease virus in free- ranging mustilids in Finland- a cross- sectional epidemiological and phylogenetic study. *Journal of general virology, 96*(6), 1423-1435.

Knuuttila, A., Uzcátegui, N., Kankkonen, J., Vapalahti, O., & Kinnunen, P. (2009). Molecular epidemiology of Aleutian mink disease virus in Finland. *Veterinary microbiology, 133*(3), 229-238.

Korber, B. T., Kunstman, K. J., Patterson, B. K., Furtado, M., Mcevilly, M. M., Robert, L., & Wolinsky, S. M. (1994). Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *Journal of virology*, *68*(11), 7467-7481.

Kowalczyk, M., Jakubczak, A., Horecka, B., & Kostro, K. (2018). A comparative molecular characterization of AMDV strains isolated from cases of clinical and subclinical infection. *Virus genes*. *54*, 561-569.

Ladner, J. T., Beitzel, B., Chain, P. S. G., Davenport, M. G., Donaldson, E. F., & Frieman, M. (2014). Standards for sequencing viral genomes in the era of high throughput sequencing. *mBio* 5(3), e01360-14. doi:10.1128/mBio.01360-14

Lam, T. T., Hon, C. C., & Tang, J. W. (2010). Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical. laboratory sciences*, *47*(1), 45-49.

Lefeuvre, P., Lett, J. M., Varsani, A., & Martin, D. P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *Journal of virology*, *83*(6), 2697-2707.

Legend, D. S., & Ginocchio, C. C. (2007). Role of cell culture for virus detection in the age of technology. *Clinical microbiology*, *20*(1), 49-78.

Leimann, A., Knuuttila, A., Maran, T., Vapalahti, O., & Saarma, U. (2015). Molecular epidemiology of Aleutian mink disease virus (AMDV) in Estonia, and a global phylogeny of AMDV. *Virus research, 199*, 56–61.

Leng, X., Liu, D., Li, J., Shi, K., Zeng, F., Zong, Y., Liu, Y., Sun, Z., Zhang, S., Liu, Y., & Du, R. (2018). Genetic diversity and phylogenetic analysis of Aleutian mink disease virus in north - east China. *Archives of virology*, *163*(5), 1241-1251.

Li, Y., Huang, J., Jia, Y., Du, Y., Jiang, P., & Zhang, R. (2012). Genetic characterization of Aleutian mink disease viruses isolated in China. *Virus genes*, *45*(1), 24-30.

Li, L., Pesavento, P. A., Woods, L., Clifford, D. L., Luff, J., Wang, C., & Delwart, E. (2011). Novel amdovirus in gray fox. *Emerging infectious diseases*, *17*(10), 1876-1878.

Lyons, D. M., & Lauring, A. S. (2017). Evidence for the selective basis of transition to transversion bias in two RNA viruses. *Molecular biology and evolution*, *34*(12), 3205-3215.

Mañas, S., Cena, J. C., Ruiz-Olmo, J., Palazons, S., Domingo, M., & Wolfinbarger, J. B. (2001). Aleutian mink disease parvovirus in wild riparian carnivores in Spain. *Journal of wildlife diseases*, *37*(1), 138-144.

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., & Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses, 3*, 1699-1738.

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution*, *1*(1), vev003. doi:10.1093/ve/vev003

McCormack, G. P., & Clewley, G. P. (2002). The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Reviews in medical virology*,*12(*4), 221-238.

McKenna, R., Olson, N. H., Chipman, P. R., Baker, T. S., Booth, T. F., Christensen, J., Asted, B., Fox, J. M., Bloom, M. E., & Mckenna, M. A. (1999). The three-dimensional structure of Aleutian mink disease parvovirus: implication for disease pathogenicity. *Journal of virology*, *73*(8), 6882-6891.

Murphy, W. J., E. Eizirik, S. J., & O'Brien et al. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, *294* (5550), 2348– 2351.

Nei, M., & Kumar, S. (2000). *Molecular evolution and phylogenetics*. New York, NY: Oxford University Press.

Nituch, L. A., Bowman, J., Wilson, P., & Schulte-Hostedde, A. I. (2012). Molecular epidemiology of Aleutian disease virus in free-ranging domestic, hybrid, and wild mink. *Evolutionary applications*, *5*(4), 330–340.

Nobusawa, E., & Satio, K. (2006). Comparison of mutation rates of human influenza A and B viruss. *Journal of Virology. 80*(7), 3675-3678.

Nüesch, J. P., & Rommelaere, J. (2006.) NS1 interaction with CKII alpha: novel protein complex mediating parvovirus-induced cytotoxicity. *Journal of virology*, *80*(10), 4729-4739.

Ogata, N., Alter, H. J., Miller, R. H., & Purcell, R. H. (1991). Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. *Proceedings of the national academy of sciences of USA*, *88*(8), 3392-3396.

Oie, K. L., Durran, G., Wolfinbarger, J. B., Martin, D., Costello, F., Perryman, S., Hogan, D., Hadlow, W. J., & Bloom, M. E. (1996). The relationship between capsid protein (VP2) sequence and pathogenicity of Aleutian mink disease parvovirus (AMD): a possible role for raccoons in the transmission of AMD infection. *Journal of virology, 70*(2), 852-861.

Olofsson, A., Mittelholzer, C., Berndtsson, L. T., Lind, L. Mejerland, T., & Belák, S. (1999). Unusual, high genetic diversity of Aleutian mink disease virus. *Journal of clinical microbiology*, *37*(12), 4145-4149.

Persson, S., Jensen, T. H., Blomström, A. L., Appelberg, M. T., & Magnusson, U. (2015). Aleutian mink disease virus in free-ranging mink from Sweden. *PloS one*, *10*(3), e0122194. doi:10.1371/journal.pone.0122194

Posada, D. & Crandall, K. A.(2001_a). Selecting the best-fit model of nucleotide substitution. *Systematic biology*, *50*(4), 580–601.

Posada, D., & Crandall, K. A. (2001_b). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the national academy of sciences of the United States of America*, *98*(24), 13757–62.

Qiu, J., Cheng, F., Burger, L. R., & Pintel, D. (2006). The transcription profile of Aleutian mink disease virus (AMDV) in CRFK cells is generated by alternative processing of pre-mRNAs produced from a single promotor. *Journal of virology*, *80* (2), 654-662.

Rasler, M., Querfurth, R., Warnatz, H. J., Lehrach, H., Yaspo, M., & Krobitsch, S. (2006). An efficient and economic enhancer mix for PCR. *Biochemical and biophysical research communication*, *347*(3), 747-751.

Reichert, M., & Kostro, K. (2014). Effect of persistent infection of mink with Aleutian mink disease virus on reproductive failure. *Bulletin of the veterinary institute in pulawy. 58*, 369-373.

Ronquist, F. & Huelsenbeck, J. P. (2005). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572-1574.

Ryt-Hansen, P., Hagberg, E. E., Chriél, M., Struve, T., Pedersen, A. G., Larsen, L. E., & Hjulsager, C. K. (2017a). Global phylogenetic analysis of contemporary Aleutian mink disease viruses (AMDVs). *Virology journal, 14*, 1. doi: 10.1186/s12985-017-0898-y

Ryt-Hansen, P., Hjulsager, C. K., Hagberg, E. E., Chriél, M., Struve, T., Pedersen, A. G., & Larsen, L. E. (2017b). Outbreak tracking of Aleutian mink disease virus (AMDV) using partial NS1 gene sequencing. *Virology journal, 149*, 119. doi: 10.1186/s12985-017-0786-5

Saifuddin, M., & Fox, J. G. (1996). Identification of a DNA segment in ferret Aleutian disease virus similar to a hypervariable capsid region of mink Aleutian disease parvovirus. *Archives of virology*, *141*(1), 1329-1336.

Sanderson, M. J., & Shaffer, H. B. (2002). Troubleshooting molecular phylogenetic analyses. *Annual review of ecology and systematics*, *33*, 49-72.

Sang, Y., Ma, J., Hou, Z., & Zhang, Y. (2012). Phylogenetic analysis of the VP2 gene of Aleutian mink disease parvoviruses isolated from 2009 to 2011 in China. *Virus genes*, *45*(1), 31-37.

Sanjuán, R. (2012). From Molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. *PLoS pathogens*, *8*(5), e1002685. doi:10.1371/journal.ppat.1002685

Sanjuán, R., & Domingo-Calap, P. (2016). Mechanism of viral mutation. *Cellular and molecular life sciences*, *73*(22), 4433-4448.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation rates. *Journal of virology, 84*(19), 9733-9748.

Schuierer, S., Bloom, M. E., Kaaden, O. R., & Truyen, U. (1997). Sequence analysis of the lymphotropic Aleutian disease parvovirus ADV-SL3. *Archives of virology, 142*(1), 157-166.

Shackelton, L. A., Hoelzer, K., Parrish, C. R., Holmes, E. C. (2007). Comparative analysis reveals frequent recombination in the parvoviruses. *Journal of general virology, 88*(Pt12), 3294-3301.

Shackelton, L. A., & Holmes, E. C. (2006). Phylogenetic evidences for the rapid evolution of human B19 erythrovirus. *Journal of virology*, *80*(7), 3666-3669.

Shackelton, L. A., Parrish, C. R., & Holmes, E. C. (2006). Evolutionary basis of codon usage and composition bias in vertebrate DNA viruses. *Journal of molecular evolution*, *62*(5), 551-563.

Shackelton, L. A., Parish, C. R., Truyen, U., & Holmes, E. C. (2005). High rate of evolution associated with the emergence of carnivore parvovirus. *Proceedings of the national academy of science of USA, 102*(2), 379-384.

Shao, X., Wen, Y., Ba, H., Zhang, X., Yue, Z., Wang, K., Li, C., Qiu, J., & Yang, F. (2014). Novel amdoparvovirus infecting farmed raccoon dogs and arctic foxes. *Emerging infectious diseases*, *20*(12), 2085-2088.

Som, A. (2014). Causes, consequences and solutions of phylogenetic incongruence. *Briefings in bioinformatics*, *16*(3), 536-548.

Statistics Canada, (2017). Number and value of mink pelts produced by color type. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210011501

Stavrinides, J., & Guttman, D. S. (2004). Mosaic Evolution of the severe acute respiratory syndrome coronavirus. *Journal of virology*, *78(*1),76-82.

Streck, A. F., Bonatto, S. L., Homeier, T., Souza, C. K., Goncalves, K. R., Gava, C. W. C., & Truyen, U. (2011). High rate of viral evolution in the capsid protein of procine parvovirus. *Journal of general virology*, *92*, 2628-2636.

Swofford, D. L. (2003). PAUP. *Phylogenetic analysis using Parsimony and other methods*, Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tarrío R, Rodríguez-Trelles, F., & Ayala, F. J. (2000). Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. *Molecular phylogenetic and evolution*, *16*(3), 344–349.

Troesch, M., Meunier, I., Lapierre, P., Lapointe, N., Alvarez, F., Boucher, M., & Soudeyns, H. (2006). Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein. *Virology*, *352*(2), 357-367.

Truyen, U., Gruenberg, A., Chang, S. F., Obermaier, B., Veijalaimen, P., & Parrish, C. R. (1995). Evaluation of the feline- subgroup parvoviruses and the control of canine host range *in vivo*. *Journal of virology*, *69*(8), 4702-4710.

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins: structure function and bioinformatics, 48*(2), 227-241.

Wang, Z., Wu, W., Hu, B., Zhang, H., Bai, X., Zhao, B. J., Zhang, L., Yan, X. (2014). Molecular epidemiology of Aleutian mink disease virus in China. *Virus research*, *184*, 14-19.

Wang, L., Zhi, B.,Wu, W., & Zhang, X. (2008). Requirement for shrimp caspase in apoptosis against virus infection. *Development and comparative immunology*, *32*(6), 706–715.

Whiley, D. M., & Sloots, T. P. (2005). Sequence variation in primer targets affects the accuracy of viral quantitative PCR. *Journal of clinical virology*, *34*(2), 104-107.

Williams, T. M. (2014). Evolution: rooting the eukaryotic tree of life. *Current biology*, *24*(4), 465-470.

Willems, L., Thienpont, E., Kerkhofs, P., Burny. A., Mammerickx, M., Kettmann, R. (1993). Bovine leukemia virus, an animal model for the study of intra strain variability. *Journal of virology*, *67*(2), 1086-1089.

Wohi, S., Schaffner, S., & Sabeti, P. C. (2016). Genome analysis of viral outbreak. *Annual Review of virology*, *3*(1), 173-195.

Wu, H., & Rossmann, M. G. (1993). The canine parvovirus empty capsid structure *Journal of molecular biology*, *233* (2), 231–244.

Xi, J., Wang, J., Yu, Y., Zhang, X., Mao, Y., Hou, Q., & Liu, W. (2016). Genetic characterization of the complete genome of an Aleutian mink disease virus isolated in north China. *Virus genes*, *52*(4), 463-473.

Yang, Z., & Rannala, B., (2012). Molecular phylogenetics: principles and practice. *Nature review genetics*, *13*, 303–314.

Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of molecular evolution*, *50*(1),56–68.