1-1-2013

Isolation and Characterization of a Full Length Retrotransposon: CR1

Cassandra Michelle Weason
University of Denver

ISOLATION AND CHARACTERIZATION OF A FULL LENGTH RETROTRANSPOSON: CR1

_____

A Thesis

Presented to

The Faculty of Natural Sciences and Mathematics

University of Denver

_____

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

_____

by

Cassandra M. Weason

June 2013

Advisor: Thomas W. Quinn

Author: Cassandra M. Weason
Title: Isolation and Characterization of a Full Length Retrotransposon: CR1
Advisor: Thomas W. Quinn
Degree Date: June 2013

Abstract

Transposable elements (TE) have been found in all genomes and have clearly had a major impact on genomic evolution. The described research takes advantage of an abundant transposable element present in the genomes of Anseriformes, called Chicken Repeat 1 (CR1). Previous studies in Anseriformes suggest that CR1 is presently active in recent evolutionary time (St. John, 2004). A fully functional CR1 element itself is approximately 4.5kb long (Kajikawa, 1997), where almost all inserts are truncated at the 5' end. Because of this, it has been a challenge to isolate a full length, active element. In this study, two CR1 sequences were obtained after screening a genomic library of Cape Barren Goose using probes complimentary to the flanking regions of the element. The findings unveiled sequences with a complete ORF1, ORF2, 3' untranslated region and a portion of the 5' untranslated region. This study gets one step closer a further understanding the transposition mechanism that are adopted by this class of TE's, a non-long terminal repeat (non-LTR) retransposons. Eventually, the capture of an active element could make it especially valuable for future research by investigating their ability to transpose in living cells.

Acknowledgements

First and foremost, I would like to give a big thank you to my advisor, Dr. Tom Quinn for giving me the opportunity to do research in his lab. His advice and guidance were extremely valuable throughout my time in the master's program. My committee members Dr. Nancy Sasaki and Dr. Robert Dores were also helpful. Both members gave me valuable feedback during the progression of my project. Nancy also gave me the platform to discuss my research by inviting me to give a poster presentation at the Society for Microbiologist symposium. I am also very grateful for the support and encouragement of previous Quinn lab members, Judy St. John and Mandip KC. Each were very interested in the progression of my project and offered me great advice. I would like for David Traylor for providing the DNA sample. Additionally, thank you to many of my family and friends, who have been extremely supportive of me.

Table of Contents

List of Figures

Chapter One: Introduction

Only a small fraction of most eukaryotic genomes encode for functional proteins or RNA products. Within noncoding regions, repetitious DNA elements called transposable elements (TEs) are the most abundant type of DNA found (Doolittle and Sapienza, 1980). As sequencing technologies have advanced, our knowledge of the composition and structure of genomes has also progressed. TEs have now been found in almost all eukaryotic genomes studied, with one exception being *Plasmodium falciparum* (Wicker *et al*., 2007). TEs are also common constituents of prokaryotic genomes (Kleckner, 1981). The fraction of genomes comprised of TEs differs drastically from organism to organism. For instance, the proportion of the genome comprised of TEs ranges as high as 95-99% in some plant species, such as *Lilium;* whereas, it is 77% in *Rana esculenta*, 27% in *Gallus gallus* and 35% in *Homo sapien* (Biemont and Viera, 2005 and Bestor, 2007). In eukaryotes it appears that differences in genome size are not due primarily to variation in gene number. Rather, they are due to the quantity of non-coding mobile elements, where by larger genomes typically contain higher proportions of TEs (Burke *et al*., 2002 and Bowen and Jordan, 2002). Although the mechanisms promoting genome expansion have yet to be elucidated, it is clear that non-coding DNA, such as TEs, play an important role in affecting genome sizes. Originally, these elements were viewed as molecular parasites that function simply to maintain their frequency in

populations without conferring any positive advantage to the individual organism (Doolittle and Sapienza, 1980). For this reason, Orgel and Crick (1980) referred to them as "selfish DNA". Despite the implication, TEs are essential in the evolution of complex, multicellular organisms. Their significance and long-term evolutionary consequences within hosts' genomes have been a topic of discussion by authors such as Kidwell and Lisch (2001), who would argue that the term "selfish DNA" is actually misleading. It is becoming increasingly clear that organisms have developed mechanisms to suppress TE activity, and, in some cases they are conscripted by the host genome into beneficial functions (Gombart *et al*., 2009). In fact, the previously disregarded characteristics of TEs have provided insight to a series of constructive mechanisms that make significant contributions to the evolution of their hosts. For instance, novel recombination events of the *Drosophila* P-elements, a DNA transposon, produce useful variation and might be of positive selective value (Thompson-Stewart *et al.*, 1994). In 2002, Morrish *et al*. examined the dynamics of the retrotransposons L1 and Alu in *Homo sapiens*. In this study, there was evidence in human cultured cells defective for DNA-repair displayed that L1 was able to repair double stranded breaks in DNA. Additionally, it has been observed that the mobility of retrosequences can lead to the formation of novel "chimeric retrogenes", which are capable of being expressed (Buzdin *et al.*, 2003). In Drosophila, there is an alcohol dehydrogenase gene that lost its protein-coding ability, rendering it non-functional (also known as a processed pseudogene). Long and Langley (1993) discovered that it utilizes the reverse transcriptase machinery produced by other retrosequences. Here, reverse transcription into downstream unrelated exons gave rise to

a novel gene. Another adaptive use of TEs is demonstrated in *Staphylococcus aureus*, where TEs carry genes that confer antibiotic resistance on their host cells and allow adaptation to adverse environmental conditions (Murphy *et al.*, 1985).

However, TEs also have deleterious effects on organisms. Broadly speaking, the mobility of TEs can cause large genomic rearrangements, including duplications, deletions, and insertions. For example, some breast cancer has been caused in humans due to a large deletion of genomic DNA resulting from the insertion of an L1 retrotransposon (Gilbert *et al.*, 2002). In certain populations of *Drosophila melanogaster*, the inheritance of P-elements can result in an array of associated problems, including temperature-dependent infertility and elevated rates of mutation (Kidwell, 1994). Lastly, multiple nonautonomous transposable elements inserted within the introns of a low-density lipoprotein receptor gene in humans facilitate unequal crossing-over of nearby homologous sequences (Hobbs *et al.*, 1985). The resulting deletion of an exon within the gene results in an heritable autosomal disorder called hypercholesterolemia and is characterized by elevated cholesterol levels. It appears mobile elements are moving targets that have adopted strategies sometimes in concert and sometimes in opposition with each other to achieve a balance between long-term beneficial effects and detrimental effects to many species (Kazazian, 2004).

TEs are classified according to the mechanism by which they transpose (Finnegan, 1989). Class I elements, often referred to as retrotransposons or retroposons, propagate throughout the genome in a "*copy and paste*" fashion and require an RNA intermediate (DNA-RNA-DNA). They can be further divided into two groups: those

containing long terminal repeats (LTR) at their termini and those that do not, called non-LTRs. LTR retrotransposons are marked by the presence of long terminal repeats of 200-600 base pairs (bp) flanking a central coding region (Wilhelm and Wilhelm, 2001). The mechanism by which they transpose is similar to that of retroviruses (Finnegan, 1992). The main difference is that retrotransposons with LTR's lack a functional envelope protein, which would allow for the movement from one cell to another; this has led to proposals that retroviruses evolved from LTR retrotransposons (Kazazian, 2004). More specifically, the left-hand LTR contains strong promoter and enhancer sequences that, once bound by a tRNA primer, initiate the transcriptional process which extends beyond the right LTR (Lodish *et al.*, 2008). The open reading frames of LTR retrotransposons code for, among other things, a putative reverse transcriptase and an integrase, both needed for transposition. Once the RNA exits the nucleus into the cytoplasm, it is translated and then reverse transcribed to yield double-stranded DNA. At this point, it is shuttled back into the nucleus and inserted into a new location in the genome. An interesting study by Garfinkel and colleagues in 1985 describes a system for studying Ty element (an LTR) transposition in *Saccharomyces cerevisiae*. A plasmid containing a Ty element fused to a 5' galactose induced promoter was introduced to *S. cerevisiae*. The levels of reverse transcriptase activity were compared to uninduced cells containing Ty. The presence of reverse transcriptase was virtually absent in uninduced cells. A study like this provided early evidence that Ty elements produce their own reverse transcriptase and, more broadly, raised questions regarding distant familial relationships of viruses and TEs.

Conversely, the non-LTRs constitute a diverse group of elements where the mechanism by which they transpose is less understood, but a general outline of the process exists. These elements are not as closely related to viruses as LTRs are, and they do not have long directed repeat sequences at their termini. Two subfamilies of non-LTRs exist called Long INterspersed Elements (LINEs) and Short INterspersed Elements (SINEs), where one (LINES) contains all the necessary encoding to become mobile, and the others (SINES) are nonautonomous and will "borrow" the necessary machinery to become mobile. Almost all non-LTR retrotransposons that are produced are truncated at the 5' end and become 'dead on arrival' (incapable of further retrotransposition) (Malik *et al.,* 1999). Furthermore, it appears that a majority of these elements serve no function to the host and accumulate mutations over time, such that older elements are more divergent than younger ones (Voliva *et al.,* 1983). In the case with autonomous LINEs, the RNA exits the nucleus into the cytoplasm, and the element-encoded proteins required for self replication and insertion are then translated from two open reading frames. Among these proteins are an endonuclease and a reverse transcriptase (RT). The endonuclease and RT proteins bind to the RNA forming a ribonucleoprotein complex which then accompanies the single stranded RNA back to the nucleus (Boeke, 2003). The transported RNA is subsequently reverse transcribed and simultaneously inserted in a new location (Wilhelm and Wilhelm, 2001). Moreover, reverse transcription begins by nicking one of the strands by the encoded endonuclease to reveal a free 3'-OH on the template strand. The nick serves as a starting point for the reverse transcriptase of the elemental RNA. Since reverse transcription occurs at the target site following cleavage, the integration

mechanism is referred to as target primed reverse transcription (TPRT) (Luan and Eickbush, 1995). With the completion of the human genome project in 2001, more information regarding the mammalian LINE (L1) element has become available (International Genome Sequencing Consortium, 2001). L1 demonstrates the typical structural features found in most non-LTR retrotransposons, which are 4-6 kb in length; contain two open reading frames (ORF), (where ORF 2 encodes a reverse transcriptase and endonuclease), long poly-A tracks and 5' truncations of variable lengths (Haas *et al.*, 1997). Alternatively, SINEs are also non-LTR retrotransposons but do not encode proteins for mobility and transpose passively (Finnegan, 1989). They are thought to do so through trans-regulation with LINE transcripts which share a common sequence on the 3' end with SINEs (Ta and Mao, 2004 and Deininger *et al.*, 2003). One SINE element called *Alu* is abundantly found in the human genome and demonstrates how SINEs compete with LINE RNAs for integration by binding LINE encoded proteins (Gu *et al*., 2000). Consequently, SINEs can be viewed as parasites of the LINEs.

The second classes of elements, called class II elements, are referred to as DNA transposons. In 1940, Barbara McClintock was the first to discover mobile entities within the maize genome, which happened to be DNA transposons called activator (Ac) and dissociation (Ds) elements. It was not until many years later that her idea of mobile elements was accepted when a similar bacterial element called IS (insertion sequence) was identified, and a basis for its transposition was elucidated. These elements integrate at new locations using a "*cut and paste*" mechanism by transposing from DNA to DNA. By contrast, these elements do not involve an RNA intermediate but contain a gene

encoding an active transposase enzyme between flanking inverted-repeat termini (Feschotte and Pritham, 2007). The transposase enzyme will bind to the inverted repeat on the template strand, perform the necessary strand breakage reaction and insert at a new site. Daughter insertions of most DNA transposons can occur at a large number of sites throughout a genome, because the target-site recognition is limited to a small number of nucleotides (Benjamin and Kleckner, 1989). Additionally, DNA transposons tend to integrate at sites in close proximity to the parental insertion (Kazazian, 2004).

A general feature present in all transposable elements are target site duplications (TSDs) that flank either side of the newly inserted element. TSDs are created when an encoded restriction endonuclease makes a staggered cut in the DNA strand to prepare the target site for TE insertion (Kidwell and Lisch, 2001).

The non-LTR retrotransposon of focus in this study is a taxonomically widespread, middle-repetitive sequence called Chicken Repeat One (CR1) (Figure 1.1). Stumph *et al*., (1981) were the first to characterize this element in the chicken. Analysis of the recently sequenced chicken genome revealed that 3% of the genome is comprised of CR1 alone and that it is present in over 200,000 copies (IHGSC, 2001 and Wicker *et al.,* 2004). Additionally, high numbers of CR1 have also been found in other avian genomes, including waterfowl (St. John *et al.,* 2004).

This element is also present in a variety of other animals including insects (Fabrick *et al*., 2011), reptiles (Novick *et al*., 2009 and Kajikawa *et al*., 1997) and others. The general structure of CR1 is similar to other non-LTRs, except that it does not contain the 3' polyadenylic acid (poly A) configuration (Silva and Burch, 1989). They carry all of

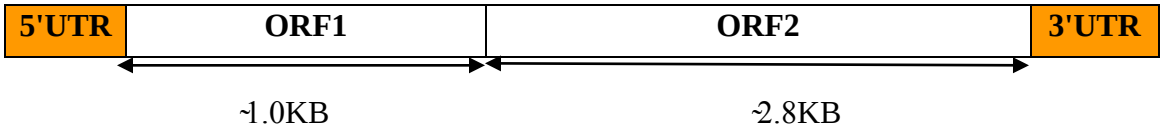


Figure 1.1. A general structure of a full-length CR1 non-LTR retrotransposon. The two open reading frames (ORF1 and ORF2) encode proteins necessary for successful retrotransposition with the approximate length of each indicated. The highlighted boxes shown are the 5'UTR and 3'UTR. In this study, the full length sequence was obtained by screening a genetic library of Cape Barren Goose and adapted from St. John and Quinn, 2008b.

the information needed for their transcription in two uninterrupted open reading frames (ORFs) flanked on either side by untranslated regions (5' and 3' UTRs). As with most non-LTRs, CR1 suffers from high rates of truncation at the 5' end. Reverse Transcriptase (RT) synthesizes DNA from the 3' end of the element, and for reasons not clearly understood, the transcriptional machinery often fails to successfully reverse transcribe a full-length element, at 4.5kb (Kajikawa *et al.*, 1997). The result is demonstrated in the Wicker *et al*. (2004) study of the chicken genome (*Gallus gallus*) where over 96,230 copies of CR1 found are truncated to various sizes less than 500 bp (IHGSC, 2001). In this study, it was discovered that only one CR1 contained two intact ORFs and was considered the candidate for a function element. However, it is not clear if it is currently active in *Gallus gallus*. Nonetheless, Haas *et al*. (2001) reported finding two complete CR1 sequences from chicken (CC and H3 in GenBank), but numerous mutations in both ORFs rendered them useless in terms of providing any further insight into the retrotransposition mechanisms. To date, there is no evidence to suggest the chicken genome contains an 'active' CR1 and harnessing one continues to escape researchers.

The UTRs appear to retain regulatory functions. Because of the numerous 5' truncations of this element, the 5'UTR is not clearly understood. It appears, however, it may serve as a promoter and further, that different CR1 families have "captured" different promoter sequences (Haas *et al*., 2001). The 3'UTR contains a conserved motif responsible for forming a hairpin structure where encoded proteins, such as RT, may dock onto and initiate reverse transcription (Haas *et al*., 2001). Adjacent to the 3'UTR (in the 3' direction), are 1-4 copies of a unique 8bp repeat sequence, 5'-NATTCTRT-3'

(Silva and Burch, 1989). The number of octomer repeats accumulates independently over time after the initial retrotranspositional event, suggesting CR1s with fewer octomer repeats are younger than CR1s with more than one (St. John and Quinn, 2008b).

The conserved structural features and proteins encoded by ORF2 are better known than ORF1 in CR1. Similar to ORF2 in L1, CR1 ORF2 encodes an RT and endonuclease (Kajikawa *et al*., 1997 and Boissinot *et al*., 2004). Evidence suggests that two domains encoded by ORF1 have been identified from the capture of a CR1 element in a swan called coscoroba (KC, 2008). Previous work done with CR1 regarding protein import into the cell nucleus suggested that ORF1 is also believed to encode a nucleic acid binding protein with localization signals to facilitate RNA transport back to the nucleus (Dingwall and Laskey, 1986).

In the previously mentioned study of the chicken genome done by Wicker *et al.* (2004), the characterizations of a majority of the 96,230 CR1 repeats were truncated, where only one copy was determined to have both intact ORF's. This element was considered to be the candidate for a functional element and is 4033 bp in length. Although it does not appear to contain a 5' promoter sequence, it is inserted into an A/T-rich region that may serve as elements for a promoter. In 2004, St. John *et al*. discovered a truncated CR1 within the third intron of the lactate dehydrogenase B gene of two Anseriforme species: coscoroba (*Coscoroba coscoroba*) and Cape Barren goose (*Cereopsis novaehollandieae*), a sister order and common ancestor to Galliformes. This discovery was absent in other Anseriformes (waterfowl) tested. Because mitochondrial DNA sequencing previously supported the taxonomic relationship between coscoroba

and Cape Barren goose (CBG), the value of this recent insertion is significant for phylogenic studies (Donne-Goussé *et al.*, 2002). Additionally, it has been clearly shown that a common ancestor of these two species diverged from other Anseriformes 9-11 million years ago, making this the most recent CR1 insertion noted to date (St. John *et al.*, 2004). This may suggest that coscoroba and Cape Barren Goose contain active element within their genomes.

It is clear that the size distribution of CR1 makes it such that very few full-length elements exist in the chicken genome. Because of this, the ability for researchers to capture a full-length CR1 element has been a challenge and has never been reported. Furthermore, because chicken (Galliforme) and Cape Barren Goose (Anseriforme) are sister orders to each other, it would be helpful to use what is known about the size distribution of CR1 in chicken and apply it to that of the CBG. In other words, one would also expect a majority of CR1s to be truncated (at the 5'-end), with very few full-length elements in the genome of CBG. However, since there is evidence that CR1 is actively transposing in CBG, one goal of the research presented here was to sequence a full-length CR1 element. The other goal was to isolate an actively transposing element as opposed to one that had been inactivated by subsequent mutation. CR1 would provide an ideal model to better understand the life cycle of other non-LTR retrotransposons, offer insight into how certain types of elusive TEs introduce new genes into genomes, or how non-LTRs disrupt genes.

Chapter Two: Materials and Methods

Tissue Source and DNA extraction

Blood was taken from a male adult Cape Barren Goose, *Cereopsis novaehollandiae*, and 450 µl was extracted according to the procedure outlined in Promega's Wizard Genomic DNA isolation kit to recover high molecular weight DNA. The sample was originally provided by David Traylor.

DNA digestion

Fifty micrograms (84.4 µl) of high molecular weight genomic DNA was fragmented in two 1125 µl partial restriction digests using the *Sau*3A I. The first reaction contained 2.5U *Sau*3A I and the second reaction contained 1.25U *Sau*3A I (4000 U/mL, Biolabs). Both digestions were set-up using 112.5 µl 10.0X NEB buffer, 12.5 µl of 100.0X BSA buffer, and 916 µl of H_2O to a total volume of 1125 µl. Reactions were incubated at 37°C for 30 minutes. After incubation, digestion was arrested via 5 µl of EDTA to inactivate the enzymes. Digested DNA was gently purified and extracted twice with phenol:cholorform and ethanol precipitated using 1/10 volume of 3M NaOAc (pH 5.2), two times the volume of cold 95% ethanol and a final rinse with 1ml of 70% ethanol. Sample was re-suspended in 200 µl of TE buffer.

Size fractionation of digested DNA

Partially digested DNA was fractionated in a 10%-40% sucrose density gradient made using a buffer containing 10mM Tris-Cl (pH 8), 10mM NaCl, and 1mM EDTA (pH8). DNA (200 µl) was loaded on top of 5.7 mL of the gradient and centrifuged at

22,400 rpm for 20 hours at 20°C in a Beckman SW41 rotor and centrifuge. Following centrifugation, fractions were collected in 350 µl aliquots, and 10 µl of each was visualized on a 0.3% agarose TAE gel poured on top of a 1% agarose support, using the molecular size markers *Xho*I lambda and *Hind*III cut lambda. The 350 µl aliquots containing the fragments with the desired range of 15,000- 23,000 bp were pooled and then dialyzed against two liters of TE buffer using the Thermo Scientific Slide-A-Lyzer MINI Dialysis device (10K MWCO). Fragments were purified via ethanol precipitation using the same procedure described above, with final resuspension in 8.5 µl of TE buffer. Using a 0.8% agarose TAE gel, the sample and known concentrations of DNA were loaded and compared and the final concentration was determined to be 0.05 µg/µl.

Ligation and Packaging into a Bacteriophage Vector

The bacteriophage Lambda DASH II/*Bam*H I vector kit by Agilent Technologies which preferentially size selects for larger inserts, was used to clone the fragmented genomic DNA. The target DNA that was previously digested with *Sau*3A I restriction enzyme resulted in 5'-GATC-3' overhangs complimentary to the *Bam*H I sites of the Lambda phage arms. Since DNA concentration was limited, an adjustment from the Agilent protocol was made for the ligation reaction in order to maintain an equimolar ratio of Lambda DASH II vector (0.5 µl, 0.5 µg), and fragmented DNA inserts (3.0 µl, 0.16 µg). The reaction also contained 0.5 µl of 10X ligase buffer, 0.5 µl of 10mM rATP, and 2U of T4 DNA Ligase (0.5 µl) for a total reaction volume of 5 µl. The reaction was incubated at 4°C overnight. The recombinant lambda phage was then packaged into two Gigapack III packaging extracts included in the kit (2.5 µl of ligation reaction mixture/packaging reaction). The reactions were incubated for 90 minutes at 22°C and

were then combined into a single tube. Finally, 500 µl SM buffer and 20 µl of chloroform were added, and samples were placed in storage at 4ºC.

Probe Preparation

Several complimentary DIG-labeled hybridization probes were designed through PCR. The 100 µl PCR included 5X Go Taq Buffer (Promega), 0.25 mM dNTPs, 10 µM forward/reverse oligonucleotide, 2 µl eluted DNA, 1.25U Go Taq DNA Polymerase (5U/µl, Promega), and H_2O to 100 µl. The thermal profile for the PCR was as follows: 94°C denaturation for 1 min., annealed at 55°C for 1 min., and 72°C extension for 2 min., 35 cycles. The amplification products were then extracted with phenol:chloroform and random primed labeling was done for 20 hours at 37°C according to the DIG High Prime DNA labeling and Detection Starter Kit #1 by Roche Applied Science. The DIG-labeled probes were complementary to a portion of the target sequence in ORF1 (5'end) and ORF2 (3'end) of CR1and were designed from a previously obtained consensus sequence from a putative full-length CR1 in *Coscoroba coscoroba*, a close relative of the Cape Barren Goose (KC, 2008) (Table 2.1). The primers were originally synthesized by Sigma-Aldrich.

Library Screening

Original titre and amplification of library. Screenings were done by preparing XL1-Blue MRA (P2) host cells according to the Agilent protocol. Before screenings were performed, various dilutions of the phage suspension in SM buffer were prepared in order to determine the efficiency of the library. These dilutions were added to 200 µl of prepared host cells and reactions were incubated in a sterile culture tube for 15 minutes at 37°C. Following incubation, the reaction mixture was added to 3mL of molten 0.7%

Table 2.1

List of Primers Used for DIG-labeled Probes

Probe	Primer sequence (5'-3')	Primer name	Probe size (bp)	[1]T_m (C)	[2]PCR (T_a)
5'-1	AATGCGGCAGTTCAGGTCTCTG	L195'ORF1 F	438	59.7	55
	GCCTTTCTACCTCATCCTCACC	MKCORF1-2R		57.1	
5'-4	AATGCGGCAGTTCAGGTCTCTG	L195'ORF1 F	142	59.7	55
	CAGAACTCAAGGAGGAGGTGGA	TQORF1-191 R		58.1	
3'-1	AGCGGAGTCCCCCAGGGGTC	TQCR3519 F	848	66.4	55
	CCCAGGGAAGTAGTTGAGTCGCC	TQCR4319 R		61.4	

Note. The first two probes were designed from primers in the region upstream of ORF1 on the 5' end. The last probe was designed from a region in ORF2 on the 3' end.
[1]Tm (°C) indicates melting temperature of primer.
[2]PCR (T_a) indicates annealing temperature and was lowered to yield PCR products.

agarose and poured onto pre-warmed 10cm NZY agar plate. Plates were incubated for 6 hours at 37°C and the plaque forming units (pfu) on each plate were counted. The total calculated recombinant plaques were determined to be 9.4 X 10^5 pfu/µg. An amplification of the primary library was completed according to the manufacturer's protocol and the resultant 35mL library was stored in 1mL aliquots in 7% DMSO at -80°C.

Primary screening. A post-amplification titre was performed using the same procedure as described above to determine the concentrations needed for plating in the primary screening and was determined to be at 6.7 x 10^6 pfu/µl. For the primary screening, plating densities of 3X 10^4 pfu/plate were used rather than the recommended 5 X 10^4 pfu/plate. A 3 µl aliquot of the amplified library (@6.7 X 10^6 pfu/µl) was added to 997 µl of SM buffer. 6 µl of this dilution was added to 600 µl of XL1-Blue MRA (P2) prepared host cells to yield the desired 3 x 10^4 PFU/plate. These reactions were incubated in sterile culture tubes for 15 minutes at 37°C. Following incubation, each reaction was added to 6.5mL of molten 0.7% NZY agarose and poured onto pre-warmed 15 cm NZY agar plate and cooled. Plates were inverted, incubated for 6 hours at 37°C, and stored at -4°C to cool. To increase the probability that rare clones were isolated; the primary screening of the CBG genome was plated to represent the genome 5 times. It was determined that using 10.4 plates would be sufficient, so 12 master plates were used.

Plaque Lifts. Plaque lifts from the primary screening were done using 136 mm diameter nylon membranes (Roche) and according to DIG High Prime DNA Labeling and Detection kit by Roche Applied Science. An additional two rounds of plating were

necessary to further purify and isolate the candidate clones and is described below. Subsequent rounds of plating required smaller 10cm NZY plates and smaller 82 mm diameter nylon membranes (Roche). DNA was fixed to filters through UV-crosslinking with a UVC-508 ultraviolet crosslinker using the automatic preset at 120,000 microjoules.

Hybridization and post-stringency washes. Hybridization and post stringency washes were performed using a ProBlot 6 Labnet Hybridization oven with six sealable Robbins tubes. Each Robbins tube was able to comfortably fit two larger (136 cm^2) nylon membranes for the primary screen and three smaller (82cm^2) nylon membranes in subsequent screenings. Initially, DIG-Easy Hyb granules were prepared according to the protocol. There were adjustments in volumes of the DIG-Easy Hyb solution for the pre-rinse and hybridization steps. 13.6mL of DIG-Easy Hyb solution was added to each Rollins tube containing two membranes (6.8mL/membrane) and pre-rinsed at the hybridization temperatures (Table 2.2). Following the pre-rinse, 9.5 mL of pre-warmed DIG-Easy Hyb solution (4.76mL/membrane) containing 25 ng/ml of denatured probe (237.5 ng total) were incubated at its hybridization temperature for nine hours. During the 5' screening, half of the tubes were screened with the 5'-1 probe and the other half were screened with the 5'-4. To reduce the impurities and non-specific DNA:DNA binding to the membranes, two post stringency washes were completed after hybridization. The first wash was performed twice using 30 mL of 2X SSC, 0.1% SDS at 25°C for 5 minutes and the second wash was performed twice using 30 mL of 0.5X SSC, 0.1% SDS at 67°C for

Table 2.2

Hybridization Temperatures of DIG-labeled Probes

Probe	**[1]T_m (◦C)**	**[2]T_{opt} (◦C)**	**Hyb. Temp (◦C)**
5'-1	73.9	48.9-53.9	[3]48
5'-4	76.18	51.18-56.18	[3]48
3'-1	63.04	38.04-43.04	44

Note. Hybridization of the probes to the membranes is calculated based on GC content of the probe to target according to the equation outline in the DIG High Prime DNA Labeling and Detection Starter Kit I.
[1]$T_m = 49.82 + 0.41(\%\ G+C) - (600/l)$ (l = length of hybrid in base pairs)
[2]$T_{opt} = T_m - 20$ to 25°C
[3]Hyb. Temperatures were lowered to yield successful hybridization.

15 minutes. Both the primary and subsequent screening followed the same volumes as listed above.

Immunological detection. The working solutions are part of the kit and the preparation of additional reagents required for immunological detection was completed according to the manufacturer's directions. Agitations of the all membranes were done on a Hoefer Red Rotary shaker, set at 4. Following all the necessary post-washes described in the protocol, hybridized DIG-labeled probes are detected with NBT/BCIP color substrate solution. A volume adjustment from the protocol was made and only 140 µl of color substrate solution was added to 7mL of detection buffer. The membranes were removed from light, and color development was monitored periodically until completion.

Plaque isolation. Following color development of the membranes, the positive clones were identified and removed from their respective NZY agar plate. Each clone was plucked using the wide-end of a sterile Pasteur pipette and stored in 1mL of sterile SM buffer with 50 µl of chloroform.

Plaque Purification

In order to isolate individual plaques of the clone, several additional rounds of screening were required. Following plaque isolation and storage in SM buffer, dilutions from each candidate were made by plating densities of approximately 400-700 pfu per 100 mm^2 plate achieved by adding 10 µl of the plaque filtrate to 990 µl of SM buffer. 2 µl of the dilution were added to 600 µl of prepared host cells. Reactions were incubated and plated using the same protocol previously mentioned in the primary screening. Plates were inverted, incubated for 8 hours at 37ºC and then put into -4ºC to cool. Plaque lifts,

probe hybridization, post hybridization/detection and plaque isolation were done according to the methods previously described.

DNA purification from Bacteriophage Plaques

The DNA from a total of 8 clones was purified from Bacteriophage plaques using a protocol as described in from St. John (1998).

Isolation of Full Length CR1

Initially, PCR using internal 10 µM primers flanking ORF1 and ORF2 were used in an attempt to confirm the presence of the complete CR1 sequence from two clones. Again, these primers were obtained from KC (2008) and TQ and ordered from Sigma-Aldrich. Following this, multiple primer pairs were used to isolate fragments of the CR1 element. 50 µl reaction mixtures were performed with the addition of a 5X Colorless Go Taq Buffer (Promega), 25 mM $MgCl_2$ (Promega), 10 µM forward/reverse oligonucleotide, 1.25U Go Taq DNA Polymerase (500U, 5U/µl) and H_2O to 50 µl. The thermal profile for each of these PCR reactions included a 94°C preheat for 1 min., 94°C denaturation for 1 min., annealing temperature T_a for 1 min., 72°C extension for 2 min., and a 72°C post heat for 10 min. for 35 cycles (Table 2.3A/B). The results of the PCR were visualized on a 1% SB agarose gel stained with ethidium bromide. PCR products were prepared for sequencing by the addition of 5U exonuclease I (10 U/µl, USB) and 0.5U shrimp alkaline phosphatase (1 U/µl, USB) followed by two 45 minute incubations, 37°C and 80°C, respectively. The samples were sent off-site to the North American Eurofins MWG Operon sequencing company meeting their general recommendations for concentrations of template and primer (http://www.operon.com/default.aspx).

Table 2.3A

Primer pairs used to isolate and sequence 1H clone in CBG.

Primer Pair	Primer Sequence (5'-3')	[2]T_m (°C)	Annealing Temperature (ºC)
L19 5'	[1]GAGCAACCAGGGTGGGCAAC	61.6	55
FLANKF	GCCTTTCTACCTCATCCTCACC	57.1	
MKC ORF1-2R			
L195'ORF1F	AATGCGGCAGTTCAGGTCTCTG	59.7	55
ORF ZN R1	GTACAGAGAGGACAAGGAAAGC	55.0	
CR1LK360F	GCCCAATATGCCGACCTGAC	58.5	55
MKC 1690R	CTGTGCTGCCTGTCCCTTCT	59.7	
5' ORF2 F1	ACACTAATGCACGTAGCATGG	55.3	55
LK 2275R2L	ATGGTCGAGTTCTCTATTCTT	50.6	
ORF2 F.2	GACTTCAACTTCCCAGACATATC	53.0	55
LK2076R2L	AGGTGCTTAATGCCTTCTT	51.7	
INK 557F	AGAGAAGGTCTGGTGGGAGATG	58.0	55
LK 2076R REDO	AAGAGGAGGACTAAGGAGAATCTC	55.0	
LK 2275F.2	GAAGAAGACTGGCCTGGCTAAAC	57.8	55
LK 3198R	CAACCAGGCGATCAGGC	56.6	
LK 2076F.2	CCGAGGGTACTGAGAGAACTGGG	60.4	55
LK 3378R	ACTGGCTGCTCGTTGGCTT	60.1	
TQ CR 3162F	CCTGGGAAGATCATGGAGCAG	57.5	55
TQ CR 3601R	GTTTGCAGATGACACCAAGCT	56.0	
CR1 long C2F	ACGCTTCGTTGGGTTAGAAACTGGC	61.6	55
TQ CR 3601R	GTTTGCAGATGACACCAAGCT	56.0	
CR1 long C2F	ACGCTTCGTTGGGTTAGAAACTGGC	61.6	55
CR1 LK1R	TCTGGTAAGGCCGCACCTC	59.7	
TQ CR 3519F	AGCGGAGTCCCCCAGGGGTC	66.4	55
CR1 LK1R	TCTGGTAAGGCCGCACCTC	59.7	
ORF2 BF	GCACCCTCAGTAAGTTTGCAGA	57.4	55
CR1 LK1R	TCTGGTAAGGCCGCACCTC	59.7	
ORF2 BF	GCACCCTCAGTAAGTTTGCAGA	57.4	55
TQ CR 4319R	CCCAGGGAAGTAGTTGAGTCGCC	61.4	
ORF2 BF	GCACCCTCAGTAAGTTTGCAGA	57.4	55
I 3'UTR R.2	GTCCAACCACTGACCTAACACTA	56.1	

Note. [1]Primer was not used in sequencing.
[2]T_m (°C) indicates melting temperature of primer.

Table 2.3B

Primer pairs used to isolate and sequence 7C clone in CBG.

Primer Pair	Primer Sequence (5'-3')	[2]T_m (°C)	Annealing Temperature (ºC)
L19 5'	[1]GAGCAACCAGGGTGGGCAAC	61.6	55
FLANKF	GCCTTTCTACCTCATCCTCACC	57.1	
MKC ORF1-2R			
L19 5'UTR	CACTAGCTAGGCTGCAATGGTCT	58.8	50
ORF1 F	GCCTTTCTACCTCATCCTCACC	57.1	
MKC ORF1-2R			
L19 5' ORF1F	AATGCGGCAGTTCAGGTCTCTG	59.7	50
MKC ORF1-1R	CTATCCTCTTTTGATAGTCCAGG	52.2	
L195'ORF1F	AATGCGGCAGTTCAGGTCTCTG	59.7	50
ORF ZN R1	GTACAGAGAGGACAAGGAAAGC	55.0	
INK ORF1 F	CTAGCCCAGGAGCTGGCAGG	62.6	50
MKC 1690R	CTGTGCTGCCTGTCCCTTCT	59.7	
5' ORF2 F1	ACACTAATGCACGTAGCATGG	55.3	50
LK 2275R2L	ATGGTCGAGTTCTCTATTCTT	50.6	
ORF2 F.2	GACTTCAACTTCCCAGACATATC	53.0	55
LK 2076 R REDO	AAGAGGAGGACTAAGGAGAATCTC	55.0	
ORF2 F.2	GACTTCAACTTCCCAGACATATC	53.0	50
LK2076R2L	AGGTGCTTAATGCCTTCTT	51.7	
LK2076F.2	ATGGTCGAGTTCTCTATTCTT	50.6	50
LK 3378R	ACTGGCTGCTCGTTGGCTT	60.1	
TQ CR 3162F	CCTGGGAAGATCATGGAGCAG	57.5	55
TQ CR 3601R	GTTTGCAGATGACACCAAGCT	56.0	
CR1 long C2	ACGCTTCGTTGGGTTAGAAACTGGC	61.6	55
CR1 LK1R	TCTGGTAAGGCCGCACCTC	59.7	
ORF2 B	GCACCCTCAGTAAGTTTGCAGA	57.4	55
I 3'UTR R.2	GTCCAACCACTGACCTAACACTA	56.1	

Note. [1]Primer was not used in sequencing.
[2]Tm (°C) indicates melting temperature of primer.

Data Analysis

Sequences from Eurofins MWG Operon Sequencing Company were initially analyzed using the Applied Biosystems Sequence Scanner™ v1.0 program. The selected sequences were imported and aligned in the GeneCodes Corporation Sequencher v4.5 software. Searches were made for sequences homologous to CR1 in GenBank using the BLAST databases. Specifically, the BLASTN (nucleotide), BLASTX (protein) databases were used. Regions that were found to be homologous to any known CR1s were further examined using Mega 5.1 using the constructed consensus sequences (1H and 7C), the KC (2008) CR1 consensus sequence in *Coscoroba coscoroba,* and the Wicker *et al.* (2004) CR1 consensus sequence in *Gallus gallus.* Programs, such as the NCBI BLASTP database and the ExPASy Prosite tools were used to do further analysis on the potential protein domains found within the consensus CR1 sequences. Further protein identification and analysis tools on the ExPASy server that were used included the ProtParam and AACompldent (constellation 0). A parsimony analysis using Mega 5.1 was done to help determine the subfamily in which the consensus sequences belonged.

Chapter Three: Results

Isolation of CR1 Clone

The genomic DNA extraction yielded a population of intact, high molecular mass genomic DNA to use in the building of a Cape Barren Goose (CBG) genomic library (Figure 3.1). In order to obtain discrete, small regions of genomic DNA suitable for recombination into the Lambda DASH vector, the partial restriction digest was analyzed by agarose gel electrophoresis (Figure 3.2A). Fragments between 15-23 kb were fractionated and recovered from a sucrose gradient (Figure 3.3B). Prior to ligation, fractions were purified of sucrose, pooled and analyzed via gel electrophoresis.

According to the Agilent protocol regarding ligation and packaging, 0.4 μg of fragmented DNA, in sizes of 20 kb ligated into 1ug of vector is expected to yield a library with 1 X 10^6-1.5 X 10^7 recombinant plaques. For this study, 0.16 μg of DNA with average sizes of 16 kb was ligated into 0.5 μg of vector. Making the assumption that this would be approximately half as efficient, one might expect a yield of 0.5 X10^5 - 7.5 X 10^5 recombinant plaques. The library titre obtained was 9.4 x 10^5 pfu/μg in a total volume of 1mL (1.5 x 10^5 recombinant plaques). The result was a representation just over 2 genomic equivalents of the CBG genome, assuming its genome to be similar to other birds, 1 X 10^9bp (Wicker *et al.*, 2004). An additional titre was performed following the amplification of the library to ensure its stability and to also determine the concentration

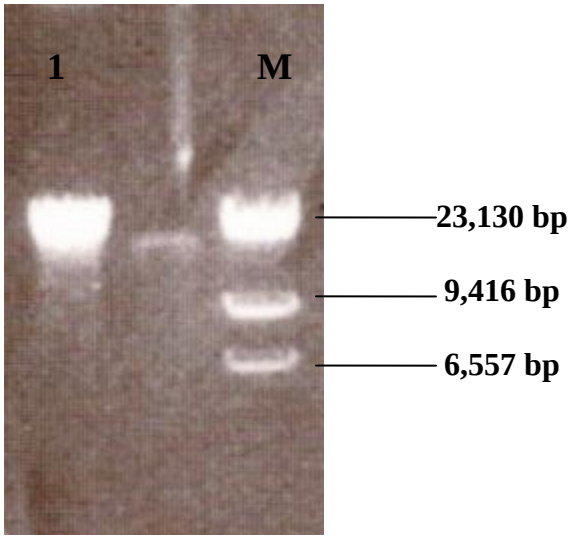
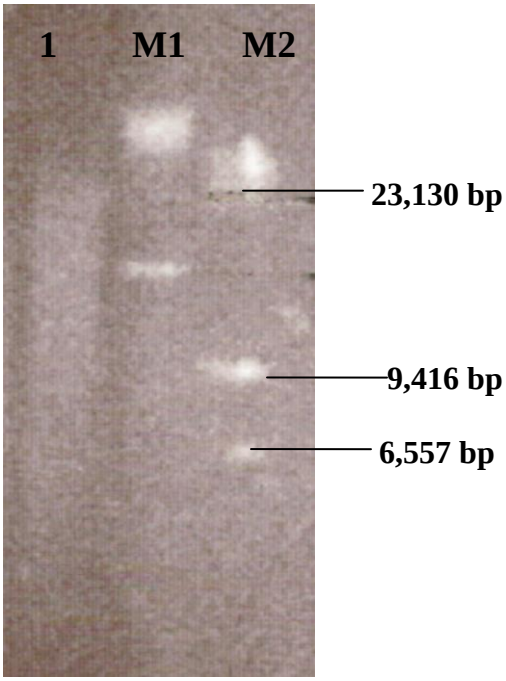


Figure 3.1. Extracted CBG DNA (lane 1). DNA was visualized on an 0.8% TAE gel stained with ethidium bromide using 5µl of *Hind*III cut lambda as the marker (lane M).

A)



B)

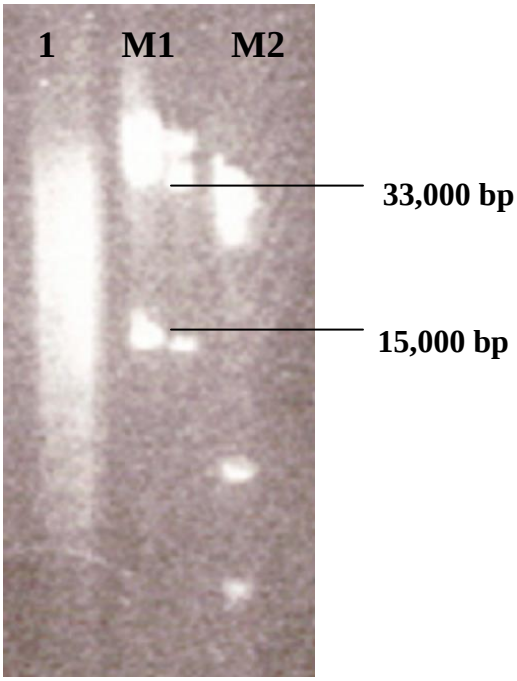


Figure 3.2 A. The partial restriction digests of CBG genomic DNA. Lane 1 contains genomic DNA digested with the *Sau*3 A I, 2.5 units. *B.* Lane 1 contains genomic DNA digested with the *Sau*3 A I, 1.25 units. DNA fragments from both were separated by gel electrophoresis and were visualized on a 0.8% TAE agarose gel stained with ethidium bromide. *Xho*I and *Hind*III cut bacteriophage lambda DNA were used as markers, M1 and M2 respectively.

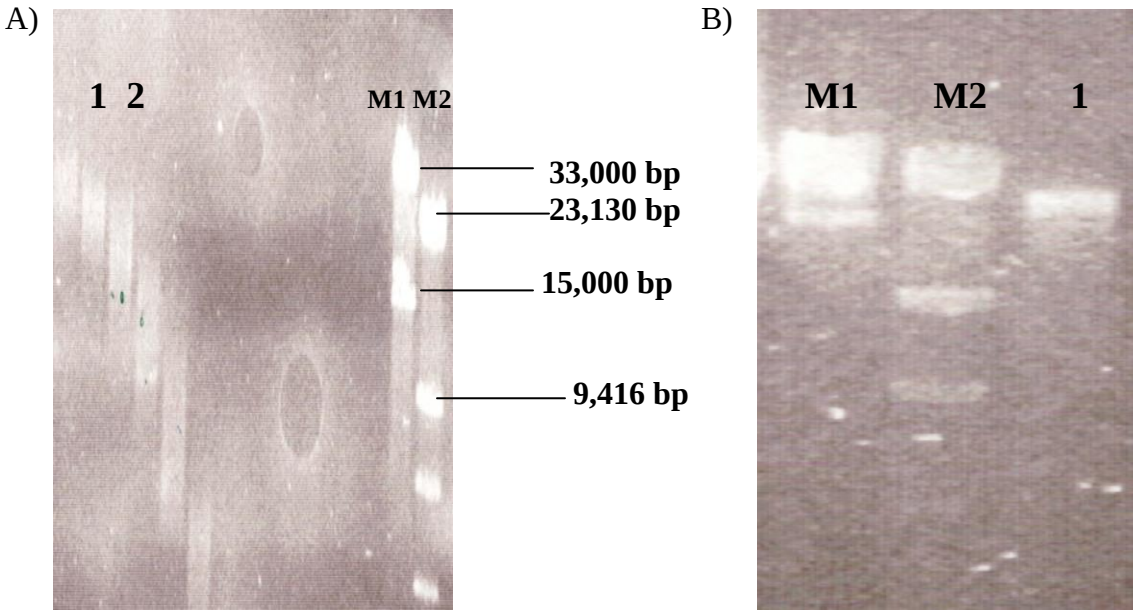


Figure 3.3 A. Partially digested genomic DNA following fractionation from a sucrose density gradient. Lane 1 and 2 contained the fractions that were of ideal size (15-23kb) for ligation into the bacteriophage Lambda-DASH vector. DNA fragments were separated by gel electrophoresis visualized on a 0.3% TAE agarose gel poured on top of a 1% agarose support stained with ethidium bromide. *Xho*I and *Hind*III cut bacteriophage lambda DNA were used as markers, M1 and M2 respectively. *B.* Samples from previous image were pooled and dialyzed and purified using the Thermo Scientific Slide-A-Lyzer MINI Dialysis device (10K MWCO). Sample was visualized on a 0.8% TAE agarose gel stained with ethidium bromide *Xho*I and *Hind*III cut bacteriophage lambda DNA were used as markers, M1 and M2 respectively.

needed for plating. The post amplification titre was 6.7 x 10^6 pfu/µl, 37 mL volume. Plating strategies and plaque lifts previously discussed were completed.

Probe Design and Hybridization

Based on the finding of only one intact element (CR1-F) in chicken and limited functional elements in other organisms, it was presumed that only a handful of intact (4-6) CR1 sequences exist in the entire CBG genome (Wicker *et al.*, 2004 and Kazazian, 1998). As previously discussed, work done on the CR1 locus in coscoroba provided several primers that allowed for the design of two probes on the 5' end and one probe on the 3' end of CR1 (Table 2.1). The use of probes on each end of the element of interest (in a 2-step hybridization process) allowed the identification of clones that contained both ends of the targeted element, and were therefore appropriate for further study. As 3' UTRs of CR1 are highly variable between subfamilies (St. John and Quinn, 2008a), probes were designed in the more conserved flanking regions of the ORFs (Figure 1.1). The efficiency of the labeled probes was determined after a 20 hour labeling reaction where radioactive markers were incorporated into 1 µg of experimental DNA. Quantification was performed through a direct detection method using a dot blot in which a series of dilutions of DIG-labeled DNA is applied to a small strip of nylon membrane. Immunological detection is complete on both the experimental DNA and a control DNA. If the 0.1pg/µl dilution dot of the control and experimental probes are visible, then the probe has reached labeling efficiency. Dot intensities were then compared. The observed efficiency of the experimental labeled probes ranged from 3 pg/µl to 0.3 pg/µl based on comparison to intensities of the known concentrations to the control (Figure 3.4). The intensity of the experimental and control spots were compared and determined to have an

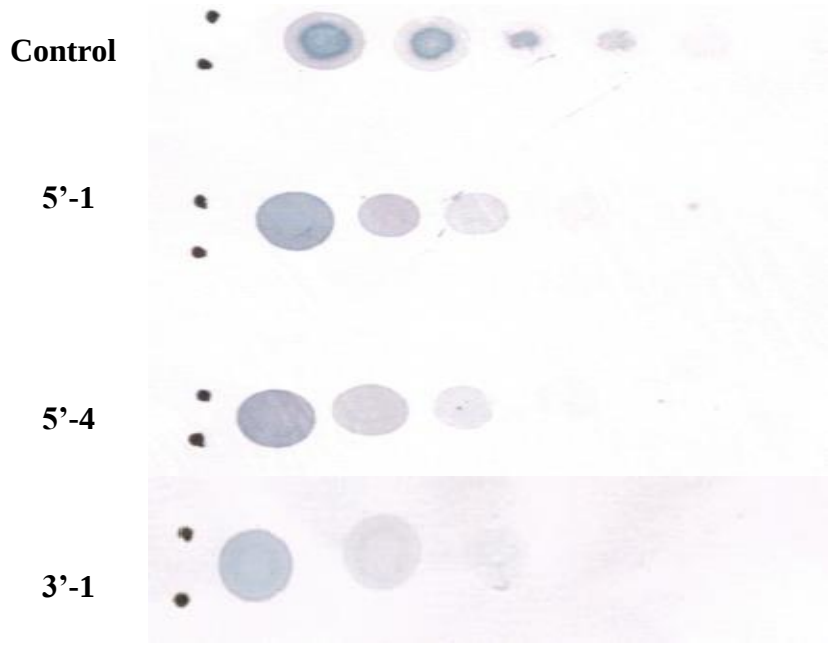


Figure 3.4. Dot blots of the labeled probes were compared against a control with known concentrations. The concentrations of the DIG-labeled control DNA are 10 pg/µl, 3 pg/µl, 1pg/µl, 0.3pg/µl, and the barely visible dot 0.1pg/µl (left to right). It was observed that the experimental dot 3 of the 5'-1 and 5'-4 probes roughly matched the 0.3pg/µl dot of the control (dot 4). The expected yield of the 3rd dot according to the control is 1pg/µl, so there was an approximate 3-fold difference in intensity.

approximate 3-fold difference in intensity with a final probe concentration of 33 ng/µl. As a step towards optimizing the screening process and techniques, two separate screens with each probe (5'/3') were complete. It was observed that the 3' screening had an approximate 30-fold difference in the number of positive signals compared to the 5' screen. Additionally, the final screen using the 5' probe yielded 24 candidates clones from a plating that included approximately 360,000 plaques (12 master plates). These observations would support the suggestion that a small amount of full-length CR1 elements exist in the genome. Hybridization, detection and extraction of candidate clones were completed as discussed in the materials and methods.

Full-length CR1 Clone Sequencing

A total of eight CR1 clones were recovered using the methods discussed in the previous chapter, and two were fully sequenced. Before further characterization of the clones began, DNA extracts of clones were retested to confirm the presence of the full-length element. PCR tests were performed using the same primer pairs that were used to design the 5'/3' probes. The amplification products for eight clones displayed single, distinct bands with the expected molecular weight. While one goal of this study was to sequence a full length CR1 element, another was to isolate an actively transposing element as opposed to one that had been inactivated by subsequent mutations. In order to eliminate elements that had mutated extensively, an area of ORF1 was sequenced and checked for premature stop codons. PCR was performed using the 5’1 probe primer pair internally located within ORF1 of CR1 (Table 2.1). The resulting sequence data were analyzed in BLASTX for both clones. The BLASTN (nucleotide) search tool yielded a significant match with ORF1 of chicken CR1 for one clone (1H), with an E value of 7e-

151 (Accession number AF308606). Clone 7C had a significant match to chicken CR1 with an E value of 6e-74. The ExPASy translation tool was then used to translate the nucleotide query. It was observed that reading (5'3') frame 1 in 1H and the (3'5') frame 1 in 7C contained no stop codons. Upon this confirmation, PCR reactions were completed using internal primers spanning the full length of the element (Table 2.3 A/B).

Analysis

Sequences from the PCR products displaying high confidence according to Phred quality scoring were used in the analysis. Scores of 30 or better (0-50) were used. They were assembled using Sequencher v4.1 (GenCodes) with a minimum match of 85% among sequences and a minimum overlap of twenty base pairs (Figure 3.5 and 3.6). The resultant consensus sequence from the multiple sequences spanning CR1 yielded 4417 bp of sequence for clone 1H and a 4067 bp of sequence for clone 7C. These were aligned with several other CR1 sequences by ClustalW in Mega 5.1 (Figure 3.7 and 3.8, *also* see appendix A). The consensus sequences were also subjected to a NCBI BLASTN search to look for similarities with other known CR1 sequences. The BLASTN algorithm shows a match with chicken CR1 starting at position 147 in the 1H consensus sequence at the first codon, ATG. Likely, this is the start codon of the chicken CR1. Clone 7C also shows significant homology to chicken CR1, but with an $A^{156} \rightarrow G$ mutation in the start codon and a premature stop codon in ORF2. The BLASTX algorithm, yielded significant matches with turtle CR1 (*Acanthochelys spixii*) and Zebrafish (*Danio rerio*) in ORF1, with an E value of 2e-50 and 1e-09, respectively.

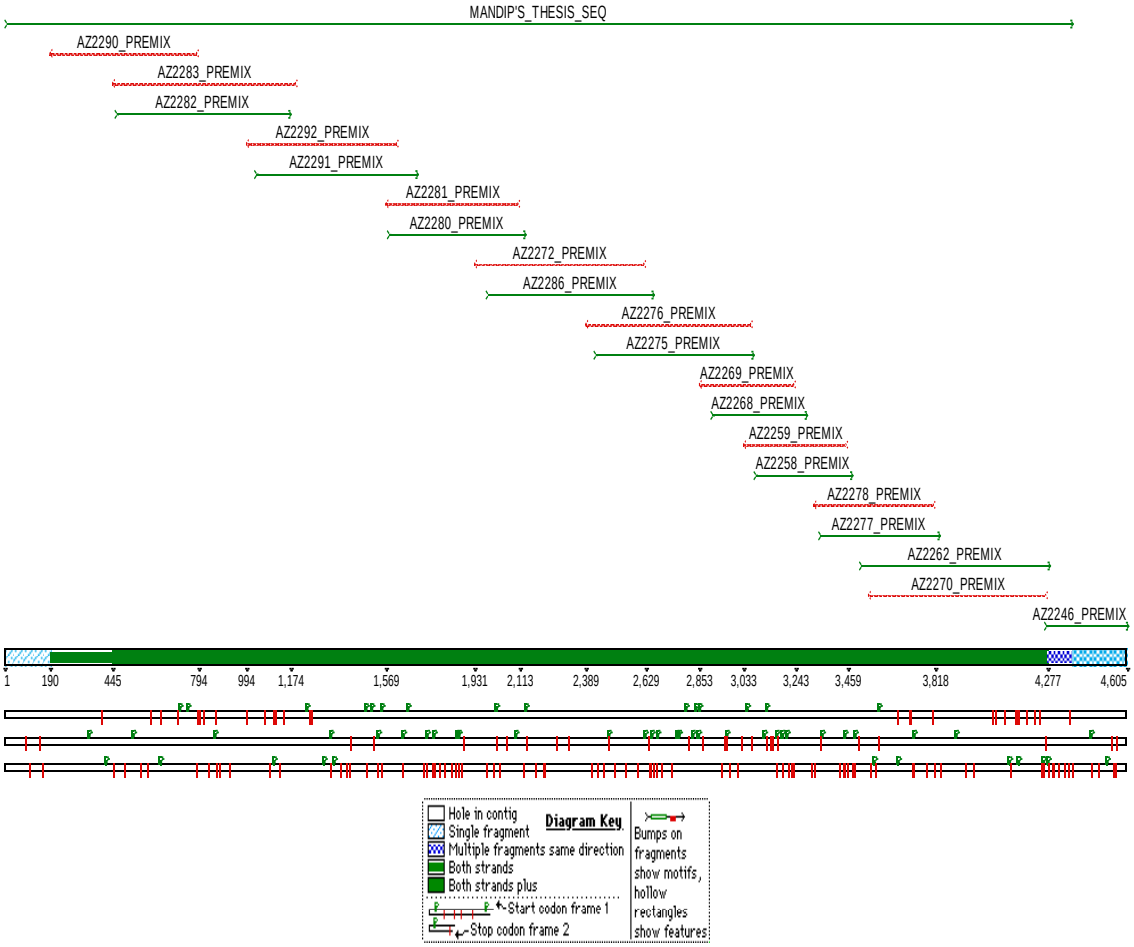


Figure 3.5. The overlay of all the sequenced fragments for 1H. The MKC_Thesis_Seq was obtained from MKC (2008). The overlaps have at least 85% matches. The consensus sequence is shown in figure 3.7.

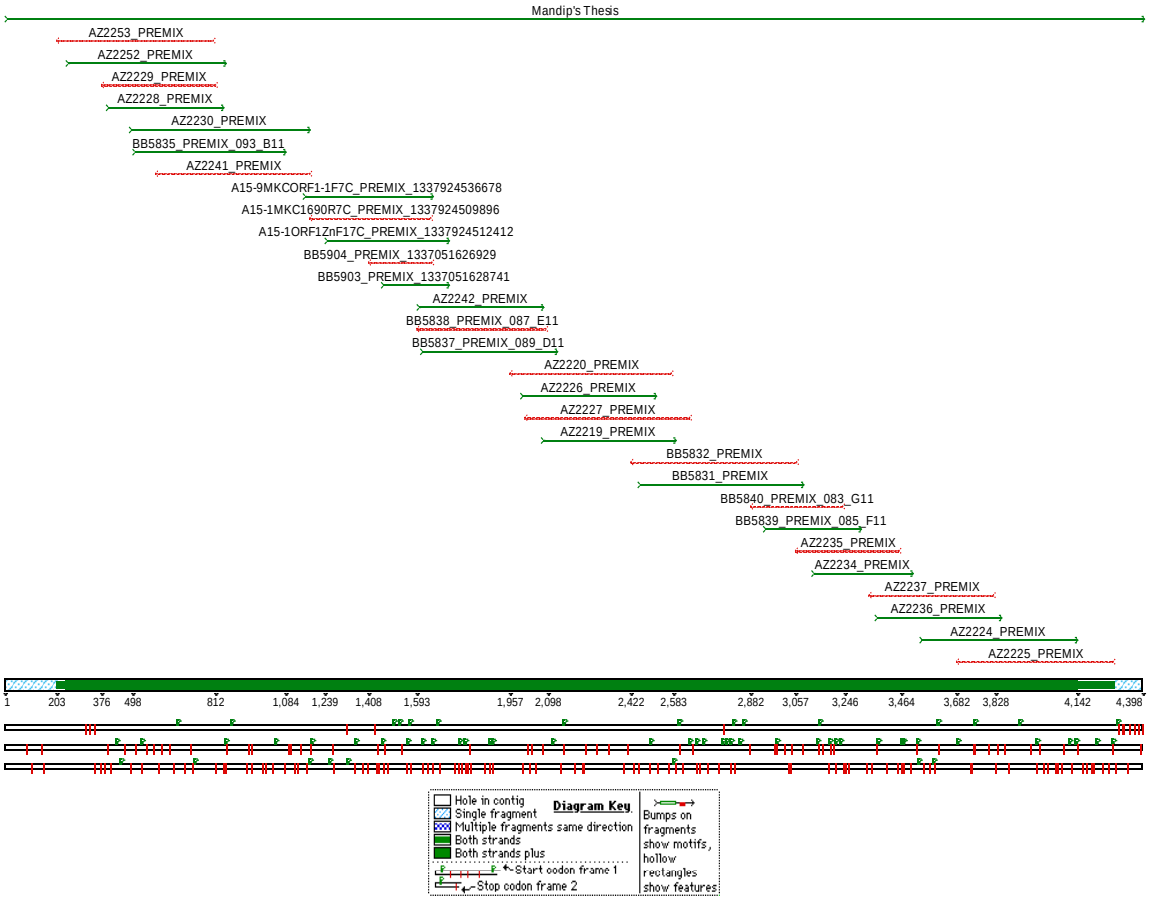


Figure 3.6. The overlay of all the sequenced fragments for 7C. The MKC_Thesis_Seq was obtained from MKC (2008). The overlaps have at least 85% matches. The consensus sequence is shown in figure 3.8.

1 TGAGCAACCC AGGGTGGGCA AACAGGGAGC GACGCGGCAG TTAGCGAGGC
51 AGTTAGCGCA GGCAGGGCGG GCAGGCAGGG CGGAGCGCTC ACCCCCGCCC
101 ATAGGGACAC CTCCTCTAAG GGCACTCTCC CGTCAGCTGG GCTACA**ATG**G
151 TCTCCAGCAG GCGCGGTGCT TTCTCTAGGA AGTCCGTACA CACCCAGACT
201 GACTGCCCGT CCAAACATGC GGCAGTTCAG GTCTCCGGAT GCGGGGAGTG
251 TCTGAGCCTC TTGCTGCCAT CGGCGGGAGG CAGGGGGACT GCTTGCGTGA
301 GGTGCGAGCA GGTGGACGAC CTGGTCCGCA TGGTGGCGGA ACTCCAGGAG
351 GAGGTGCAGA GGTTGAGGGA TATCAGGGAG TGCGAGCGGG AGATAGACTG
401 GTGGAGTGAC TCCCTGCAGG ACCGGAGGGA GAGGGACCAG GATGAGACGC
451 CCCAAAGGGG GGTGGACCCC CTGCCCTGTT GCCATCGGGC AGAGGGAGGG
501 GACCTGCGAG TTGAGGAGGA ATGGAGACAG GTCCCTTCTC GACCTCGCAG
551 GAGATGCTCT CCCCTTCCGG CGCTGCCTTC CCAGGTGCCC TTGAACAATA
601 GGTTTGAGGC CCTGGAGCTT GAGAGACTGG TGGGGGAGGA CGAGGTAGGA
651 AGCCTACCCA GGAGGATGCC TGAGGTGAGG AAGTTGACTC CACGCCTCAG
701 GACTGCCTCC ACCAAGAAAG AAAGAAGGGT GATTGTTGTG GGCGACTCCC
751 TCCTCAGGGG AACNGAGGGC CCTATTTGTC GGCCTGCCCC CACGCATAGG
801 GAAGTCTGCT GCCTCCCTGG GGCCAGGGTC AGAGACGTTG CCAGGAAGCT
851 TCCAAACCTG GTTCGGCCCT CTGACTATTA CCCGCTTTTG ATAGTCCAGG
901 CTGGCAGTGA TGATCTTGAA AAGAGAAGCC TGAAGGCTAT CAAACAGGAC
951 TATAGGGGGC TGGGACGATT GGTGGAGGGA GCGGGAGTGC AGGTGGTGTT
1001 TTCGTCTATC CCTACGGGGG AAGGCAGAGG CACGGAGAGG ACATGGAAAG
1051 CTCACGTGGT TAATAGGTGG CTCAGAGGCT GGTGCCAGCA CAGAAATTTT
1101 GGGTTTTTTC ACCATGGGGA GTTTTACTCA GCACCTGGCC TGATGGCCCC
1151 AGATGGGTCC CTATCTCCAA GGGGAAAACG GATCCTAGGC CAGGAGCTGG
1201 CGGGGCTCAT AGAGAGAGCT TTAAAC**TAG**G CACGAAGGGG GACGGGGCTC
1251 AAACAAGGCT TGTTGGAGCC GTGCCGGGGG AAACA**ATG**GC TAGGCTGGGG
1301 AAGAAGGCGA TGGCCCAGCT GAAGTGCATC TACACTAATG CACGCAGCAT
1351 GGGTAACAAA CAGGAGGAGC TGGAAGCCAT CGTGCAGCAG GAAGGCTACG
1401 ACTTGGTTGC CATCACGGAG ACGTGGTGGG ACCGCTCTCA TGACTGGAGT
1451 GCTGCAATGC CTGGCTATCG GCTCTTCAGG AGGGACAGGC AGCACAGAAG
1501 GGGTGGTGGC GTGGCTCTCT ACATTAGAGA ATCTTTTGAT GTTGTGGAAC
1551 TCCAGGCTGG GAATGATAAG GTTGAATCCC TGTGGGTTAG GATCCGCGGG
1601 AAGGCCGGCA AGGCTAACGT CCTGGTCGGG GTCTGTTATA GACCGCCGAA
1651 CCAGGATGAG GAGACGGATG AGGAGTTTTA TAGGCAACTG ACAGAAGTTG
1701 CAAAATCGTC GGCGCTTGTC CTCGTGGGGG ACTTCAACTT CCCGGATATA
1751 TCCTGGAAGC ACAACACGGC ACAGAGAAAG CAGTCTAGGA GGTTTCTGGA
1801 GAGTGTGGGA GATAGCTTCC TGATGCAGCT GGTCAGTGAA CCTACCAGGG
1851 GTGGTGCCCC GCTAGACCTT CTCTTCACAA ACAGAGAAGG ACTGGTGGGA
1901 GATGTGGTGG TCGGAAACCA TCTTGGACAG AGTGACCACG AAATGGTAGA
1951 GTTCTCTATT GTTGGCGAGG CCAGGAAGGG GACCAGTAAA ACTGCTGTCT
2001 TGGACTTCCG GAGGGCTGAC TTTGAGCTGC TCAGGACGCT GGTTGGCCGA
2051 GTCCCTTGGG AGGCGGTTCT GAAGGGCAGA GGAGTCCGGG AAGGCTGGGC
2101 GCTCCTCAAG AAGGAAATCT TAACGGCACA GGAGCGGTCC GTCCCCACGT
2151 GCCCAAAGAC GAGCCGGCGT GGAAGAAGAC CGGCCTGGCT CAACAGAGAG
2201 TTGCGGCTTG TGCTTAGCAG GAAAAAGAGG TTTTATAATC TTTGGAAAAA

2251 AGGGCGGGCC ACTGAGGAGG ACTACAAGGA TGTAGCGAGG CTGTGCAGGG
2301 AGAAAATTAG AAAGGCCAAA GCTCATCTGG AGCTCAAGCT GGCTACTGCT
2351 GTTAAAGACA ACAAAAAATC CTTTTACAAA TATATCAACG CAAAACGGAG
2401 GACTAAGGAG AATCTCCATC CTTTACTGGA TGCGAGGGGA AACCTAGTCA
2451 CTAAGGATGA GGAAAAGGCT GAGGTGCTTA ATGCTGCCTT TGCCTCAGTC
2501 TTTAGCGGCA ATACCGGTTG TTCTCTGGAT ACCCAGTGCC CTGAGCTGGC
2551 GGAAGGGGAT GGGGAGCAGG ATGTGGCCCT CACTGTCCAC AAGGAAATGG
2601 TTGGCGACCT GCTAAGGAGC TTGGATGTGC GCAAGTCGAT GGGGCCGGAT
2651 GGGATGCACC CGAGGGTACT GAAAGAACTG GCAGAGGAGC TGGCCGAGCC
2701 GCTTTCCATC ATTTATCGGC AGTCCTGGCT ATCGGTGGAG GTCCCAGTTG
2751 ACTGGCGGCT AGCCAATGTG ACGCCCATCT ATAAGAAGGG CCGGAGGGCA
2801 GACCCGGGGA ACTATAGGCC TGTCAGTTTG ACCTCAGTGC CGGGAAAGCT
2851 CATGGAGCAG ATTATCTTGA GGGTCATCAC GCGGCACTTG CAGGGCAAGC
2901 AGGCGATCAG GCCCAGTCAG CATGGGTTTA TGAAAGGCAG GTCCTGCTTG
2951 ACCAACCTGA TCTCCTTCTA TGACAAAGTG ACGCGCTTGG TGGATGAGGG
3001 AAGGGCGGTG GATGTGGTCT ACCTTGACCT CAGTAAGGCT TTTGACACCG
3051 TTCCCCACAA CATTCTCCTC AAGAAACTGG CTGCTCGGGG CTTGGACTGG
3101 CGTACGCTTC GCTGGGTTAG AAACTGGCTG GATAGCCGGG CCCAGAGAGT
3150 TGTGGTGAAT GGAGTCAAGT CCAGTTGGAG GCTGGTCACA AGTGGTGTCC
3201 CCCAGGGCTC GGTACTGGGG CCGGTCCTCT TTAATATCTT TATTGATGAT
3251 CTGGACGAGG GGATTGAGTG CACCCTCAGT AAGTTTGCAG ATGACACCAA
3301 GCTAAGTGCG TGTGTCGATC TGCTCGAGGG CAGGAAGGCT CTGCAGGAGG
3351 ATCTGGATAG GCTGGAGCGA TGGGCTGAGG TCAACTGTAT GAAGTTCAAC
3401 AAGGCCAAGT GCCGGGTCCT GCACCTGGGG CGCAACAACC CCAAGCAGAG
3451 CTACAGGCTG GGAGATGAGT GGTTGGAAAG CTGCCTGGCC GAGAAGGACC
3501 TGGGAGTATT GGTTGATAGT CGGCTGAATA TGAGCCAGCA GTGTGCTCAG
3551 GTGGCCAAGA AGGCCAACAG CATCCTGGCC TGTATAAGAA GCAGTGTGGC
3601 CAGCAGGTCT AGAGAGGTGA TTGTCCCCCT GTACTCGGCT CTGGTGAGGC
3651 CGCACCTCGA GTACTGTGTT CAGTTTTGGG CCCCTCACTA CAGGAAGGAC
3701 ATGGACGTGC TCGAGCGAGT CCAGAGAAGG GCGACCAAGC TGGTGAGGGG
3751 TCTGGAGAAC AAGTCTTACG AGGAGCGGCT GAGGGAGCTG GGGTTGTTCA
3801 GCCTGGAGAA GAGGAGGCTC AGGGGCGACC TTATCGCTCT CTACAGTTAC
3851 CTTAAAGGAG GCTGTAGTGA GGTGGGGGTT GGTCTGTTCT CCCACGTGCC
3901 TGGTGACAGG ACGAGGGGGA ATGGGCTAAA GTTGCGACAG GGGAGGTTTA
3951 GGTTTGATGT TAGGAAGTAC TTCTTTACCG AAAGGGTTAT TAAGCATTGG
4001 AACGGGCTGC CCAGGGAGGT GGNTGAGTCA CCATCCCTGG AGGTCTTTAA
4051 AAGACGTTTA GATGTAGAGC TTAGCGATAT GGTT**TAG**AGC TTAGCGATAT
4101 GGTTTAGTGG AGTACTTAGT GTTAGGTCGG AGGTTGGACT CGATGATCTT
4151 GAGGTCTCTT CCAACCTAGA AATCTGTGTC TGTGTCTGTG TCTGTGACTG
4201 CGCAGGGCCC GCGTTGTCCC AGCAGCAGGG CAGGGTGCCA GATCTCGCTC
4251 TCCACAAAGC CCTGCGCAAT GCTGACAGCA TTTGCTGCTT CCCCCAGCGT
4301 GACCATTTCC CTGTTTCCTT TTGGGTTGGA TGTTGTGGTG CCAAACAAGT
4351 AACAAAACTG TTTAAAATAC CTAAGCTAAA CTTTTTTTTT TTTTTTTTAC
4401 TTTTTTTTTT TNAAAAC 4017

Figure 3.7. The consensus sequence of CR1(1H). The total length is 4417 bases. ORF1 begins at position 148 and ends at position 1229 (1082 bases long). The ORF2 begins at position 1286 and ends at 4087 (2801 bases long). The start and stop positions for each ORF are shown in bold. The highlighted 8bp sequences are conserved regions in the 3' UTR, separated by 22bps. The accurate length of 5′ UTR is still unknown.

1 GGGTGGGCAA ACAGGGCGTG GTGAGTCAGA AGGGCGCGGC GCGGCAGTTT
51 GCTCGGCAGT TCACGCGTGC AGGCCACGCT CAGGCAGGGC AGAGTGTTCT
101 GCCCCGCCCA TAAGAACGAC TCATCTAACG GCTGTCTACC ACTAGCTAGG
151 CCACA**GTG**GC CTACACTAGG CAGAGAGCTC TCTCTAGAAA GTCTGTAACC
201 ACCCAGACTG ACTGCCCACT CAAAAATGCG GCAGTTCAGG TCTCTGGATG
251 CAGGGAGTGT CTGAGCCTGT TGCTGCCCTC TGAGGGAGGC AGAGATACTG
301 TGTGTGTGAG GTGCGAGCAG GCGGATGACC TGGTCTGCCT GGTGACAGAG
351 CTCAGGGAGG AGGTGGAGAG GTTGAGGGCC ATCAGGGAGT GTGAGCGGGA
401 GATAGACTGG TGGAGCAACT CCCTGCAAGG CCTGAAGGAG AGGTACCGGG
451 GTGAGACACC CCAAATGGGG GTGGACCCCC TGTCCTGTCG GGCTGAGGGA
501 GGGGACCTAG GAGTTGAGGA GGAATGGAGA CAGGTCCCTG CTCGACCTCG
551 CAGGCAACTA CCTGCCCCAC CTTGCCAGGT GCCCTTACAC AACAGATTTG
601 AGGCCCTGGA ACTCGAGAGA CCGGTGGGTG AGGATGAGGT AGAAAGTCTG
651 CCCAGGAGGA TGCCGAGGGC GAAGAAGTCG ACTCCACGCC TCAAGACTGC
701 CTCCACCAAG AAAGATAGAA GGGTAATCGT TGTAGGTGAT TCCCTTCTCA
751 GGGGAACAGA GGGCCCTATT TGTCAGCCTG ACCCTACCGG TAGGGAAGTC
801 TGCTGCCTCC CTGGGGATAG GGTCAGGGAT GTTGCCAGAG AGCTTCCCAA
851 TCTGGTTCAC CCATCTGATT ACTATCCTCT TTTGATAGTC CAGTCTGGCA
901 GTGATAATAT TGAAGAGAGA AGCCTGAAGG CTATCAAACT GGACTTTAGG
951 GGACTGGGAT GGTTAGTGGA TGGAGSRGGA GTACAGGTGG TGTTTTCGTC
1001 CATCCCTACA GTGGCAGGGA GGGGTACAGA CAGGACATGG AAAGCCCACC
1051 TGATTAACAC GTGGCTCAGA GGCTGGTGCC AACACAGGCA TTTTGGTTTT
1101 TTTTGACCAT GGGGCGCTTT ACTCGGCACC CGGCCTGATG GCCGCAGACG
1151 GGTCCCACCG ATCTCTAGGG GGAAAACGGA TCCTAGCCCA GGAGCTGGCA
1201 GGGCTTATTG AGAGGGCTTT AAAC**TAG**GTA AGAAGGGGGA TGGGGCTGAA
1251 ATAACGCTTG TTAGAGGTGT GCCAGGGGGA ACA**ATG**GCAA GGCTGGGGGA
1301 GAAGGCAATG GCCCAGCTGA AGTGCATCTA CACTAATGCA CTCAGCATGG
1351 GTAACAAGCA AGAGGAGCTG GAATCAATTG TGCAGCAGGC AGGCTATGAC
1401 TTGGTTGCCA TCACTGAAAC GTGGTGGGAC CACTCTCATG ACTGGAGTGC
1451 TGCAATGTCT GGCTATAGGC TCTTCAGAAG GGACAGGCAG CACGGAAGGG
1501 GTGGTGGTGT GGCTCTCTAT ATTAGAGAGT GTTTTGATGT TGTGGAACTT
1551 GAGGCTGGGA ATGATAAGGT TGAGTCCCTA AGGGTTAGGA TCAGCAGGAA
1601 GGCCAACAAG GCAAGCATCC TGGTGGGGGT CTGTTATAGA CCGCCAAACC
1651 AGGATGAGGA GACGGATGAG GAGTTCTACA GGCAGCTGGC AGAAGTTGCG
1701 AAATCATCAG CACTTGTTCT CGTGGGGGAC TTCAACTTCC CAGACATATC
1751 CTGGAAGCAC AAAACAGCCC AGAGAAAGCA GTCTAGGAGG TTTCTGGAGA
1801 GCATAGAAGA TAGCTTCCTG ACGCAGCTGG TTAGAGAGCC TACCAGGGGA
1851 GTTGTCCTGC TAGACCTTCT GTTCACAAAC AGAGAAGGAC TGGTGGGAGA
1901 TGTGGTGGTC GGGAGCTGTC TTGGGCAGAG TGACCACAAA ATGGTAGAGT
1951 TCTCTATTCT TGGCAAAGTC AGGAAGGGGA CCAGTAAAAC CGCTGTCTTG
2001 GACTTCCGGA GGGCTGACTT TGAGCTGTTA AGGACACTGG TTGGCAGAGT
2051 CCCTTGGGAG GTGGTTCTGA AGGGCAGAGG AGTCCAGGAA GGCTGGGCAC
2101 TCTTCAAGAA GGAAATCTTA GTGGCTCAGG AGCGGTTTGT CCCCACGTGC
2151 CCAAAGACGA GCCAGCGTGG AAGTAGACCG GCCTGGCTGA ACAGANNGTT
2201 GTGGCTCGAG CTTAGGAGAA AAAAGAGGGT TTATAATCTT TGGAAAAGAG

```
2251 GGCGGGCCAC TCAGGAGGAC TATAAGGATG TTGCGAGGCT TTGCAGGGAC
2301 AAAATTAGAA AAGCCAAAGC TCATCTGGAG CTCAATCTGG CTACTGCCAT
2351 TAAAGATAAC AAAAANTGCT TTTATAAATA CATCAACATG AAAAGGAGGA
2401 CTAAGGAGAA TCTCCATCCT TTACTGGATG CGGGGGGAAG CTTAGTTACA
2451 AGAGATGAGG AAAAGGCTGA GGTGCTCAAT GCCTTCTTTG CCTCAGTCTT
2501 TAGCGGCAAG ACCAGTTGTT CTCTGGATAC CCAGTCCCCT GAGCTGGTGG
2551 AAGGGGATGG GTAGAAGAAT GTGGCCCTCA CAATCCATGA GGAAATGGTT
2601 GGTAACCTGC TACAGCACTT GGATGTACAC AAGTCGATGG GTCTGGACGG
2651 GATCCACCTG AGGGTGCTGA NNGAACTGGC AGAGGAGCTG GGCAAGCTGC
2701 TTTCCATCAT TTATCGGCAG TCCTGGCTAT CAGGGGAGGT CCCAGTCAAC
2751 TGGCAGCTAG CAAATGTGAC ACCCATCTAC AAGAAGGGCT GGAGGGTAGA
2801 CCCGGGGAAC TATAGGCCTG TTAGTTTGAC CTCAGTGCCA GGNAAGCTCT
2851 TGGAGCAGAT TATCTTGAGT GTCATCACGT GGCACTTGCA GGGCAACCAG
2901 GCGATCAGGC CCAGTCAGCA TGGGTTTATG AAAGGCAGGT CCTGCTTGAC
2951 GAACCTGGTC TCCTTCTATG ACAAAGTGAC GCGCTTAGTG GATGAGGGAA
3001 AGGCTGTGGA TGTGGTCTAC CTTGACTTCG GTAAGGCTTT TGACACCATT
3051 TCCCACAACA TTCTCCTCAA GAAACTGGCT GCTAGTGACT TGGACTGGCG
3101 TACGCTTCAT TGGGTTAAAA ACTGGCTGGA TAGCCGGACC CAAAGAGTTG
3151 TGGTGAATGG AGTCAAATCC AGTTGGAGGC TGGTCACTAG TGGAGTCCCT
3201 CAGGGCTCAG TGCTGGGGCC AGTCCTCTTT AATATCTTTA TCGATGATCT
3251 GGATGAAGGG ATCGAGTGCA CCCTCAGTAA GTTTGCAGAC GACACCAAGT
3301 TAGATGCATG TGTCGATCTG CTCGAGGGTA GGAAGGCTCT GCAGGAGGAT
3351 CTGGATAGGC TGGACCGATG GGCTGAGGCC AACTGTATGA AGTTCAACAA
3401 GGCCAAGTGC CGGGTCCTGC ACCTGGGGCA CAACAACCCC AAACATCACT
3451 ACAGACTGGG AGATGGGTGG TTGGAAATCT GCCTGGCAGA GAAGGACCTG
3501 GGAGTATTGG TTGATAGTTG GCTGAATATG AGCCAGCAGT GTGCTCAGGT
3551 GGCCAAGAAG GCCAACAGCA TCCTGGCTTG TCTAAGAAGC AGTGTGGCCA
3601 GCAGGGCTAG GGAAGTGATT GTCCCGCTGT ACTCGGCTCT GGTGAGGCCG
3651 CACCTTGAGT ACTGTGTTCA GTTTTGGGCC CCTTGCTTCA AGAAGGACAT
3701 GGAGGTGCTC GAGAGAGTCC AGAGAAGGGC GACGAAGCTG GTGAGGGGTC
3751 TGGAGAACAA GTCTTATGAG GAGCGGCTGA GGGAGCTGGG GTTGTTTAGC
3801 CTGGAGAATA GGAGGCTCAG GGGCGACCTT ATCGCTCTCT ATAGGTACAT
3851 TAAAGGAGGC TGTAGCGAGG TAGGAATTGG TCTATTCTCC CATGTGCCTG
3901 CTGACAGGAC GAGAGGGAAT GGGTTAAAGT TGCGCCAGGG GAGGTTTAGG
3951 TTGGATATTA GGAAGAACTT CTTTACTGAA AGAGTTGTTA GGCATTGGAA
4001 TGGGCTGACC AGGGAAGTGG TTGAGTCGCC ATCCCTGGAG GTCTTTAAAA
4051 GACGTTTAAA TGTA***GCA*** 4067
```

Figure 3.8. The consensus sequence of CR1 (7C). The total length is 4067 bases. ORF1 begins at position 156 and ends at position 1227 (1071 bases long). The ORF2 begins at position 1284 and ends at 4067 (2783 bases long). The remaining bases in ORF2 and the 3'UTR are not available. The start and stop positions for each ORF are shown in bold. The accurate length of 5′ UTR is still unknown.

Similarly, there were also significant matches with ORF2 in turtle (6e-167), chicken (0), Zebra finch (*Taeniopygia guttata),* 0 and purple sea urchin (*Strongylocentrotus purpuratus),* 1e-150. The proteins encoded in ORF2 have been previously characterized and are known to harbor two highly conserved endonuclease and reverse transcriptase domains (Feng *et al.,* 1996), which the BLASTX confirmed.

The ORF1 and ORF2 nucleotide sequences were then added to NCBI Open Reading Frame finder, using the standard genetic code to identify all open reading frames. The ORF1 for clone 1H was translated in the +1 frame starting at ATG and yielded a 1083bp sequence (361 amino acids). The ORF2 was translated in the +1 frame starting at ATG and yielded a 2802 bp sequence (934 amino acids). When comparing 7C to the aligned clones, it was seen that ORF1 contained a substitution and ORF2 had a stop codon (appendix A).

For ORF regions, a protein-protein BLAST (BLASTP) search and a protein-domain search was completed using the ExPASy Prosite tool. ORF1 BLASTP results for 1H showed a match to a nucleic acid binding protein (Figure 3.8) in the Bastard halibut (*Paralichtys olivaceus*) with an E value of 1e-10. The ExPASy Prosite amino acid search suggested a significant match to a bipartite nuclear localization signal profile (Figure 3.9A). The ExPASy Prosite nucleotide search had matched to a ferredoxin-type iron sulfur binding region, Von Willebrand factor type C (VWFC) domains, epidermal growth factor (EGF) -like domains, tubulin subunits and a cysteine-rich integrin beta domain (Figure 3.9B). Additionally, both the ExPASy ProtParam and AACompldent tools were used in junction to search the Swiss-Prot and TrEMBL protein databases for potential

```
gb|AAN15746.1|  putative nucleic acid binding protein [Paralichthys olivaceus]
Length=294

 Score = 69.7 bits (169),  Expect = 1e-10, Method: Compositional matrix adjust.
 Identities = 49/150 (33%), Positives = 72/150 (48%), Gaps = 11/150 (7%)

Query  187  ASTKKERRVIVVGDSLLRGXEGPICRPAPTHREVCCLPGARVRDVARKLPNLVRPSDYYP  246
            + T  E++ +V+GDS+LR  +  +  PA T   V C+PGAR  DV   L  L +    Y
Sbjct  125  SDTPAEQQTLVIGDSILRNVK--LATPATT---VKCIPGARAGDVESYLKLLAKGKRKYS  179

Query  247  LLIVQAGSDDLEKRSLKAIKQDYRGLGRLVEGAGVQVVFSS-IPTGEGRGT-ERTWKAHV  304
             +I+  GS+DL  R  +  K +   +    +     V+FS  +P      T  RT   H
Sbjct  180  KIIIHVGSNDLRLRQSEITKINIDSVCTYAKTMSDSVIFSGPLPNVSSDDTFSRTSSFH-  238

Query  305  VNRWLRGWCQHRNFGFFHHG-EFYSAPGLM  333
              RWL  WC   N GF ++   F+  PGL+
Sbjct  239  --RWLSRWCPDNNVGFVNNWRTFWGKPGLI  266
```

Figure 3.9. Results obtained from BLASTP algorithm of the ORF1 (1H) region showing a putative nucleic acid binding protein, with an E value of 1e-10.

A.

NLS_BP *Bipartite nuclear localization signal profile* :

USERSEQ1  (360 aa)

178 - 194: **score** = 4.000

RKLTPRLRTASTKKERR

B.

4FE4S_FER_1 *4Fe-4S ferredoxin-type iron-sulfur binding region signature* :

6 - 17: [level tag: (-1)] CtCCAgCAgGCG

VWFC_1 *VWFC domain signature* :

23 - 70: [level tag: (-1)]

CtttCTCtaggaagtccgta.CacacCcaga......Ctgactgcccgt....CCaaaC

394 - 436: [level tag: (-1)]

CgacCTCgcaggagatg....Ctct.Ccc.......Cttccggcgctg....CCttccC

EGF_1 *EGF-like domain signature 1* :

266 - 277: [level tag: (-1)] CcCtgCagGAcC

387 - 398: [level tag: (-1)] CcCttCtcGAcC

667 - 678: [level tag: (-1)] CtCccTggGGcC

885 - 896: [level tag: (-1)] CaCggAgaGGaC

1002 - 1013: [level tag: (-1)] CcCagAtgGGtC

TUBULIN *Tubulin subunits alpha, beta, and gamma signature* :

527 - 533: [level tag: (-1)] AGGTGAG

INTEGRIN_BETA *Integrins beta chain cysteine-rich domain signature* :

634 - 649: [level tag: (-1)] CgGcctGcCcCcaCgC

Figure 3.10. *A.* A significant match was seen within amino acids 178-194 as a bipartite nuclear localization signal profile for a motif within ORF1 of the CR1 (1H) returned by the ExPASy Prosite tool. *B.* Significant matches were seen with nucleotides 6-17 as a ferredoxin-type iron sulfur binding region, two Von Willebrand factor type C (VWFC) domains from bp 23-70 and 394-436, multiple epidermal growth factor (EGF) -like domains, tubulin subunits from bp 527-533 and an integrin beta chain, cysteine-rich domain from bp 634-649.

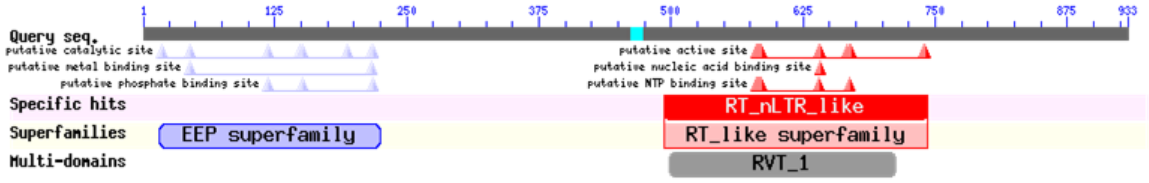


Figure 3.11. The conserved domains of ORF2 of CR1 (1H) are an endonuclease domain (EEP superfamily) and a reverse transcriptase domain (RT_like superfamily). The figure was obtained from NCBI using a BLASTP database.

hits. Interestingly, results showed similarities to GAG protein, which include nucleocapsid proteins. Although much more is known about ORF2, BLASTP results for 1H confirmed a statistically significant match to an endonuclease and reverse transcriptase (RT) domains with an E value of 8.10e-23 (Figure 3.10). Because 7C did not have ORF1 present, no information regarding potential proteins/domains was elucidated.

High levels of sequence divergence in the 3'UTR made it difficult to align. However, since there are highly conserved motifs just upstream of the 3' UTR in ORF2, it was determined that alignment using ORF2 would be best to give insight into which subfamily CR1 belongs. Parsimony analysis of the 1H clone was performed, in addition to 119 CR1 ORF2's isolated from coscoroba (I-VI) (Figure 3.11) (St. John *et al.*, 2004). Boot strap analysis of the 1H CR1 supported a monophyly within clade I, but only with a bootstrap support of 68. In review, the 1H CR1 sequence in this research is 4417 bases long, making it the longest CR1 isolated from waterfowl, so far. ORF1 begins at position 147 of the consensus and ends at position 1229 making ORF1 1082 base pairs long. ORF2 is 2801 long and begins at position 1286 and ends at 4087. Lastly, 330 base pairs of the 3'UTR are present. The 7C CR1 sequence is 4067 bases long. ORF1 begins at position 156 of the consensus sequence and ends at position 1227 making ORF1 1071 bases in length. The incomplete ORF2 is 2783 long and begins at position 1284 and ends at base 4067.

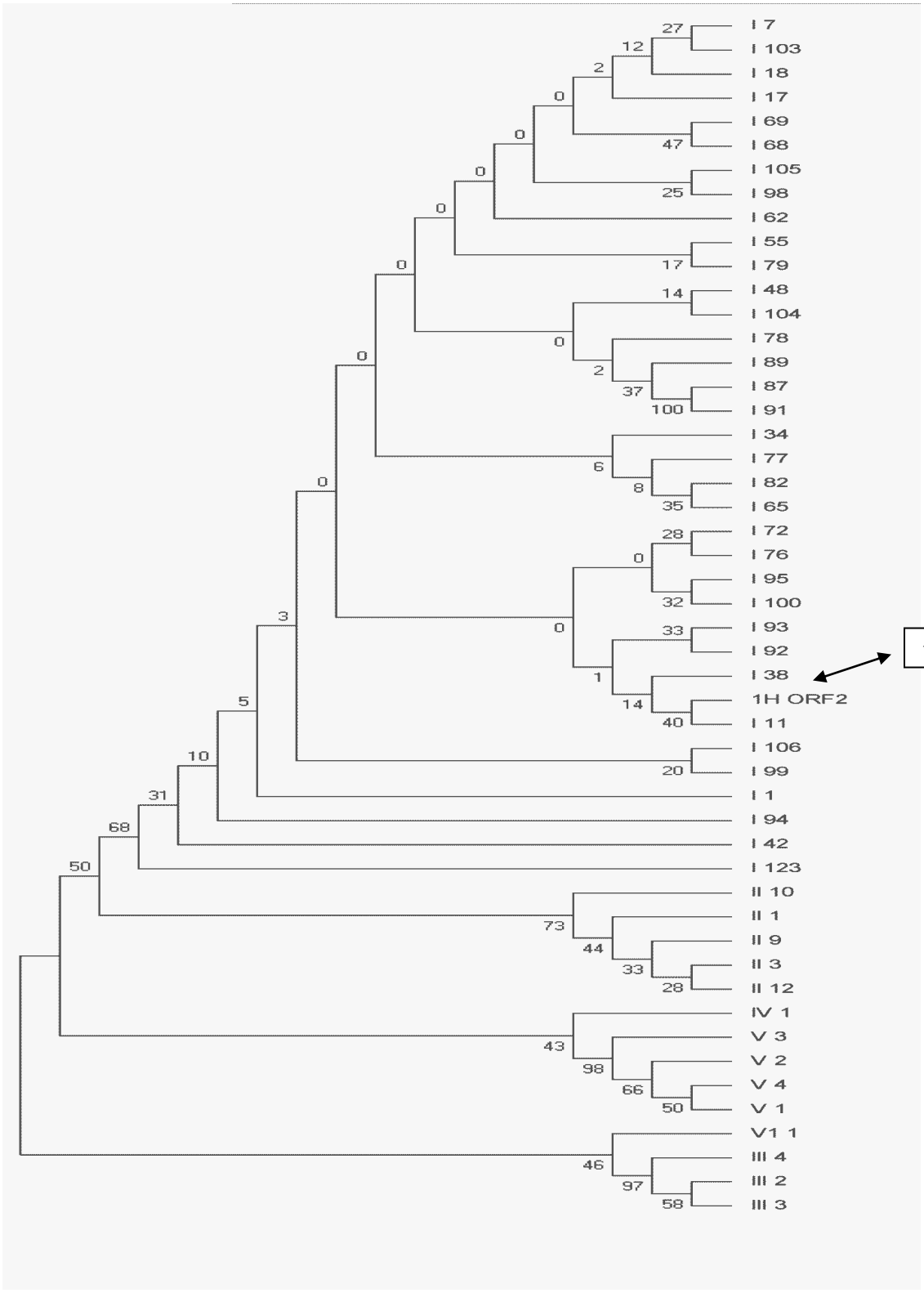

I 7
I 103
I 18
I 17
I 69
I 68
I 105
I 98
I 62
I 55
I 79
I 48
I 104
I 78
I 89
I 87
I 91
I 34
I 77
I 82
I 65
I 72
I 76
I 95
I 100
I 93
I 92
I 38
1H ORF2
I 11
I 106
I 99
I 1
I 94
I 42
I 123
II 10
II 1
II 9
II 3
II 12
IV 1
V 3
V 2
V 4
V 1
V1 1
III 4
III 2
III 3
1H

Figure 3.12. Parsimony analysis of the 1H clone and a dataset derived from consensus sequences from coscoroba subfamilies I-VI (St. John, 2008a). Labeled with Roman numerals to represent the subfamily following an Arabic numbers to represent the clone number in each subfamily. Clone 1H belongs to subfamily 1, with a bootstrap value of 68.

Chapter Four: Discussion

As a step toward sequencing an actively transposing full-length CR1 element in CBG, a library screen was performed to isolate full-length elements containing potentially functional ORFs. The size distribution of CR1s in the chicken genome found that >98% are truncated at the 5' end to 2000 bp or less (Wicker *et al.,* 2004). The two step hybridization process, initially using the 5'-end probe and finishing with the 3'-end probe, narrowed down the clones to a limited few potential elements present. In fact, there was a total yield of eight clones, which may be a reflection of the few targets present in the CBG genome. The isolation and characterization of two CR1 clones (1H and 7C) was completed here. Both contained adequate sequence similarity to bind to the probes, however internal sequence analysis varied significantly between the two. 7C had at least one frameshift or termination mutation, indicating that it is not retrotransposition-competent. However, 1H had two ORF's, indicating that it may be an active element.

For clone 1H, the full-length ORF1 was sequenced in addition to ORF2 and the 3'UTR. For the second clone, 7C, the full length ORF1 was sequenced in addition to most of ORF2. Conceivably, these sequences could provide more insight into the mechanisms of CR1 transposition.

Although inactive, the full length sequence of CR1 in several other organisms is available in GenBank. Wicker *et al.,* (2004) have reported 2 complete CR1 sequences in

chicken, only one with intact ORFs. The element with intact ORFs has the potential to produce proteins and was used as a comparison with the sequences in this research. This "mother" sequence of CR1 in chicken is referred to as CR1-F. The putative intact CR1-F element is 4033 bases in length and consists of two closely spaced ORFs out of which ORF1 is 1073 bases and ORF2 is 2946 bases. KC (2008) has reported a full-length CR1 sequence from a closely related waterfowl, *Coscoroba coscoroba*, which proved valuable in this study. The KC CR1 has a total length of 4378 bases, where ORF1 is 1083 base pairs long, ORF2 is 2841 bases long and 112 bases comprise the 3'UTR. Unlike the chicken CR1 and the sequences reported in this study, the KC sequence is undoubtedly a conglomeration of several different CR1s as it was produced from direct sequencing of multiple independent PCR amplifications of total genomic DNA. The sequences described here for the CBG are the first to be collected from single long elements within Anseriformes. Both the consensus sequence for chicken CR1-F and coscoroba CR1 allowed the identification of a number of conserved sequence motifs.

The end of ORF1 and the beginning of ORF2 of both CR1 clones is separated by a 56 base pair region. This region in both clones has a high content of purine bases (68% in 1H and 71% in 7C). Interestingly, there is currently a debate about the origins of TEs: whether non-LTR retrotransposons gave rise to retroviruses (Burke *et al.*, 2002) or if TEs were derived from endogenous retroviruses (Kazazian, 2004). A study by Peters *et al.*, (2008) suggests that an exogenous retrovirus, the lentivirus, contains several regions of poly purine tracks (PPT) within the genome that are important for viral function. Specifically, a mutagenesis experiment was performed in the various PPT regions of the

virus to further elucidate a function. It was observed that complex roles, ranging from the facilitation of RNA-pol interactions, production of *gag* proteins, and viral replication and transfer were all affected. Notably, similar PPTs are also seen in several LTR-retrotransposons, such as the Ty retroelement in yeast (Heyman *et al*., 2003). Perhaps, the debate of the relationship between non-LTR retrotransposons and retroviruses could be better understood by isolating active elements.

The first 146 bases of 1H and 155 bases of 7C in Figure 3.7 and Figure 3.8 are rather ambiguous and presumably include some or all of the 5' UTR. As a vast majority of CR1 elements in the avian genome are severely truncated, finding sufficient information to identify the boundaries of an intact 5' UTR is problematic. It has been previously noted that the 5'UTR serves as a promoter to facilitate the transcription of the element. In support of this, Minakami *et al.* (1992) observed that the 5' UTR in the human L1 non-LTR retrotransposon was able to turn on unrelated genes downstream. It was discovered that turtle CR1 contained sequence homology to a portion of the L1 promoter and short DNA sequences called E-box (CANNTG) and c-Myb (CAGTTA) that bind transcription factors to initiate transcription (Kajikawa *et al*., 1997 and Howe *et al.,* 1990). Interestingly, this study reports that the short portion of the 5' UTR in 1H contained two similar motifs that could be potential binding sites for transcription factors (not shown). Additionally, Haas *et al.* (2001) observed that the 5'UTR between chicken CR1 subfamilies shows no homology and predicts that each subfamily arose by acquiring different promoter sequences. As the case may be, a better conclusion of the exact function will be made using larger data sets.

Not much is known about the function of ORF1 of CR1. In 2008, KC reported a nucleic acid binding domain, which had been previously predicted to exist by Haas *et al.* (1997) simply by analogy with families of non-retrotransposons. In database searches, the current 1H CR1 was successful at revealing a significant match to a putative nucleic acid binding domain in studies related to the CR1-like family of non-LTR retrotransposons from the Bastard halibut fish (*Paralichthys olivaceus*). This domain is encoded by base pairs 559-1008 of ORF1 region in 1H (Figure 3.8). To support this find, additional analysis of the amino acid content in ORF1 (1H) resulted in close matches to proteins found in viral *gag* genes. Encoded in most *gag* genes are nucleoproteins that contain zinc finger-like motifs that structurally associate with nucleic acids (nucleic acid binding domain). An unusual arrangement of cysteine residues is reported in ORF1 of the turtle CR1 element, $CX_2CX_{14}CX_2C$. This motif has a similar amino acid composition to those found in transcription factors SL1 and TIF-IB (Kajikawa *et al.*, 1997, Comai *et al.*, 1994 and Heix *et al*., 1997). It is also reported here that the waterfowl ORF1 of CR1 contains this zinc finger motif from amino acid residues 32-53. Additionally, ORF1 has a very high probability of including a bipartite nuclear localization signaling profile, observed in 17 amino acid sequence coded by base pairs 178-194 (Figure 3.9A). The uptake of proteins into the nucleus is highly discerning and mediated by their selective entry through the pores of the nuclear envelope (Dingwall and Laskey, 1986). Perhaps, the transcribed RNA molecule, in association with all the necessary retrotranpositional proteins, is translocated back to the nucleus using these two motifs. More specifically, the nucleic acid binding domain could bind the RNA while using the localization signal to

translocate into the nucleus. Lastly, a signature for tubulin binding domain was found in ORF1. Tubulins are globular proteins that comprise the cytoskeletal network (microtubules) and facilitates traffic of materials within the cell. It would be interesting to think that tubulins further facilitate the movement of RNA into the nucleus.

ORF2 of CR1 and CR1-like elements is relatively well characterized and its contribution to the insertion mechanism of retrotransposons is understood. It is reported here to encode for a putative endonuclease phosphatase and reverse transcriptase domain (Figure 3.10). In review, these two domains work to catalyze the phosphodiester bond on the target site DNA and to reverse transcribe the RNA into the host genome as dsDNA (Luan and Eickbush, 1995).

Overall, the 3'UTR is known to be distinct between subfamilies of CR1. While there are conserved blocks of sequence within the 3' UTR, regions outside of these blocks are highly variable (St. John and Quinn, 2008a). One such example of a conserved region at the sequence level in the 3' UTR belongs to a set of inverted repeats. It is proposed that they form a stem-loop structure that is recognized as the *cis*-acting docking site for ORF1 proteins (Haas *et al.*, 2001). The CR1 reported here contains an 8 bp sequence AGGTTGGA and its counterpart TCCAACCT separated by a 22 bp region, as seen in the highlighted regions (4101-4151) in figure 3.7. It has been previously speculated that an alternative (or additional) function of the inverted repeats in the 3'UTR may some how be involved in the interfering RNA (RNAi) pathway by forming double stranded DNA (St. John and Quinn, 2008a). This pathway might serve to silence transposable elements in the genome by binding to a RISC complex. In turn, the bound transcript/RISC complex

will cleave other transposable element containing complementary sequences which prevents integration into the host genome (Slotkin and Martienssen, 2007). After subsequent parsimony analysis of 3'UTR region (along with the conserved 3' end of ORF2), it is suggested that the 1H CR1 in this study belongs to subfamily I (Figure 3.11). Notably, this subfamily is distinct within Anseriformes and believed to be the same subfamily that has been actively undergoing retrotransposition in evolutionarily recent times (St. John *et al.,* 2004). The flanking region of the 3'UTR is defined by the presence of an 8-bp direct repeat, 5'-TTCTGTGA-3' (St. John and Quinn, 2008b and Silva and Burch, 1989). An internal segment of this repeat, 5'-TCTGTG-3' is present in four copies in the 1H clone sequence. In a biochemical demonstration of the L1 endonucleolytic activity of ORF2, it was hypothesized that the 8-bp direct repeat sequence is detected and cleaved on the host strand. This becomes the site that this repeat unit within CR1 hybridizes and then undergoes reverse transcription (Feng *et al.*, 1996).

Chapter Five: Summary

In summary, a complete ORF1, ORF2, 3' UTR and likely 5' UTR was sequenced from Cape Barren goose. It was found that ORF1 encodes a domain that may be involved in binding to nucleic acids and another that may be a signal to translocate the complex into the nucleus. ORF2 was found to encode an endonuclease and reverse transcriptase domain that allows the incorporation of the RNA transcript into the host genome. It is quite possible that the clone containing two open reading frames could be an active element (or at least is close to an accurate representation of a waterfowl CR1) and provides further information regarding the exact transposition mechanism. In a way, the isolation, characterization and elucidation of a transport mechanism of an active CR1 element can contribute to the understanding of broader questions that surround the world of transposable elements. This includes how genomes have been influenced by mobile elements and a potential shift in the paradigm how genomes are constructed.

The technique used in this study resulted in the successful isolation of two full length CR1 elements within CBG, with 6 more clones waiting for further sequencing and analysis. The cloning and sequencing of CR1 from a single source DNA template serves to be much more accurate than the previous technique of "PCR walking" and overlapping smaller pieces from multiple CR1s in the genome in the sense that the former approach generates sequence from a single cohesive element. However, the sequence and primers

obtained from that earlier "MKC" research made it possible to accomplish the clone isolation and characterization much quicker than otherwise possible. There are however, a number of drawbacks to this approach that still need to be resolved. Due to the fact that the flanking regions the CR1 element is currently unknown, there was difficulty in obtaining reliable sequence of this area. Several attempts were made at running long PCRs using the known T3 and T7 promoters of the vector to resolve the flanking regions. There was little success, as only a small segment of the 3' UTR was resolved. Additionally, there are many steps involved in getting to the end goal of isolating a complete sequence. The repeated manipulation of the clone DNA may have introduced base errors (deletions and insertions) that would not otherwise occur. Perhaps, this could explain the novel 6 bp repeat present in the 3'UTR of the 1H clone.

Future research around CR1 involves resolving the other 6 clones obtained during this study from the vector. Once the 5' and 3'-ends are clarified, oligonucleotide primers can be designed from sequences on the flanking regions of the candidate clones. Once this is complete, these flanking primers can be used to reamplify perspective CR1 transposons in a genomic sample. Beyond this, active CR1 elements can be removed from the genome with modified primers that have rare-cutting restriction endonucleases at their 5' end. With an isolated CR1, cloning the element into a vector that would be inserted into a certain mouse cell line assay could test for the presence of retrotransposition activity (Martin *et al.*, 2005). This novel method could then allow CR1 to be used to better understand how TEs affect genomes.

Bibliography

Benjamin, H.W., and Kleckner, N. (1989) Intramolecular transposition by Tn 10. *Cell* 59:373-383.

Bestor, T.H. (2007) The host defense function of genomic methylation patterns, in Novartis Foundation Symposium 214 - Epigenetics (eds D. J. Chadwick and G. Cardew), John Wiley and Sons, Ltd., Chichester, UK.doi: 10.1002/9780470515501.ch11

Biemont, C., and Vieira, C. (2005) What transposable elements tell us about genome organization and evolution: the case of Drosophila. *Cytogenet. Gen. Res.* 110*:*22-34.

Boeke, J.D. (2003) The unusual phylogenetic distribution of retrotransposons: a hypothesis. *Gen. Res.* 13:1976-1983.

Boissinot, S., Emtezam, A., Young, L., Munson, P.J., and Ferano, A.V. (2004) The insertional history of an active family of 1 retrotransposons in humans. *Gen. Res.* 14*:*1221-1231.

Bowen, N.J., and Jordan, I.K. (2002) Transposable elements and the evolution of eukaryotic complexity. *Curr. Issues. Mol. Biol.* 4:65-76.

Burke, W.D., Malik, H.S., Rich, S.M., and Eickbush, T.H. (2002) Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, *Giardia lamblia*. *Mol. Biol. Evol*. 19*:*619-630.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003) The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acid Res.* 31:4385-4390.

Comai, L., Zomerdijk, J.C., Beckmann, H., Zhou, S., Admon, A., and Tjian, R. (1994) Reconstitution of transcription factor SL1: exclusive binding of TBP by SL1 or TFIID subunits. *Science* 266:1966-1972.

Deininger, P.L., Moran, J.V., Batzer, M.A., and Kazazian Jr., H. H. (2003) Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.* 13:651-658.

Dingwall, C., and Laskey, R.A. (1986) Protein import into the cell nucleus. *Annu. Rev. Cell Biol.* 2:367-390.

Donne-Goussé, C., Laudet, V., and Hanni, C. (2002) A molecular phylogeny of anseriformes based on mitochondrial DNA analysis. *Mol. Phylogenet. Evol.* 23:339-356.

Doolittle, W.F., and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601-603.

Fabrick, J.A., Mathew, L.G., Tabashnik, B.E., and Li, X. (2011) Insertion of an intact CR1 retrotransposons in a cadherin gene linked Bt resistance in the pink bollworm, *Pectinophora gossypiella*. *Insect Biochem. Mol. Biol*. 20:651-665.

Feng, Q., Moran, J.V., Kazazian Jr., H.H. and Boeke, J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87:905-916.

Feschotte, C. and Pritham, E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*. 41:331-368.

Finnegan, D.J. (1989) Eukaryotic transposable elements and genome evolution. *TIG* 5:103- 107.

Finnegan, D.J. (1992) Transposable elements. *Curr. Biol.* 2:861-867.

Garfinkel, D.J., Boeke, J.D., and Fink, G. R. (1985) Ty element transposition: reverse transcriptase and virus-like particles. *Cell* 42:507-517.

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell* 110:315-325.

Gombart, A.F., Saito T., and Koeffler, H.P. (2009) Exaptation of an ancient Alu short interspersed element provides a highly conserved vitamin D-mediated innate immune response in humans and primates. *BMC Genomics* 10:321.

Gu, Z., Nekrutenko, A., and Li, W. (2000) Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabytes of genomic sequence. *Gene* 259:81-88.

Haas, N.B., Grabowski, J.M., Sivitz, A.B., and Burch, J.B. (1997) Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* 197:305-309.

Haas, N.B., Grabowski, J.M., North, J., Moran, J., Kazazian Jr, H. H., and Burch, J. B. (2001) Subfamilies of CRI non-LTR retrotransposons have different 5'UTR sequences but are otherwise conserved. *Gene* 265:175-183.

Heix, J., Zomerdijk, J.C., Ravanpay, A., Tjian, R., and Grummt, I. (1997) Cloning of murine RNA polymerase I specific TAF factors: Conserved interactions between the subunits of the species-specific transcription initiation factors TIF-IB/SL1. *PNAS* 94:1733-1738.

Heyman, T., Wilhelm, M. and Wilhelm, F.X. (2003) The central PPT of the yeast retrotransposon Ty1 is not essential for transposition. *J. Mol. Biol.* 331:315-320.

Hobbs, H.H., Lehrman, M.A., Yamamoto, T., and Russell, D.W. (1985) Polymorphism and evolution of Alu sequences in the human low density lipoprotein receptor gene. *PNAS* 82:7651-7655.

Howe, K.M., Reakes, C.F.L., and Watson, R.J. (1990) Characterization of the sequence-specific interactions of mouse c-myb protein with DNA. *EMBO* 9:161-169.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* 432:860-921.

Kajikawa, M., Ohshima, K., and Okada, N. (1997) Determination of the entire sequence of turtle CR1: The first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. *Mol. Biol. Evol.* 14:1206-1217.

Kazazian, Jr., H.H. (1998) Mobile elements and disease. *Curr. Opin. Genet. Dev.* 8:343-350.

Kazazian Jr., H.H. (2004) Mobile Elements: Drivers of Genome Evolution. *Science* 303:1626-1632.

KC, M. (2008) Study of retrotransposon CR1 in waterfowl: Isolation of the element and its application as a genetic marker for phylogenetic studies. Honors Thesis, Department of Biological Sciences, University of Denver.

Kidwell, M.G. (1994) The Wilehlmine E. Key 1991 invitational lecture. The evolutionary history of the p family of transposable elements. *J Hered.* 85:339-346.

Kidwell, M.G., and Lisch, D.R. (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55:1-24.

Kleckner, N. (1981) Transposable elements in prokaryotes. *Ann. Rev. Genet*. 15:341-404.

Lodish, H., Berk, A., Kaiser, C.A., Scott, M.P., Bretscher, A., and Matsudaira, P. (2008) Molecular Cell Biology (6th Edition) New York: W.H. Freeman and Company.

Long, M., and Langley, C.H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* 260: 91-95.

Luan, D.D., and Eickbush, T.H. (1995) RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol. Cell Biol*. 15:3882-3891.

Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999) The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* 16:793-805.

Martin SL, Li PW-l, Furano AV, Boissinot S. (2005) The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenet. Genome Res.* 110:223-228.

McClintock, B. (1956) Controlling elements and the gene. *Cold Spring Harbor Symp. Quant. Biol*. 21:197-216.

Minakami, R., Kurose, K., Etoh, K., Furuhata, Y., Hattori, M., and Sakaki, Y. (1992) Identification of an internal *cis*-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nuc. Acid Res*. 20:3139-3145.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet*. 31:159-165.

Murphy, E., Huwyler, L., and Do Carmo de Freire Bastos, M. (1985) Transposon Tn554: complete nucleotide sequence and isolation of transposition-defective and antibiotic-sensitive mutants. *EMBO J*. 4:3357-3365.

Novick, P.A., Basta, H., Floumanhaft, M., McClure, M.A., and Boissinot, S. (2009) The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard *Anolis Carolineasis* shows more similarity to fish than mammals. *Mol. Biol. Evol.* 26:1811-1822.

Orgel, L.E., and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* 284:604-607.

Peters, K., Barg, N., Gärtner, K., and Rethwilm, A. (2008) Complex effects of foamy virus central purine-rich regions on viral replication. *J. Virol.* 373:51-60.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) Molecular cloning: a laboratory manual, 2nd ed. Cold Springs Harbor Laboratory Press, Cold Springs Harbor, N.Y.

Silva, R., and Burch, J.B. (1989) Evidence that chicken CR1 elements represent a novel family of retroposons. *MCB* 9:3563-3566.

Slotkin, K.R., and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genomes. *Nature* 8:272-285.

St. John, J., Cotter J.P., and Quinn, T.W. (2004) A recent chicken repeat 1 retrotransposition confirms the Coscoroba-Cape Barren goose clade. *Mol. Phylogent. Evol.* 37:83-90.

St. John, J., and Quinn, T.W. (2008a) Identification of novel CR1 subfamilies in an avian order with recently active elements. *Mol. Phylogent. Evol.* 43:1008-1014.

St. John, J., and Quinn, T.W. (2008b) Recent CR1 non-LTR retrotransposon activity in coscoroba reveals an insertion site preference. *BMC Genomics*. 9:567.

Stumph, W.E., P. Kristo, M.J. Tsai, and B.W. O'Malley. (1981) A chicken middle repetitive DNA sequence which shares homology with mammalian ubiquitous repeats. *Nucleic Acids Res*. 9:5383-5397.

Ta, Z., Li, S., and Mao, C. (2004) The changing tails of a novel short interspersed element in *Aedes aegypti*; Genomic evidence for slippage retrotransposition and the relationship between 31 tandem repeats and the poly (dA) tail. *Genetics* 168*:*2037-2047.

Thompson–Stewart D., Karpen G.H., and Spradling, A.C. (1994) A transposable element can drive the concerted evolution of tandemly repetitious DNA. *Proc Natl Acad Sci*. 91:9042-9046.

Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchinson III, A., and Edegell, M.H. (1983) The L1 Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. *Nucleic Acid Res.* 11:8847-8859.

Wicker, T., Robertson, J.S., Schulze, S.R., Feltus, F.A., Magrini, V., Morrison, J.A., Mardis, E.R., Wilson, R.K., Peterson, D.G., Paterson, A.H., and Ivarie, R. (2004) The repetitive landscape of the chicken genome. *Genome* 15*:*126-136.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J., Capy, P., Chalhoub, B., Flavel, A., Leroy, P., Morgante, M., Panuad, O., Paux, E., and Schulman, A.H. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8*:* 973-982.

Wilhelm, M., and Wilhelm, F. X. (2001) Reverse transcription of retroviruses and LTR retrotransposons. *Cell. Mol. Life. Sci.*, 58:1246-1262.

Appendix A

```
AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [  48]
MKC_CR1_Cosc  GCG CCC CCT CCC CAG CTG TTC AAA AGC AGC CCC GAG GGA GTG CGA GCA  [  48]
1H_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- T.. ...  [  48]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --. ..T .G. .A. A..  [  48]
CR1-F_Chicken --- --- --- --. T.A ACA AAG TG. GAT GCT TTT .GA ..G C.. TAG ATC  [  48]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [  96]
MKC_CR1_Cosc  AC- CAG GGT GGG CAA ACA --G GGC GTG GCG TGT CAG AAG GGC --G TGG  [  96]
1H_CR1_CBG    ..C ... ... ... ... ... --. ..A .C. A.. C.G ... TTA .CG AG. CA.  [  96]
7C_CR1_CBG    GGG .GT ... .A. TC. GA. --. ... .C. ... C.G ... TTT .CT CG. CA.  [  96]
CR1-F_Chicken T.T ..T .T. CTC .CC C.. TTA .T. C.. A.- .AA TGT .G. A.. AA. ...  [  96]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [ 144]
MKC_CR1_Cosc  -TG CAG CAG TGC G-- -CG CGG ACA GAG AGG GCA GAG TGT CT- -CC CTG  [ 144]
1H_CR1_CBG    -.T AGC GCA G.. AG- -G. ... G.. .GC ... ..G ... C.C TCA C.. .C.  [ 144]
7C_CR1_CBG    -.T ..C GC. ... AGG C.A ..C T.. .GC ... ... ... ... TCT G.. .C.  [ 144]
CR1-F_Chicken G.C TT. TG. CAG TAT A.. T.A GTG TTT CTT TTG TTC CT. TCG T.T TC.  [ 144]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [ 192]
MKC_CR1_Cosc  CCC ATA CGA ACA ACG CCT CTG --- --- --- --C -CA CTA GCT AGG CTG  [ 192]
1H_CR1_CBG    ... ... G.G ... C.T ... ..A AGG GCA CTC TC. -.G TC. ... G.. ..A  [ 192]
7C_CR1_CBG    ... ... A.. ..G ..T .A. ..A ACG GCT GTC TA. -.. ... ... ... .CA  [ 192]
CR1-F_Chicken .TT .GG A.. C.. G.T .TC ..A CGG ACA GTC CCT GTG ACC T.C ..T T..  [ 192]

AA MKC Cosc   ---  M   V   Y   T   R   Q   S   A   L   S   R   Q   S   V   T   [ 240]
MKC_CR1_Cosc  CA- ATG GTC TAC ACC AGG CAG AGT GCT CTC TCC AGA CAG TCT GTA ACC  [ 240]
1H_CR1_CBG    ..- ... ... .C. .G. ... .GC G.. ... T.. ..T ..G A.. ..C ... CA.  [ 240]
7C_CR1_CBG    ..- G.. .C. ... ..T ... ... ..A ... ... ..T ... A.. ... ... ...  [ 240]
CR1-F_Chicken A.C ... ... .CT ... ... ... C.G ... TG. A.. ..G A.. A.. ..G G.A  [ 240]

AA MKC Cosc    T   Q   T   D   C   L   L   N   N   A   A   V   Q   V   S   G   [ 288]
MKC_CR1_Cosc  ACC CAG ACT GAC TGC CTG CTC AAC AAT GCG GCA GTT CAG GTC TCT GGA  [ 288]
1H_CR1_CBG    ... ... ... ... ... .C. TC. ..A C.. ... ... ... ... ... ..C ...  [ 288]
7C_CR1_CBG    ... ... ... ... ... .CA ... ..A ... ... ... ... ... ... ... ...  [ 288]
CR1-F_Chicken ..T ... ..A ..G G.. ... .C. .GA ... .T. ..C ..G ... ... ... ..C  [ 288]

AA MKC Cosc    Y   R   E   C   L   S   L   L   L   P   S   E   G   G   R   G   [ 336]
MKC_CR1_Cosc  TAC AGG GAG TGT CTG AGC CTG TTG CTG CCA TCT GAG GGA GGC AGA GGG  [ 336]
1H_CR1_CBG    .G. G.. ... ... ... ... ..C ... ... ... ..G .C. ... ... ..G ...  [ 336]
7C_CR1_CBG    .G. ... ... ... ... ... ... ... ... ..C ... ... ... ... ... .AT  [ 336]
CR1-F_Chicken .G. ... .GT ..C .AC ... ... C.. ... ..G AGG ... .AT ..G .AG .AT  [ 336]

AA MKC Cosc    T   V   C   V   R   C   E   Q   V   D   D   L   V   R   L   V   [ 384]
MKC_CR1_Cosc  ACT GTG TGT GTG AGG TGC GAG CAG GTG GAT GAC CTG GTC CGC CTG GTG  [ 384]
1H_CR1_CBG    ... .CT ..C ... ... ... ... ... ... ..C ... ... ... ... A.. ...  [ 384]
7C_CR1_CBG    ... ... ... ... ... ... ... ... .C. ... ... ... ... T.. ... ...  [ 384]
CR1-F_Chicken G.C ACT ... ... ... ... ... ... ..A ... ..G ... C.. A.. ... ...  [ 384]

AA MKC Cosc    A   E   L   K   E   E   V   E   R   L   R   A   I   R   E   C   [ 432]
MKC_CR1_Cosc  GCA GAA CTC AAG GAG GAG GTG GAG AGG TTG AGG GCT ATC AGG GAG TGT  [ 432]
1H_CR1_CBG    ..G ... ... C.. ... ... ... C.. ... ... ... .A. ... ... ... ..C  [ 432]
7C_CR1_CBG    A.. ..G ... .G. ... ... ... ... ... ... ... ..C ... ... ... ...  [ 432]
CR1-F_Chicken .T. ..G ... .G. ... ... ... ... ..A ..A ..A A.C ... ..A ... ..C  [ 432]
```

```
AA MKC Cosc    E   R   E   I   D   W   W   S   D   S   L   Q   G   L   K   E   [ 480]
MKC_CR1_Cosc  GAG CGG GAG ATA GAC TGG TGG AGC GAC TCC CTG CAG GGC CTG AAG GAG  [ 480]
1H_CR1_CBG    ... ... ... ... ... ... ... ..T ... ... ... ... .A. .G. .G. ...  [ 480]
7C_CR1_CBG    ... ... ... ... ... ... ... ... A.. ... ... ..A ... ... ... ...  [ 480]
CR1-F_Chicken ... ... ... ..T ... ... ... ..T ... ..T ... --- --- --- .GA ..A  [ 480]

AA MKC Cosc    R   Y   R   G   E   T   P   Q   M   G   V   D   P   L   P   C   [ 528]
MKC_CR1_Cosc  AGG TAC CGG GGT GAG ACA CCC CAA ATG GGG GTA GAC CCC CTG CCC TGT  [ 528]
1H_CR1_CBG    ... G.. .A. .A. ... ..G ... ... .G. ... ..G ... ... ... ... ...  [ 528]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ..G ... ... ... T.. ...  [ 528]
CR1-F_Chicken ..A .G. A.A .A. TCC GTC ... ..G .CC .TA ..G ... ..T T.C ..T ..C  [ 528]

AA MKC Cosc    R   C   R   A   E   G   G   D   L   G   V   E   E   E   W   R   [ 576]
MKC_CR1_Cosc  CGC TGT CGG GCA GAG GGA GGG GAC CTA GAA GTT GAG GAG GAA TGG AGA  [ 576]
1H_CR1_CBG    T.. CA. ... ... ... ... ... ... ..G CG. ... ... ... ... ... ...  [ 576]
7C_CR1_CBG    ..- --- --. ..T ... ... ... ... ... .G. ... ... ... ... ... ...  [ 576]
CR1-F_Chicken T.T A.. ..A ... C.. A.. .CT ..T TC. AG. CAA ... A.. ... ... CAG  [ 576]

AA MKC Cosc    P   V   P   A   R   P   R   R   R   R   P   P   L   L   A   P   [ 624]
MKC_CR1_Cosc  CCG GTC CCT GCT CGA CCT CGC AGG CGA CGC CCC CCC CTA CCG GCC CCA  [ 624]
1H_CR1_CBG    .A. ... ... T.. ... ... ... ... A.. T.. T.T ... ..T ... ..G .TG  [ 624]
7C_CR1_CBG    .A. ... ... ... ... ... ... ... .A. .-- --- --- -.. ..T ... ...  [ 624]
CR1-F_Chicken .A. ... ..A ... ..G AAA T.. ... ... .C. ..A ... .A. ..T A.. A.G  [ 624]

AA MKC Cosc    P   S   Q   V   P   L   Y   N   R   F   E   A   L   E   L   E   [ 672]
MKC_CR1_Cosc  CCT TCC CAG GTG CCC TTA TAC AAC AGA TTT GAG GCC CTG GAG CTT GAG  [ 672]
1H_CR1_CBG    ... ... ... ... ... ..G A.. ..T ..G ... ... ... ... ... ... ...  [ 672]
7C_CR1_CBG    ... .G. ... ... ... ... C.. ... ... ... ... ... ... ..A ..C ...  [ 672]
CR1-F_Chicken GT. C.T ... ... T.T C.. C.T ..T ..G ... ..T ... ... A.A .CG ...  [ 672]

AA MKC Cosc    R   P   V   G   E   D   E   V   X   R   L   P   R   R   M   P   [ 720]
MKC_CR1_Cosc  AGA CCA GTG GGT GAG GAT GAG GTA GAA AGK CTA CCC AGG AGG ATG CCT  [ 720]
1H_CR1_CBG    ... .TG ... ..G ... ..C ... ... .G. ..C ... ... ... ... ... ...  [ 720]
7C_CR1_CBG    ... ..G ... ... ... ... ... ... ... ..T ..G ... ... ... ... ..G  [ 720]
CR1-F_Chicken G.. GAG ..A A.. ... ... A.. ..G .G. ..A ..G ..T CCA ..T G.. A..  [ 720]

AA MKC Cosc    R   A   R   K   P   T   S   R   L   E   T   A   S   I   K   K   [ 768]
MKC_CR1_Cosc  AGG GCG AGG AAG CCG ACT CCA CGC CTC GAG ACT GCC TCC ATC AAG AAA  [ 768]
1H_CR1_CBG    GA. .T. ... ... TT. ... ... ... ... AG. ... ... ... .C. ... ...  [ 768]
7C_CR1_CBG    ... ... .A. ... T.. ... ... ... ... A.. ... ... ... .C. ... ...  [ 768]
CR1-F_Chicken .A. .T. ... .G. T.A G.. ... ... T.A A.. ... ..T ... TC. ..A ...  [ 768]

AA MKC Cosc    D   R   R   V   I   V   V   G   D   S   L   L   R   G   T   E   [ 816]
MKC_CR1_Cosc  GAC AGA AGG GTG ATT GTT GTA GGC GAC TCC CTT CTC AGG GGA ACG GAG  [ 816]
1H_CR1_CBG    ..A ... ... ... ... ... ..G ... ... ... ..C ... ... ... ..N ...  [ 816]
7C_CR1_CBG    ..T ... ... ..A ..C ... ... ..T ..T ... ... ... ... ... ..A ...  [ 816]
CR1-F_Chicken ..A ..G .A. ... ..A ... ... ..T ... ... ..A ..G ..A ... ..A ...  [ 816]

AA MKC Cosc    G   P   I   C   R   P   D   P   T   H   R   E   V   C   C   L   [ 864]
MKC_CR1_Cosc  GGC CCC ATT TGT CGA CCT GAC CCT ACC CAC AGG GAA GTC TGC TGC CTC  [ 864]
1H_CR1_CBG    ... ..T ... ... ..G ... .C. ..C ..G ..T ... ... ... ... ... ...  [ 864]
7C_CR1_CBG    ... ..T ... ... .AG ... ... ... ... GGT ... ... ... ... ... ...  [ 864]
CR1-F_Chicken ... .TT ..A ... .AG ... ... ... ... .G. ... ..G ..G ... ... ...  [ 864]

AA MKC Cosc    P   G   A   R   V   R   D   V   A   R   K   L   P   N   L   V   [ 912]
MKC_CR1_Cosc  CCT GGG GCC AGG GTC TGG AAT GTT GCC AGG AAG CTT CCC AAC CTG GTT  [ 912]
1H_CR1_CBG    ... ... ... ... ... A.A G.C ... ... ... ... ... ..A ... ... ...  [ 912]
7C_CR1_CBG    ... ... .AT ... ... A.. G.. ... ... ..A G.. ... ... ..T ... ...  [ 912]
CR1-F_Chicken ... ... ... C.. ... A.. G.. A.. T.. ... ..A ... ..T GGT ... A.C  [ 912]

AA MKC Cosc    R   P   S   D   Y   Y   P   L   L   I   V   Q   A   G   S   D   [ 960]
MKC_CR1_Cosc  CGC CCC TCT GAC TAC TAC CCC CTT TTG ATA GTC CAG GCT GGC AGT GAT  [ 960]
1H_CR1_CBG    ..G ... ... ... ..T ... ..G ... ... ... ... ... ... ... ... ...  [ 960]
7C_CR1_CBG    .A. ..A ... ..T ... ..T ..T ... ... ... ... ... T.. ... ... ...  [ 960]
CR1-F_Chicken .A. ... ... ..T ... ..T ..A T.A ... ... ..T ... ... ..A ..C ...  [ 960]
```

```
AA MKC Cosc    D   I   E   E   R   S   L   K   A   I   K   R   D   F   R   G   [1008]
MKC_CR1_Cosc  GAT ATT GAA GAG AGA AGC CTG AAG GCT ATC AAA CGG GAC TTT AGG GGA  [1008]
1H_CR1_CBG    ... C.. ... A.. ... ... ... ... ... ... ... .A. ... .A. ... ..G  [1008]
7C_CR1_CBG    A.. ... ... ... ... ... ... ... ... ... ... .T. ... ... ... ...  [1008]
CR1-F_Chicken ..G G.. .CT ..C ... ... ... .G. ... ..A ..G AAT T.T ... ... ...  [1008]

AA MKC Cosc    L   G   R   L   V   D   G   A   G   V   Q   V   V   F   S   S   [1056]
MKC_CR1_Cosc  CTG GGA TGA TTA GTA GAT GGA GCG GGG GTA CAG GTG GTG TTT TCG TCT  [1056]
1H_CR1_CBG    ... ... C.. ..G ..G ..G ... ... ..A ..G ... ... ... ... ... ...  [1056]
7C_CR1_CBG    ... ... ..G ... ..G ... ... .SR ..A ... ... ... ... ... ... ..C  [1056]
CR1-F_Chicken ... ..G A.G ... ..T ..C AAT ..A ..C A.G ..A ..A ..A ... G.A GG.  [1056]

AA MKC Cosc    I   P   A   V   A   G   R   G   T   E   R   T   R   K   A   H   [1104]
MKC_CR1_Cosc  ATC CCT GCA GTG GCA GGG AGG GGT ACC GAG AGG ACA CGG AAA GCC CAC  [1104]
1H_CR1_CBG    ... ... A.G .G. .A. ..C ..A ..C ..G ... ... ... T.. ... ..T ...  [1104]
7C_CR1_CBG    ... ... A.. ... ... ... ... ... ..A ..C ... ... T.. ... ... ...  [1104]
CR1-F_Chicken ..T ... A.. ... ... ... .AA .A. .AT .CA GT. ..C A.. ... A.. ..T  [1104]

AA MKC Cosc    L   I   N   T   W   L   R   G   W   C   H   H   R   N   F   G   [1152]
MKC_CR1_Cosc  CTG ATT AAC ACG TGG CTC AGA GGC TGG TGT CAC CAC AGA AAT TTT GGG  [1152]
1H_CR1_CBG    G.. G.. ..T .G. ... ... ... ... ... ..C ..G ... ... ... ... ...  [1152]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ..C ..A ... ..G C.. ... ..T  [1152]
CR1-F_Chicken .AC ..A ... ..A ... ..G ... ... ... ..C A.A .G. .AG ... C.C ...  [1152]

AA MKC Cosc    F   F       D   H   G   A   L   Y   S   A   P   G   L   V   T   [1200]
MKC_CR1_Cosc  TTT TTT --- GAC CAT GGG GCG CTT TAC TCG GCA CCT GGC CTG GTG ACT  [1200]
1H_CR1_CBG    ... ... --- C.. ... ... .A. T.. ... ..A ... ... ... ... A.. G.C  [1200]
7C_CR1_CBG    ... ... T-- ... ... ... ... ... ... ... ... ..C ... ... A.. G.C  [1200]
CR1-F_Chicken ..C ... --- ..T ... ... ... A.C ... .T. ... ... ... ... AG. TT.  [1200]

AA MKC Cosc    A   D   G   S   L       S   C   R   G   K   W   I   L   A   Q   [1248]
MKC_CR1_Cosc  GCA GAT GGG TCC CT- --A TCT CTA AGG GGA AAA TGG ATC CTA GCC CAG  [1248]
1H_CR1_CBG    C.. ... ... ... ..- --. ... .C. ... ... ... C.. ... ... .G. ...  [1248]
7C_CR1_CBG    ... ..C ... ... .AC CG. ... ... G.. ... ... C.. ... ... ... ...  [1248]
CR1-F_Chicken ... ... ... .AT .AC CTG ... .A. ... ..G ... C.. .C. ..T ... ...  [1248]

AA MKC Cosc    E   L   A   G   L   I   E   R   A   L   N   *                   [1296]
MKC_CR1_Cosc  GAG CTG GCA GGG CTC ATT GAG AGG GCT TTA AAC TAG GAA AGA AGG GGG  [1296]
1H_CR1_CBG    ... ... ..G ... ... ..A ... ..A ... ... ... ... .C. C.. ... ...  [1296]
7C_CR1_CBG    ... ... ... ... ..T ... ... ... ... ... ... ... .T. ... ... ...  [1296]
CR1-F_Chicken ... ... ... ..A ..T G.A ... ... T.. ..G ... ... AC. T.. G.. ...  [1296]

AA MKC Cosc                                                           M    [1344]
MKC_CR1_Cosc  ATG GGG CTG AAA TTA GGC TTG TTG GAG CTG TAC CAG GGG GAA YAA TGG  [1344]
1H_CR1_CBG    .C. ... ..C ... CA. ... ... ... ... .C. .G. .G. ... A.. C.. ...  [1344]
7C_CR1_CBG    ... ... ... ... .A. C.. ... ..A ... G.. .G. ... ... ... C.. ...  [1344]
CR1-F_Chicken .A. ... A.A ... CC. ... C.. C.A ... A.. AG. .T. ... ... CG. .AA  [1344]

AA MKC Cosc    A   R   P   G   E   K   A   M   A   Q   L   K   C   I   Y   T   [1392]
MKC_CR1_Cosc  CAA RGC CRG GGR AGA AGG CAA TGG CCC AAC TGA AGT GCA TCT ACA CTA  [1392]
1H_CR1_CBG    .T. G.. TG. ..A ... ... .G. ... ... .G. ... ... ... ... ... ...  [1392]
7C_CR1_CBG    ... G.. TG. ..G ... ... ... ... ... .G. ... ... ... ... ... ...  [1392]
CR1-F_Chicken T.G GAT TA. ..G T.. G.C G.G A.. ... .G. .A. ... .TG ... .T. .C.  [1392]

AA MKC Cosc    N   A   R   S   M   G   N   K   Q   E   E   L   E   A   I   V   [1440]
MKC_CR1_Cosc  ATG CAC GCA GCA TGG GTA ACA AAC AAG AGG AGC TGR AAG CCA TCG TGC  [1440]
1H_CR1_CBG    ... ... ... ... ... ... ... ... .G. ... ... ..G ... ... ... ...  [1440]
7C_CR1_CBG    ... ... T.. ... ... ... ... .G. ... ... ... ..G ..T .A. .T. ...  [1440]
CR1-F_Chicken ... ... A.. ... ... .C. ... ... ... .A. ..T .AG ... ... .T. ...  [1440]

AA MKC Cosc    Q   Q   A   G   Y   D   L   V   A   I   T   E   T   W   W   D   [1488]
MKC_CR1_Cosc  AGC AGG CAG GCT ATG ACT TGG TTG CCA TCA CGG AAA CGT GGT GGG ACC  [1488]
1H_CR1_CBG    ... ... A.. ... .C. ... ... ... ... ... ... .G. ... ... ... ...  [1488]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... .T. ... ... ... ... ...  [1488]
CR1-F_Chicken ... ... ..A ... ... ..C .A. ... ... .T. .A. ... ... ... ... ...  [1488]
```

```
AA MKC Cosc   H   S   H   D   W   S   A   A   M   S   G   Y   R   L   F   R    [1536]
MKC_CR1_Cosc  ACT CTC ATG ACT GGA GTG CTG CAA TGT CTG GCT ATA GGC TCT TCA GAA  [1536]
1H_CR1_CBG    G.. ... ... ... ... ... ... ... ..C ... ... ..C ... ... ... .G.  [1536]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  [1536]
CR1-F_Chicken G.. .C. ... ... ... ... ... ... ..G A.. ... .C. AA. ... ... .G.  [1536]

AA MKC Cosc   R   D   R   Q   H   R   R   G   G   G   V   A   L   Y   I   R    [1584]
MKC_CR1_Cosc  GGG ACA GGC AGC ACA GAA GGG GTG GTG GTG TGG CTC TCT ATA TTA GAG  [1584]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... .C. ... ... ... .C. ... ...  [1584]
7C_CR1_CBG    ... ... ... ... ..G ... ... ... ... ... ... ... ... ... ... ...  [1584]
CR1-F_Chicken A.. .T. ... .AG GA. .G. .A. ... ... .A. ... ... ... ..G ... A..  [1584]

AA MKC Cosc   E   C   F   D   V   V   E   L   E   A   G   N   D   K   V   E    [1632]
MKC_CR1_Cosc  AGT GTT TTG ATG TTG TGG AAC TTG AGG CTG GGA ATG ATA AGG TTG AGT  [1632]
1H_CR1_CBG    .A. C.. ... ... ... ... ... .CC ... ... ... ... ... ... ... .A.  [1632]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  [1632]
CR1-F_Chicken ... .C. ... GG. ..A CT. ... ..A T.A ... ..G ..A ... ... ... .A.  [1632]

AA MKC Cosc   S   L   W   V   R   I   R   G   K   A   N   K   A   S   I   L    [1680]
MKC_CR1_Cosc  CTC TAT GGG TTA GGA TCA GAG GGA AGG CCA ACA AGG CAA GCA TCC TGG  [1680]
1H_CR1_CBG    .C. .G. ... ... ... ..C .C. ... ... ..G G.. ... .T. A.G ... ...  [1680]
7C_CR1_CBG    .C. ..A ... ... ... ... .CA ... ... ... ... ... ... ... ... ...  [1680]
CR1-F_Chicken .CT ... ... ... A.. ... .G. .A. GA. .TG ... ... ..G A.. ... ...  [1680]

AA MKC Cosc   V   G   V   C   Y   R   X   P   N   Q   D   E   E   T   D   E    [1728]
MKC_CR1_Cosc  TGG GGG TCT GTT ATA GAC CKC CAA ACC AGG ATG AGG AGA CGG ATG AGS  [1728]
1H_CR1_CBG    .C. ... ... ... ... ... .G. .G. ... ... ... ... ... ... ... ..G  [1728]
7C_CR1_CBG    ... ... ... ... ... ... .G. ... ... ... ... ... ... ... ... ..G  [1728]
CR1-F_Chicken ... .C. ... ... ... ... .G. ... ... ... ... .A. ... .A. ... ..T  [1728]

AA MKC Cosc   X   F   Y   R   Q   L   A   E   A   A   K   S   S   A   L   V    [1776]
MKC_CR1_Cosc  AGT TCT ACA GGC AGC TGG CAG AAG CCG CGA AAT CGT CAG CGC TTG TTC  [1776]
1H_CR1_CBG    ... .T. .T. ... .A. ..A ... ... TT. .A. ... ... .G. ... ... .C.  [1776]
7C_CR1_CBG    ... ... ... ... ... ... ... ... TT. ... ... .A. ... .A. ... ...  [1776]
CR1-F_Chicken T.. ... .TG A.. ... ... TG. ... .T. .AC G.. ..C .G. .CT ... .C.  [1776]

AA MKC Cosc   L   V   G   D   F   N   F   P   D   I   S   W   K   H   N   T    [1824]
MKC_CR1_Cosc  TCG TGG GGG ACT TCA ACT TCC CAG ACA TAT CCT GGA AGC ACA ACA CAG  [1824]
1H_CR1_CBG    ... ... ... ... ... ... ... .G. .T. ... ... ... ... ... ... .G.  [1824]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .A.  [1824]
CR1-F_Chicken ..A ... ... ... ... ... ... .CA ... ... G.. ..G .AT ... .TT T..  [1824]

AA MKC Cosc   A   Q   R   K   Q   S   R   R   F   L   E   S   V                [1872]
MKC_CR1_Cosc  CCC AGA GAA AGC AGT CTA GGA GGT TTC TGG AGA GCG T-- --- --- -GG  [1872]
1H_CR1_CBG    .A. ... ... ... ... ... ... ... ... ... ... .T. .-- --- --- -..  [1872]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ..A .-- --- --- -A.  [1872]
CR1-F_Chicken .A. ... AG. ... ... ... ... ... ... ... .AT ATA CAT TCC AGA A..  [1872]

AA MKC Cosc   E   D   S   F   L   T   Q   L   V   S   E   P   T   R   G   G    [1920]
MKC_CR1_Cosc  AAG ATA GCT TCC TGA CGC AGC TGG TTA GAG AGC CTA CCA GGG GAG GTG  [1920]
1H_CR1_CBG    G.. ... ... ... ... T.. ... ... .C. .T. .A. ... ... ... .T. ...  [1920]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... T..      [1920]
CR1-F_Chicken .G. .C. ... ..T ... T.. ..A ..A .A. ... .A. ... .G. .A. ... C..  [1920]

AA MKC Cosc   A   P   L   D   L   L   F   T   N   S   E   G   L   V   G   D    [1968]
MKC_CR1_Cosc  CCC CGC TAG ACC TTC TGT TCA CAA ACA GTG AMG GAC TGG TGG GAG ATG  [1968]
1H_CR1_CBG    ... ... ... ... ... .C. ... ... ... .A. .A. ... ... ... ... ...  [1968]
7C_CR1_CBG    T.. T.. ... ... ... ... ... ... ... .A. .A. ... ... ... ... ...  [1968]
CR1-F_Chicken ... .AT ... ..T .G. ... ... ... ... .A. .A. .T. ... ... ... ...  [1968]

AA MKC Cosc   V   E   V   G   S   C   L   G   Q   S   D   H   E   M   V   E    [2016]
MKC_CR1_Cosc  TGG TGG TCG GGA GCT GTC TTG GGC AGA GTG ACC ACG AAA TGG TWG AGT  [2016]
1H_CR1_CBG    ... ... ... .A. A.C A.. ... .A. ... ... ... ... ... ... .A. ...  [2016]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ..A ... ... .A. ...  [2016]
CR1-F_Chicken .A. A.A .T. ..G ... ... ... ... ... ... ... ... .C. ... .A. ...  [2016]
```

```
AA MKC Cosc   F   S   I   H   G   E   V   R   K   G   I   S   K   T   A   V    [2064]
MKC_CR1_Cosc  TCT CTA TTC TTG GCG AAG TCA GGA AGG GGA CCA GTA AAA CCG CTG TAT  [2064]
1H_CR1_CBG    ... ... ..G ... ... .G. C.. ... ... ... ... ... ... .T. ... .C.  [2064]
7C_CR1_CBG    ... ... ... ... ..A ... ... ... ... ... ... ... ... ... ... .C.  [2064]
CR1-F_Chicken ... .G. ... ... .T. G.. ... ... G.A ... A.. .C. ... .T. ..A CC.  [2064]

AA MKC Cosc   L   D   F   R   R   A   D   F   E   L   F   R   T   L   V   G    [2112]
MKC_CR1_Cosc  TGG ACT TCC GGA GGG CTG ACT TTG AGC TGT TCA GGA CAC TGG TTG GCA  [2112]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... ..C ... ... .G. ... ... ..C  [2112]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... .A. ... ... ... ... ...  [2112]
CR1-F_Chicken ... ... ... A.. ... .A. ... ... .AT ... ... ... G.. .A. .A. .G.  [2112]

AA MKC Cosc   R   V   P   W   E   A   V   L   K   G   R   G   V   Q   E   G    [2160]
MKC_CR1_Cosc  GAG TCC CTT GGG AGG TGG TTC TGA AGG GCA GAG GAG TCC AGG AAG GCT  [2160]
1H_CR1_CBG    ... ... ... ... ... C.. ... ... ... ... ... ... ... G.. ... ...  [2160]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  [2160]
CR1-F_Chicken ... ... .C. ... GTT CA. .CT .AG ..A .T. A.. .G. .A. .A. .T. ...  [2160]

AA MKC Cosc   W   A   L   F   K   K   E   I   L   M   A   Q   E   R   S   V    [2208]
MKC_CR1_Cosc  GGG CAC TCT TCA AGA AGG AAA TCT TAA TGG CTC AGG AGC GGT CTG TCC  [2208]
1H_CR1_CBG    ... .G. ..C ... ... ... ... ... ... C.. .A. ... ... ... .C. ...  [2208]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ..G ... ... ... ... ... T.. ...  [2208]
CR1-F_Chicken ..T TG. ... ... ... ... ..G ... ... A.. .G. ... ... A.G ... .A.  [2208]

AA MKC Cosc   P   R   C   P   K   T   S   R   R   G   R   R   P   A   W    L   [2256]
MKC_CR1_Cosc  CCA CGT GCC CAA AGA CGA GCC GGC GCG GAA GAA GAC CGG CCT GGC TGA  [2256]
1H_CR1_CBG    ... ... ... ... ... ... ... ... .T. ... ... ... ... ... ... .C.  [2256]
7C_CR1_CBG    ... ... ... ... ... ... ... ... A.. .T. ... .T. ... ... ... ...  [2256]
CR1-F_Chicken ..C T.A ... GC. .A. T.. ... ... ... ... ... ... ... TG. ..A ...  [2256]

AA MKC Cosc   N   R   E   L   W   L   E   L   R   R   K   K   R   V   Y   N    [2304]
MKC_CR1_Cosc  ACA GAG AGT TGT GGC TCG AGC TTA GGA GAA AAA AGA GGG TTT ATR AYC  [2304]
1H_CR1_CBG    ... ... ... ..C ... .T. T.. ... .C. .G. ... ... ..T ... ..A .T.  [2304]
7C_CR1_CBG    ... ..- -.. ... ... ... ... ... ... ... ... ... ... ... ..A .T.  [2304]
CR1-F_Chicken .T. .G. .AC .A. TCT .GA G.. .CC A.G AG. ... ... .AA .C. .CC TC.  [2304]

AA MKC Cosc   L   W   K   R   G   R   A   T   Q   E   D   Y   K   D   V   A    [2352]
MKC_CR1_Cosc  TTT GGA AAA GAG GGC GGG CYA CTC AAG AGG ACT ATA AGG ATA TTG CAA  [2352]
1H_CR1_CBG    ... ... ... A.. ... ... .C. ..G .G. ... ... .C. ... ..G .A. .G.  [2352]
7C_CR1_CBG    ... ... ... ... ... ... .C. ... .G. ... ... ... ... ..G ... .G.  [2352]
CR1-F_Chicken .G. ... .G. AG. .A. ... .A. .C. GGA .A. .A. .C. .A. .AG ..A TT.  [2352]

AA MKC Cosc   R   L   C   R   D   K   I   R   K   A   K   A   H   L   E   L    [2400]
MKC_CR1_Cosc  GGC TGT GCA GGG ACA AAA TYA GAA AGG CCA ARG CTC ATC TGG AGC TCA  [2400]
1H_CR1_CBG    ... ... ... ... .G. ... .T. ... ... ... .A. ... ... ... ... ...  [2400]
7C_CR1_CBG    ... .T. ... ... ... ... .T. ... .A. ... .A. ... ... ... ... ...  [2400]
CR1-F_Chicken A.A ... ... ... .G. ... .C. ... ... .A. .A. .C. .G. .T. .A. ...  [2400]

AA MKC Cosc   N   L   A   T   A   V   K   D   N   K   K   C   F   Y   K   Y    [2448]
MKC_CR1_Cosc  ATC TGG CTA CTG CCA TTA AAG ACA ACA AAA AAC GAT TTT ACA AAT ACA  [2448]
1H_CR1_CBG    .G. ... ... ... .TG ... ... ... ... ... ..T CC. ... ... ... .T.  [2448]
7C_CR1_CBG    ... ... ... ... ... ... ... .T. ... ... .-T .C. ... .T. ... ...  [2448]
CR1-F_Chicken .C. ... ..G ... GGG .A. ... GG. ... .G. ... TC. ... ... .G. .T.  [2448]

AA MKC Cosc   I   N   T   K   R   R   T   K   E   N   L   H   P   L   L   D    [2496]
MKC_CR1_Cosc  TCA ACA AAA AAA GGA GGA CTA AGG AGA ATC TCC ATC CTT CAC TGG ATG  [2496]
1H_CR1_CBG    ... ..G C.. ..C ... ... ... ... ... ... ... ... ... T.. ... ...  [2496]
7C_CR1_CBG    ... ... TG. ... ... ... ... ... ... ... ... ... ... T.. ... ...  [2496]
CR1-F_Chicken ... ... GT. .G. ... ... .C. G.. ... ... ... ..T ..C T.. ... ...  [2496]

AA MKC Cosc   A   G   G   N   L   V   T   R   D   E   E   K   A   E   V   L    [2544]
MKC_CR1_Cosc  CGG GGG GAA ACT TAG TTA CAA RAG ATG AGG AAA AGG CTG AGG TGC TYA  [2544]
1H_CR1_CBG    ..A ... ... ..C ... .C. .T. AG. ... ... ... ... ... ... ... .T.  [2544]
7C_CR1_CBG    ... ... ... G.. ... ... ... G.. ... ... ... ... ... ... ... .C.  [2544]
CR1-F_Chicken A.. CT. .G. .TG .GA CC. .TG AG. ..A ... ... ... .A. .C. .T. .G.  [2544]
```

```
AA MKC Cosc   N   A   F   F   A   S   V   F   S   G   K   T   S   C   S   L    [2592]
MKC_CR1_Cosc  ATG CCT TCT TTG CCT CAG TCT TTA GCG GCA AGA CCA GTT GTT CTC TGG  [2592]
1H_CR1_CBG    ... .TG C.. ... ... ... ... ... ... ... .T. ..G ... ... ... ...  [2592]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  [2592]
CR1-F_Chicken ... ... ... ..A .A. .T. ... ... AAA .TC ... ... ... A.C ... A..  [2592]

AA MKC Cosc   D   T   R   Y   P   E   L   V   E   G   Y   G   E   Q   D   V    [2640]
MKC_CR1_Cosc  ATA CCC GGT ACC CTG AGC TGG TGG AAG GGW ATG GGG AGC AGG ATG TGG  [2640]
1H_CR1_CBG    ... ... A.. G.. ... ... ... C.. ... ..G ... ... ... ... ... ...  [2640]
7C_CR1_CBG    ... ... A.. C.. ... ... ... ... ... ..G ... ..T ..A ..A ... ...  [2640]
CR1-F_Chicken T.T .T. CAC T.T ... .C. ... CA. CCC T.G C.. ... ... ..A C.A AAC  [2640]

AA MKC Cosc   A   L   T   I   H   E   E   M   V   G   D   L   L   Q   H   L    [2688]
MKC_CR1_Cosc  CCC TCA CTA TCC ACG AGG AAA TGG TTG GCG ACC TGC TAC AGC ACT TGG  [2688]
1H_CR1_CBG    ... ... ..G ... ..A ... ... ... ... ... ... ... ..A G.A G.. ...  [2688]
7C_CR1_CBG    ... ... .A. ... .T. ... ... ... ... .TA ... ... ... ... ... ...  [2688]
CR1-F_Chicken .T. C.. .A. .T. ... ... ... CA. .CA .A. ... ... ... .C. .AC ...  [2688]

AA MKC Cosc   D   V   R   K   S   M   G   P   D   G   I   H   P   R   V   L    [2739]
MKC_CR1_Cosc  ATG TAC GCA AGT CGA TGG GGC CGG ATG GGA TCC ACC CGA GGG TAC TGA  [2736]
1H_CR1_CBG    ... .G. ... ... ... ... ... ... ... ... .G. ... ... ... ... ...  [2736]
7C_CR1_CBG    ... ... A.. ... ... ... .T. T.. .C. ... ... ... T.. ... .G. ...  [2736]
CR1-F_Chicken .CT GC. A.. ... .C. ... AA. .A. ... A.. .T. ... ... .A. .G. ...  [2736]

AA MKC Cosc   R   E   L   A   E   E   L   A   K   P   L   S   I   I   Y   R    [2784]
MKC_CR1_Cosc  GAG AAC TGG CGG AGG AGC TGG CCA AGC CGC TTT CCA TCA TTT ATC GGC  [2784]
1H_CR1_CBG    A.. ... ... .A. ... ... ... ..G ... ... ... ... ... ... ... ...  [2784]
7C_CR1_CBG    .-- ... ... .A. ... ... ... G.. ... T.. ... ... ... ... ... ...  [2784]
CR1-F_Chicken .G. ... ... .A. ... T.A .A. ..G ... ... ... ... ... .C. ... A..  [2784]

AA MKC Cosc   Q   S   W   L   S   G   E   V   P   V   D   W   R   L   A   N    [2832]
MKC_CR1_Cosc  AGT CCT GGC TAT CAG GGG AGG TCC CAG TCG ACT GGC GGC TAG CAA ACG  [2832]
1H_CR1_CBG    ... ... ... ... .G. T.. ... ... ... .T. ... ... ... ... .C. .T.  [2832]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ..A ... ... A.. ... ... .T.  [2832]
CR1-F_Chicken GC. ... T.T .GA .G. .T. ... ... ... AA. ... ..A ... .T. .C. .T.  [2832]

AA MKC Cosc   V   T   P   I   Y   K   K   G   R   R   V   D   P   G   N   Y    [2880]
MKC_CR1_Cosc  TGA CGC CCA TCT ACA AGA AGG GCC GGA GGG TAG ACC CGG GGA ACT ATA  [2880]
1H_CR1_CBG    ... ... ... ... .T. ... ... ... ... ... C.. ... ... ... ... ...  [2880]
7C_CR1_CBG    ... .A. ... ... ... ... ... ..T ... ... ... ... ... ... ... ...  [2880]
CR1-F_Chicken ... .T. ... ... C.. ... ... ..T .C. ... AG. .T. ... ... ... CC.  [2880]

AA MKC Cosc   R   P   V   S   L   T   S   V   P   G   K   L   M   E   Q   I    [2928]
MKC_CR1_Cosc  GGC CTG TTA GTT TGA CCT CAG TGC CAG GGA AGC TCA TGG AGC AGA TTA  [2928]
1H_CR1_CBG    ... ... .C. ... ... ... ... ... .G. .A. ... ... ... ... ... ...  [2928]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... .-. ... ..T ... ... ... ...  [2928]
CR1-F_Chicken ... ... ... .CC ... ... .G. ... .G. ... ..A .T. ... ... ... ..G  [2928]

AA MKC Cosc   I   L   S   V   I   T   R   H   L   Q   G   N   Q   A   I   R    [2976]
MKC_CR1_Cosc  TCT TGA GTG TCA TCA CGC GGC ACT TGC AGG GCA ACC AGG CGA TCA GGC  [2976]
1H_CR1_CBG    ... ... .G. ... ... ... ... ... ... ... ... .G. ... ... ... ...  [2976]
7C_CR1_CBG    ... ... ... ... ... ..T ... ... ... ... ... ... ... ... ... ...  [2976]
CR1-F_Chicken ... ... .G. AG. ... ..T ... ..G ... ... A.. ... G.. G.. ... ...  [2976]

AA MKC Cosc   P   S   Q   H   G   F   M   K   G   R   S   C   L   T   N   L    [3024]
MKC_CR1_Cosc  CCA GTC AGC ATG GGT TTA TGA AAG GCA GGT CCT GCT TGA CGA ACC TGA  [3024]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... .C. ... ...  [3024]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..G  [3024]
CR1-F_Chicken .T. .C. ... ... ... .C. C.. .G. ... ... ... ... ... .C. ... ...  [3024]

AA MKC Cosc   I   S   F   Y   D   K   V   T   R   L   V   D   X   G   K   S    [3072]
MKC_CR1_Cosc  TCT CCT TCT ATG ACA AAG TGA CGC GCT TGG TGG ATG ASG GGA AGG CTG  [3072]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... .G. .A. G.. .G.  [3072]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... .A. ... ... .G. .A. ... ...  [3072]
CR1-F_Chicken ... ... ... ... .TC C.. ... .C. .TC ... ... ... .G. .A. ... ...  [3072]
```

```
AA MKC Cosc   X   D   V   I   Y   L   D   F   S   K   A   F   D   T   V   S    [3120]
MKC_CR1_Cosc  KGG ATG TGA TCT ACC TTG ACT TCA GTA AGG CTT TTG ACA CCG TTT CCC  [3120]
1H_CR1_CBG    T.. ... ..G ... ... ... ..C ... ... ... ... ... ... ... ..C ...  [3120]
7C_CR1_CBG    T.. ... ..G ... ... ... ... ..G ... ... ... ... ... ..A ... ...  [3120]
CR1-F_Chicken TT. ... .AG ... ... .A. .T. ... .C. .A. .C. ... .A. .T. .C. ...  [3120]

AA MKC Cosc   H   N   I   L   L   X   K   L   A   A   R   G   L   D   W   R    [3168]
MKC_CR1_Cosc  ACA ACA TTC TCC TCA RGA AAC TGG CTG CTC GTG GCT TGG ACT GGC GTA  [3168]
1H_CR1_CBG    ... ... ... ... ... A.. ... ... ... ... .G. ... ... ... ... ...  [3168]
7C_CR1_CBG    ... ... ... ... ... A.. ... ... ... ..A ... A.. ... ... ... ...  [3168]
CR1-F_Chicken ... GT. ... ... .GC AA. .G. ... .A. TC. ... ... ... ..A .AT AC.  [3168]

AA MKC Cosc   T   L   R   W   V   K   N   W   L   D   G   X   A   Q   R   V    [3216]
MKC_CR1_Cosc  CGC TTC GTT GGG TTA AAA ACT GGC TGG ATG GCY GGG CCC AAA GAG TTG  [3216]
1H_CR1_CBG    ... ... .C. ... ... G.. ... ... ... ..A ..C ... ... .G. ... ...  [3216]
7C_CR1_CBG    ... ... A.. ... ... ... ... ... ... ..A ..C ..A ... ... ... ...  [3216]
CR1-F_Chicken .T. ..G .C. ... .A. GG. ... ... ... .G. ..C A.. ... .G. ... .G.  [3216]

AA MKC Cosc   V   V   N   G   A   K   X   X   S   W   X   P   X   T   S   G    [3264]
MKC_CR1_Cosc  TGG TGA ATG GAG CCA AAT CNN CAA GTT GGA GRC CGG TYA CTA GTG GAG  [3264]
1H_CR1_CBG    ... ... ... ... T.. .G. .C- --. ... ... .G. T.. .C. .A. ... .T.  [3264]
7C_CR1_CBG    ... ... ... ... T.. ... .C- --. ... ... .G. T.. .C. ... ... ...  [3264]
CR1-F_Chicken ... ... ... ... TT. ... .C- --. .C. ..G .A. ... .C. .G. ... .T.  [3264]

AA MKC Cosc   V   P   Q   G   S   V   L   G   P   V   L   F   N   I   F   I    [3312]
MKC_CR1_Cosc  TCC CCC AGG GCT CAG TGC TGG GGC CAG TCC TCT TTA ATA TCT TTA TCR  [3312]
1H_CR1_CBG    ... ... ... ... .G. .A. ... ... .G. ... ... ... ... ... ... .TG  [3312]
7C_CR1_CBG    ... .T. ... ... ... ... ... ... ... ... ... ... ... ... ... ..G  [3312]
CR1-F_Chicken .T. ... ..A .G. .G. ... ... ... .T. ... .G. ... ..G ... G.. .TG  [3312]

AA MKC Cosc   X   D   L   D   E   G   I   E   C   T   L   S   K   F   A   D    [3360]
MKC_CR1_Cosc  ATG ATC TGG ATG AGG GGA TCG AGT GCA CCC TCA GTA AGT TTG CAG ATG  [3360]
1H_CR1_CBG    ... ... ... .C. ... ... .T. ... ... ... ... ... ... ... ... ...  [3360]
7C_CR1_CBG    ... ... ... ... .A. ... ... ... ... ... ... ... ... ... ... .C.  [3360]
CR1-F_Chicken ... .C. ... ... ... .C. .T. ... ... ... ... T.. ... ... ... ...  [3360]

AA MKC Cosc   D   T   K   L   G   A   C   V   D   L   L   K   X   R   K   A    [3408]
MKC_CR1_Cosc  ACA CCA AGT TAG GTG CGT GTG TCG ATC TGC TCA AGG GRA GGA AGG CTC  [3408]
1H_CR1_CBG    ... ... ..C ..A ... ... ... ... ... ... ..G ... .C. ... ... ...  [3408]
7C_CR1_CBG    ... ... ... ... A.. .A. ... ... ... ... ..G ... .T. ... ... ...  [3408]
CR1-F_Chicken ... ... ..C ... C.. GAA ... .T. ... ... CTG ... .T. .AG ... .C.  [3408]

AA MKC Cosc   L   Q   E   D   L   D   R   L   D   Q   W   A   E   X   N   C    [3456]
MKC_CR1_Cosc  TGC AGG AGG ATC TGG ATA GGC TGG ACC AAT GGG CTG AGG YCA ACT GTA  [3456]
1H_CR1_CBG    ... ... ... ... ... ... ... ... .G. G.. ... ... ... T.. ... ...  [3456]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... G.. ... ... ... C.. ... ...  [3456]
CR1-F_Chicken .A. ..A G.. ... ... ... ..T ... .TA GC. ... ... .A. C.. .TG .G.  [3456]

AA MKC Cosc   M   X   F   N   K   A   K   C   X   V   L   H   L   G   X   N    [3504]
MKC_CR1_Cosc  TGA RGT TCA ACA AGG CCA AGT GCM GGG TCC TGC ACC TGG GGK GCA ACA  [3504]
1H_CR1_CBG    ... A.. ... ... ... ... ... ..C ... ... ... ... ... ..C ... ...  [3504]
7C_CR1_CBG    ... A.. ... ... ... ... ... ..C ... ... ... ... ... ..C A.. ...  [3504]
CR1-F_Chicken ..G GA. ... ... ..A ... .A. ..C ... ... ... ..T .T. .CC ... GT.  [3504]

AA MKC Cosc   N   P   X   Q   R   Y   R   L   G   D   E   W   L   E   S   C    [3552]
MKC_CR1_Cosc  ACC CCA WGC AGC GCT ACA GGC TGG GAG ATG AGT GGT TGG AAA GCT GCC  [3552]
1H_CR1_CBG    ... ... A.. ..A ... ... ... ... ... ... ... ... ... ... ... ...  [3552]
7C_CR1_CBG    ... ... AA. .T. A.. ... .A. ... ... ... G.. ... ... ... T.. ...  [3552]
CR1-F_Chicken ... ... G.. .AT ... ... ... .TA .G. CA. ... ..C ... ..G A.. .TG  [3552]

AA MKC Cosc   L   A   E   K   D   L   G   V   L   V   D   S   W   L   N   M    [3600]
MKC_CR1_Cosc  TGG CAG AGA AGG ACC TGG GAG TAC TGG TTG ATA GTT GGC TGA ATA TGA  [3600]
1H_CR1_CBG    ... .C. ... ... ... ... ... ..T ... ... ... ..C ... ... ... ...  [3600]
7C_CR1_CBG    ... ... ... ... ... ... ... ..T ... ... ... ... ... ... ... ...  [3600]
CR1-F_Chicken .A. AG. .A. T.. ... ... .G. .GT ..A ... .C. C.A .A. ... .C. ...  [3600]
```

```
AA MKC Cosc   S   Q   Q   C   A   Q   V   A   K   K   A   N   S   I   L   A    [3648]
MKC_CR1_Cosc  GCC AGC AGT GTG CTC AGG TGG CCA AGA AGG CCA ACA GCA TCC TGG CTT  [3648]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .C.  [3648]
7C_CR1_CBG    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...  [3648]
CR1-F_Chicken ... ..T G.. ... .C. ... ... ... ... ... ... .TG ... ... ... ...  [3648]

AA MKC Cosc   C   I   R   S   S   V   A   S   R   S   R   E   V   I   V   P    [3696]
MKC_CR1_Cosc  GTA TAA GAA GCA GTG TGG CCA GCA GGT CTA GGG AAG TGA TTG TCC CCC  [3696]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... ... .A. .G. ... ... ... ...  [3696]
7C_CR1_CBG    ..C ... ... ... ... ... ... ... ..G ... ... ... ... ... ... .G.  [3696]
CR1-F_Chicken ... .C. ... AT. ... ... ... ... ..A AC. ... ... .A. ... ..T ...  [3696]

AA MKC Cosc   L   Y   S   A   L   V   R   P   H   L   E   Y   C   V   Q   F    [3744]
MKC_CR1_Cosc  TGT ACT CGG CTC TGG TAA GGC CGC ACC TCG AGT ACT GTG TTC AGT TTT  [3744]
1H_CR1_CBG    ... ... ... ... ... .G. ... ... ... ... ... ... ... ... ... ...  [3744]
7C_CR1_CBG    ... ... ... ... ... .G. ... ... ... .T. ... ... ... ... ... ...  [3744]
CR1-F_Chicken ... ... .A. .A. ... CG. ... T.. ... .T. ... ... ... .C. ... ...  [3744]

AA MKC Cosc   W   A   P   R   Y   K   K   D   M   E   V   L   E   R   V   Q    [3792]
MKC_CR1_Cosc  GGG CCC CTC GCT ACA AGA AGG ACA TGG AGG TGC TCG AGA GAG TCC AGA  [3792]
1H_CR1_CBG    ... ... ... A.. ... G.. ... ... ... .C. ... ... ..C ... ... ...  [3792]
7C_CR1_CBG    ... ... ..T ... T.. ... ... ... ... ... ... ... ... ... ... ...  [3792]
CR1-F_Chicken ... ... ... A.. GT. ... .A. ... .T. ... CC. .G. ..T .T. ... ...  [3792]

AA MKC Cosc   R   R   X   T   K   L   V   R   G   L   E   N   K   S   Y   E  [3840]
MKC_CR1_Cosc  GAA GGG CRA CGA AGC TGG TGA GGG GTC TGG AGA ACA AGT CTT ACG AGG  [3840]
1H_CR1_CBG    ... ... .G. .C. ... ... ... ... ... ... ... ... ... ... ... ...  [3840]
7C_CR1_CBG    ... ... .G. ... ... ... ... ... ... ... ... ... ... ... .T. ...  [3840]
CR1-F_Chicken .G. ... .A. .A. ... ... ... ... ... ... .AC ... G.C ... .T. .A.  [3840]

AA MKC Cosc   E   R   L   R   E   L   G   L   F   X   L   E   K   R   R   L    [3888]
MKC_CR1_Cosc  AGC GGC TGA GGG AGC TGG GGT TGT TCA GYC TGG AGA AGA GGA GGC TCA  [3888]
1H_CR1_CBG    ... ... ... ... ... ... ... ... ... .C. ... ... ... ... ... ...  [3888]
7C_CR1_CBG    ... ... ... ... ... ... ... ... .T. .C. ... ... .T. ... ... ...  [3888]
CR1-F_Chicken ..T ... ... A.. ... ... .A. ... ... .T. ... ... ... ... ... ...  [3888]

AA MKC Cosc   R   G   D   L   I   A   L   Y   R   Y   L   K   G   C   C   S    [3936]
MKC_CR1_Cosc  GGG GAG ACC TCA TCG CTC TCT ATA GGT ACC TTA AAG GAG GCT GTA GCG  [3936]
1H_CR1_CBG    ... .C. ... .T. ... ... ... .C. .T. ... ... ... ... ... ... .T.  [3936]
7C_CR1_CBG    ... .C. ... .T. ... ... ... ... ... ..A ... ... ... ... ... ...  [3936]
CR1-F_Chicken ... ... ... .T. .T. ... ... ... AC. ... .G. .G. ... .T. ... .T.  [3936]

AA MKC Cosc   E   X   G   V   G   L   F   S   H   V   X   G   E   R   T   R    [3984]
MKC_CR1_Cosc  AGG TRG GGG TTG GTC TAT TCT CCC ACG TGC YTG GTG AGA GGA CGA GGG  [3984]
1H_CR1_CBG    ... .G. ... ... ... .G. ... ... ... ... C.. ... .C. ... ... ...  [3984]
7C_CR1_CBG    ... .A. .AA ... ... ... ... ... .T. ... C.. C.. .C. ... ... .A.  [3984]
CR1-F_Chicken ..C .C. ... .CA .C. .C. ... .T. GT. ..A C-A ... .T. .A. ... .A.  [3984]

AA MKC Cosc   G   N   X   L   K   L   H   Q   G   R   F   R   G   D   I   R    [4032]
MKC_CR1_Cosc  GGA ATG RGC TAA AGT TGC ACC AGG GTA GGT TTA GGT TGG ATA TTA GGA  [4032]
1H_CR1_CBG    ... ... G.. ... ... ... GA. ... .G. ... ... ... .T. ..G ... ...  [4032]
7C_CR1_CBG    ... ... G.T ... ... ... G.. ... .G. ... ... ... ... ... ... ...  [4032]
CR1-F_Chicken ... ... GCT .C. ..C ... G.. ... .A. .A. .C. ..C ... .C. ... ...  [4032]

AA MKC Cosc   K   N   V   F   T   E   R   V   V   R   H   W   N   G   L   P    [4080]
MKC_CR1_Cosc  AGA ACG TCT TTA CTG AAA GGG TTG TTA GGC ATT GGA ATG GGC TGC CCA  [4080]
1H_CR1_CBG    ..T ..T ... ... .C. ... ... ..A ... A.. ... ... .C. ... ... ...  [4080]
7C_CR1_CBG    ... ..T ... ... ... ... .A. ... ... ... ... ... ... ... ..A ...  [4080]
CR1-F_Chicken .AT ..T A.. .CT ... ... ... .G. .C. ... .C. ... ... ... ... ...  [4080]

AA MKC Cosc   R   E   V   V   E   S   P   S   L   E   V   F   K   R   H   L    [4128]
MKC_CR1_Cosc  GGG AAG TGG TTG AGT CAC CAT CCC TGG AGG TCT TTA AGA GAC ATT TAG  [4128]
1H_CR1_CBG    ... .G. ... .N. ... ... ... ... ... ... ... ... .A. ... G.. ...  [4128]
7C_CR1_CBG    ... ... ... ... ... .G. ... ... ... ... ... ... .A. ... G.. ..A  [4128]
CR1-F_Chicken .A. TG. ... .G. ... ... .GA G.. ... T.. .G. ... .AG AG. ... .G.  [4128]
```

```
AA MKC Cosc   D   V   A   L   S   D   M   F   *                                   [4176]
MKC_CR1_Cosc  ATG TAG CCC TTA GTG ATA TGG TTT AGT GGA GGA CTT GTT AGT GTT AGG   [4176]
1H_CR1_CBG    ... ... AG. ... .C. ... ... ... ... ... .T. ... AG. GT. AGG TC.  [4176]
7C_CR1_CBG    ... ... .A- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4176]
CR1-F_Chicken ... .T. TGT .G. .G. .C. ... ... ... .-- A.. ACC A.. G.. .AA G..  [4176]

AA MKC Cosc                                                                   [4224]
MKC_CR1_Cosc  TCA GTG GTT GGA CTA GGT GAT CTT GGA GGT CTC TTC CAA CCT AGA CGA  [4224]
1H_CR1_CBG    GAG ..T .GA CTC GAT .A. CT. GAG .TC TC. TC. AA. .T. GAA .TC T.T  [4224]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4224]
CR1-F_Chicken CG. A.. ... ... ..G .A. --C ..G T.G .TC T.T .C. A.C .T. ..C GAT  [4224]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4272]
MKC_CR1_Cosc  TTC TGT GAT CTG TGA TTC TGT GA- --- --- --- --- --- --- --- ---  [4272]
1H_CR1_CBG    G.. ... .TC TGT GTC .GT GAC TGC GCA GGG CCC GCG TTG TCC CAG CAG  [4272]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4272]
CR1-F_Chicken .CT ATG AT. ... ... ... ..A .TT TGT -TG AAA CCT GCA GAA CAG CTG  [4272]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4320]
MKC_CR1_Cosc  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4320]
1H_CR1_CBG    CAG GGC AGG GTG CCA GAT CTC GCT CTC CAC AAA GCC CTG CGC AAT GCT  [4320]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4320]
CR1-F_Chicken CCA TGT CCT TCC ACT GAG AGG CTG TGG CTA TTC CTC ATT TAA TTG TCT  [4320]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4368]
MKC_CR1_Cosc  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4368]
1H_CR1_CBG    GAC AGC ATT TGC TGC TTC CCC CAG CGT GAC CAT TTC CCT GTT TCC TTT  [4368]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4368]
CR1-F_Chicken GTC GAC ACT CTG TCA AAC AGC AAG GGA GGC ACT CTG TTT GCA CCT GTG  [4368]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4416]
MKC_CR1_Cosc  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4416]
1H_CR1_CBG    TGG GTT GGA TGT TGT GGT GCC AAA CAA GTA ACA AAA CTG TTT AAA ATA  [4416]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4416]
CR1-F_Chicken AAC AAG CTG TGC CTC AGC GAA TGC TCA TAT TCT CTC AGA AGA TGC ATG  [4416]

AA MKC Cosc   --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4464]
MKC_CR1_Cosc  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4464]
1H_CR1_CBG    CCT AAG CTA AAC TTT TTT TTT TTT TTT TTA CTT TTT TTT TTT NAA AAC  [4464]
7C_CR1_CBG    --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  [4464]
CR1-F_Chicken TTC CCA AGG AGA AAA GAC TTT CAG TGT CAA ATC TTC CAC ACC TAT TAG  [4464]
```

Appendix A. Alignment of CR1 sequences using Mega v5.1. Highlighted regions indicate the start and stop codons for ORF1 and ORF2. Several deletions, in groups of 3, were seen in 7C and are also highlight above.