THE UNIVERSITY OF SYDNEY

CHARACTERISATION OF ENDOGENOUS RETROVIRUSES IN THE SALTWATER CROCODILE (*CROCODYLUS POROSUS*)

Amanda Yoon-Yee Chong

Faculty of Veterinary Science
The University of Sydney

A thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

August, 2013

Declaration

The work presented in this thesis is original and was undertaken while I was enrolled as a PhD student in the Faculty of Veterinary Science, The University of Sydney, Australia. I certify that this thesis has not been submitted for any other degree, and that all sources of information and assistance during the experimental work and in the preparation of this thesis are duly acknowledged.

Author contribution statements signed by all co-authors to my published and submitted publications are also provided.

Amanda Yoon-Yee Chong
August 2013

Thesis Summary

This project constitutes the first comprehensive investigation of endogenous retroviruses (ERVs) in crocodilians. The studies presented in this thesis comprise two major themes. The first of these is the characterisation of ERVs from crocodilians, with a focus on those families that show potential for replication. Interwoven through the experimental chapters is an evaluation and discussion of the various methods used to identify and investigate ERVs, and their relative merits for the study of ERVs in a non-model organism.

This project was initiated prior to the availability of the crocodilian genome sequences, necessitating the use of PCR based and hybridisation based screening methods for early investigations. Prior to the studies outlined here, knowledge of ERVs in these species was mostly limited to cross species investigations of diversity based on conserved regions and a single study of the genomic organisation of a particularly divergent ERV lineage from *C. niloticus*. Thus it was important to establish baseline knowledge of the ERV diversity that may be present within a species, both from a population perspective using *C. porosus* as the model species (Chapter 2), and within individuals using samples from various tissues from *C. johnstoni* (Chapter 3). These studies provided initial insights into the range of ERV integrations that may be present in crocodilians, and revealed evidence of species specific ERV activity and evolution.

These PCR surveys also suggested that at least one lineage of ERVs had recently been active in *Crocodylidae*, prompting the targeted screening of the *C. porosus* genomic BAC library to generate additional data from other ERV domains (Chapter 4). This also allowed the reconstruction of the likely proviral genomes, and characterisation of the genomic structure of these ERVs. This study provided the first insights in to crocodilian ERVs from a genomic perspective, with a preliminary estimation of the ERV content of the *C. porosus* genome suggesting a relatively sparse, but not unusually low ERV population when compared other vertebrates. Analysis of the ERV proviral genomes suggested that many crocodilian ERVs represent intermediates between the recognised exogenous retroviral genera.

The availability of genomic sequence data facilitated further characterisation of crocodilian ERVs, providing a new perspective on the evolution of crocodilian ERVs and insights into the diversity of ERVs that may be present in non-mammalian vertebrates. Chapters 5 and 6

present a discussion of the relative merits of bioinformatics tools developed for genomic studies of ERVs, and a quantification of the ERV complement of the crocodilian genomes. From a technical perspective, these chapters highlight the difficulties of studying ERVs in taxa where these elements have not been well characterised, and offer some suggestions for the adaptation of current methodologies for such studies.

Finally, the comparative study presented in Chapter 7 represents the first comprehensive study of ERVs across non-mammalian taxa and reveals the remarkable levels of ERV diversity that may be found in the genomes of three crocodilians (*A. mississippiensis*, *C. porosus*, and *G. gangeticus*). The identification of a new class of ERVs, and a large number of intermediary ERV families is of particular significance as it lends support to theories of a gradual evolution of retroviral genera. Furthermore, these lineages are primarily basal to the other ERV classes, falling between these and the *Gypsy* transposons, suggesting that these lineages may represent preserved copies of the early predecessors of modern exogenous retroviruses. Lastly, the identification of a potentially exapted ERV in *C. porosus* is an interesting development that warrants further investigation, and further highlights the significance that ERVs and other related genomic elements play in shaping the genomes of their hosts.

Overall, this project has demonstrated that crocodilians, and likely other non-mammalian vertebrates, are a rich source of novel ERV diversity, and may provide unique insights into the evolution of modern exogenous retroviruses and their hosts. It has also highlighted the relative merits of a wide variety of ERV detection techniques, both molecular and bioinformatic, and how these may be adapted for studies of previously uncharacterised taxa. This project will provide a useful resource to facilitate further investigations into the significance of ERVs in crocodilian biology, and offers insights into how these approaches may be translated to studies of other vertebrate taxa.

Acknowledgements

There are many people to whom I owe my gratitude for their support and assistance throughout this project. First and foremost, I have to thank my supervisors Dr Jaime Gongora and Dr Sally Isberg for everything that they have done for me throughout my candidature. I thank you both for providing this opportunity to me, and allowing me to develop and expand this project according to my ideas and interests. Your encouragement and support during this process has been invaluable. I would particularly like to thank Jaime for continually challenging me with new ideas and opportunities, and for occasionally reminding me that there is a world outside of the lab.

This project would not have been possible without the assistance of many collaborators, both within the faculty and overseas. Many thanks go to my colleagues and lab mates Mr Weerachai Jaratlerdsiri, Ms Sarah Atkinson, and Ms Shannon Kjeldsen for their assistance in the lab and with the early stages of data analysis; and to staff at Darwin Crocodile Farm and Berrimah Veterinary Laboratories for the collection and provision of the tissue samples.

Thank you also to Dr David Ray and Dr Travis Glenn for providing me the opportunity to work with them in the USA. Thanks to David and Dr Daniel Peterson for allowing me access to the crocodilian BAC libraries and to their facilities at Mississippi State University; Dr Xueyan Shan and Dr Zenaida Magbanua for their assistance with preparing and screening the libraries; Travis for allowing me to work in his lab at the University of Georgia, and for his assistance and advice for the sequencing of the BAC clones; Jeffrey Wagner at Georgia Genomics Facility for setting up the libraries for sequencing; and Dr Brant Faircloth for helping me filter the resulting data.

Thanks also go to the International Crocodilian Genomics Working Group for allowing me access to the crocodilian genomes for my project. I am especially grateful to David for bringing me into this project, and to my fellow collaborators within the TE analyses: Dr Jerzy Jurka, Dr Kenji Kojima, and Dr Arian Smit for their assistance with the classification of the crocodilian ERVs.

I gratefully acknowledge the funding provided by the Rural Industries Research and Development Corporation, and the National Science Foundation (USA) for the research carried within this project. During my candidature, I was funded by a Jean Walker

Postgraduate Fellowship, and an Eric Horatio Maclean Scholarship both provided through the Faculty of Veterinary Science.

A special thank you goes to my colleagues in the office, for putting up with my crazy working habits and frequent complaints about things not working. In particular, thanks go to Jessica Fletcher, and Mette Lillie for being so supportive of all my crazy ideas and not being afraid to tell me when I've lost the plot. A big thank you also to Jane Bursill and Trung Doan for their advice with all things lab related; and David Liu and Michael Homsey for their technological assistance, particularly with my penchant for destroying computers.

Finally, a big thank you goes to my family and friends for their encouragement and support throughout this project, and for always being there to help me pull through; and to my pup Zach for keeping an eye on me throughout.

Notes on Style

The material presented in this thesis consists of a combination of published and submitted works, and traditional chapters. Where material has been published or submitted, the contents of these manuscripts have been included within the chapters and formatted accordingly. Final published and submitted versions have been included within the appendices

Some of the work presented here relates to the generation and bioinformatic analysis of large datasets. Due to the formatting of some of this data, pertinent information has been included within the chapters and appendices, and additional material has been made available in electronic form in the attached CD. These materials are as follows:

- Appendix I
  - Table S4.5
  - Table S6.1
- Appendix IV
  - Electronic version of the ImageJ macro used for densitometry analyses
- Appendix V
  - Electronic version of the python code for the ERV detection pipeline and the associated configuration file
- Appendix VI
  - Electronic version of published material including additional files
- Appendix VII
  - Electronic version of submitted material

.

Publications and Conference Presentations

Acknowledgment of contribution to the research work and/or authorship

This thesis includes one original paper published in a peer-reviewed journal, and one manuscript that has been submitted. Outlined below are the published and submitted articles associated with this thesis, conference presentations resulting from the work presented herein, and other research contributions, related and unrelated, that took place over the duration of my candidature.

The ideas, development and writing up of all the papers in this thesis were principally my responsibility, working within the Faculty of Veterinary Science under the supervision of Dr Jaime Gongora.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. For each of these papers, signed author contribution statements have been submitted to the Faculty of Veterinary Science. For brevity, unsigned copies have been provided herein.

Manuscripts included in this thesis

Chapter 2

Chong, A. Y. Y., Atkinson, S. J., Isberg, S. & Gongora, J. 2012. Strong purifying selection in endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia. *Mobile DNA*, 3, 20.

Chapter 3

Chong, A. Y., Kjeldsen, S. & Gongora, J. submitted. Lineage specific evolution of endogenous retroviruses in the Freshwater Crocodile (*Crocodylus johnstoni*). Submitted to *Journal of Herpetology* in June 2013.

Confirmation of co-authorship of published work

Chapter 2:

Chong, A. Y. Y., Atkinson, S. J., Isberg, S. & Gongora, J. 2012. Strong purifying selection in endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia. *Mobile DNA,* 3**,** 20.

Author Contributions

The candidate, Amanda Y. Chong, contributed to the design of the study, generated the data, performed all analyses, interpreted the data, and wrote the manuscript.

Ms Sarah J. Atkinson contributed some of the sequence data, performed preliminary analyses, and assisted with early drafts of the manuscript.

Dr Sally Isberg collected the samples used in this study and provided advice on the final manuscript.

Dr Jaime Gongora conceived, guided the design of the study, and provided advice regarding analysis, data interpretation, and finalising manuscript.

All authors read and approved the final manuscript.

I, as a Co-Author, endorse that this level of contribution by myself and the candidate indicated above is appropriate.

Name:

Signature: Date:

Confirmation of co-authorship of submitted work

Chapter 3:

Chong, A. Y., Kjeldsen, S. & Gongora, J. submitted. Lineage specific evolution of endogenous retroviruses in the Freshwater Crocodile (*Crocodylus johnstoni*). Submitted to *Journal of Herpetology* in June 2013.

Author Contributions

The candidate, Amanda Y. Chong, assisted with the design of the study, generated the data presented in this paper, performed all analyses, interpreted the data, and wrote the manuscript.

Ms Shannon Kjeldsen contributed to the generation of data, carried out preliminary analyses, and assisted with early drafts of the manuscript.

Dr Jaime Gongora conceived, guided the design of the study, and provided advice regarding analysis, data interpretation, and finalising manuscript.

All authors read and approved the final manuscript.

I, as a Co-Author, endorse that this level of contribution by myself and the candidate indicated above is appropriate.

Name:

Signature: Date:

Conference presentations

Atkinson, S. J., **Chong, A. Y.**, Isberg, S., Gongora, J. Characterisation of endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*). (Poster) Genetics Society of Australasia Conference 2010, Canberra, ACT, Australia, 4-8 July, 2010.

Chong, A. Y., Isberg, S., Melville, L., Gongora, J. Characterisation of endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*). (Oral presentation) Faculty of Veterinary Science Postgraduate Conference, Camperdown, NSW, Australia, November 1-2, 2010.

Chong, A. Y., Isberg, S., Melville, L., Ray, D. A., Glenn, T. C., Gongora, J. Exploring endogenous retroviruses in the crocodilian genome. (Oral presentation) Faculty of Veterinary Science Postgraduate Conference 2011, Camden, NSW, Australia, November 2-3, 2011.

Chong, A. Y., Isberg, S., Melville, L., Ray, D. A., Glenn, T. C., Gongora, J. Exploring endogenous retroviruses in the crocodilian genome. (Poster) Genomic Impact of Eukaryotic Transposable Elements, Pacific Grove, CA, USA, February 24-28, 2012.

Chong, A. Y., Atkinson, S. J., Isberg, S., Melville, L., Gongora, J. Diversity of endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia. (Oral presentation) 4th International Workshop on Crocodilian Genetics and Genomics, Darwin, NT, Australia, May 16-18, 2012.

Chong, A. Y., Isberg, S., Melville, L., Ray, D. A., Glenn, T. C., Peterson, D. G., Shan X., Gongora, J. Genome-wide identification of endogenous retroviruses from a crocodilian genome. (Oral presentation) 4th International Workshop on Crocodilian Genetics and Genomics, Darwin, NT, Australia, May 16-18, 2012.

Chong, A. Y., Atkinson, S. J., Isberg, S., Melville, L., Gongora, J. Diversity of endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia. (Poster) Annual Meeting of the Society for Molecular Biology and Evolution, Dublin, Ireland, June 23-26, 2012.

Chong, A. Y., Isberg, S., Melville, L., Ray, D. A., Glenn, T. C., Peterson, D. G., Shan X., Gongora, J. Genome-wide identification of endogenous retroviruses from a crocodilian genome. (Poster and oral presentation) 33rd Conference of the International Society of Animal Genetics, Cairns, QLD, Australia, July 15-20, 2012.

Chong, A. Y., Ray, D. A., Gongora, J. *De novo* detection of endogenous retroviruses from crocodilian genome sequencing data. (Oral presentation) Faculty of Veterinary

Science Postgraduate Conference, Camperdown, NSW, Australia, November 7-8, 2012.

Other research contributions

Gongora, J., Biondo, C., Cooper, J. D., Taber, A., Keuroghlian, A., Altrichter, M., do Nascimento, F. F., **Chong, A. Y.**, Miyaki, C. Y. & Bodmer, R. 2011. Revisiting the species status of *Pecari maximus* van Roosmalen et al., 2007 (Mammalia) from the Brazilian Amazon. *Bonn zoological Bulletin*, 60, 95-101.

Gongora, J., Swan, A. B., **Chong, A. Y.**, Ho, S. Y., Damayanti, C. S., Kolomyjec, S., Grant, T., Miller, E., Blair, D. & Furlan, E. 2012. Genetic structure and phylogeography of platypuses revealed by mitochondrial DNA. *Journal of Zoology*, 286, 110-119

Luck, N. L., Thomas, K. C., Morin-Adeline, V. E., Barwick, S., **Chong, A. Y.**, Carpenter, E. L., Wan, L., Willet, C. E., Langford-Salisbury, S. M., Abdelsayd, M., Ang, R. A., Atkinson, S. J., Barcelo, F. G., Booth, M. E., Bradbury, E. J., Branighan, T. L., Brown, J., Castillo, L. E., Chandler, N. D., Chong, J. Y., Collits, K. J., Cook, E., Cruz, R. E., Farrugia, C. A., Fletcher, J. L., Fletcher, S., Gamaliel, N. S., Gurr, J. F., Hallett, N. J., Hargreaves, G., Harris, T., Hollings, S., Hopcroft, R. L., Johinke, D., Kern, P. L., Kiddell, J. L., Kilby, K. E., Kragic, B., Kwan, J. H., Lee, J. I., Liang, J. M., Lillie, M. C., Lui, B. C., Luk, S. W., Lun, K. H., Marshall, K. L., Marzec, J. A., Masters, K. T., Mazurkijevic, L. J., Medlock, J., Meoli, C., Morris, K. M., Noh, Y. H., Okazaki, H., Orourke, T. J., Payne, E. M., Powell, D. J., Quinlivan, A. R., Reeves, T. J., Robson, K., Robson, K. L., Royle, L. J., Stevenson, R., Sellens, T., Sun, Z., Sutton, A. L., Swan, A., Tang, J. M., Tinker, J. E., Tomlinson, S. C., Wilkin, T., Wright, A. L., Xiao, S. T., Yang, J., Yee, C., Jaratlerdsiri, W., Isberg, S. R., Miles, L., Higgins, D., Lane, A. & Gongora, J. 2012. Mitochondrial DNA analyses of the saltwater crocodile (*Crocodylus porosus*) from the Northern Territory of Australia. *Australian Journal of Zoology*, 60, 18-25.

St John, J., Braun, E., Isberg, S., Miles, L., **Chong, A.**, Gongora, J., Dalzell, P., Moran, C., Bed'Hom, B., Abzhanov, A., Burgess, S., Cooksey, A., Castoe, T., Crawford, N., Densmore, L., Drew, J., Edwards, S., Faircloth, B., Fujita, M., Greenwold, M., Hoffmann, F., Howard, J., Iguchi, T., Janes, D., Khan, S., Kohno, S., de Koning, A. J., Lance, S., McCarthy, F. & McCormack, J. 2012. Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology*, 13, 415.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALV	Avian leukaemia virus
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide
BAC	bacterial artificial chromosome
bp	bases, base pairs
CA	capsid protein
CERV	Crocodilian endogenous retrovirus, later reclassed as CrocERV
CITES	Convention on International Trade in Endangered Species
DNA	deoxyribonucleic acid
EAV	endogenous avian retrovirus
env	envelope genes
ERV	endogenous retrovirus
gag	group specific antigens
Gb	gigabase, billion bases
HERV	Human endogenous retrovirus
HIV	Human immunodeficiency virus
HTLV	Human T-cell lymphotropic virus
IAP	intracisternal A-particle
IN	integrase
indel	insertion-deletion
IUCN	International Union for Conservation of Nature
JSRV	Jaagsiekte sheep retrovirus
kb	kilobase, thousand bases
KoRV	Koala retrovirus
LINE	long interspersed element
LRT	likelihood ratio test
LTR	long terminal repeat
MA	membrane associated protein
Mb	megabase, million bases
MHC	major histocompatibility complex
MID	molecular identification tag
MLV	Murine leukaemia virus
MMTV	Mouse mammary tumour virus
MPMV	Mason-Pfizer monkey-type virus
mRNA	messenger RNA
mtDNA	mitochondrial DNA
MYA	million years ago
NC	nucleocapsid protein
NCBI	National Center for Biotechnology Information
onc	oncogene
ORF	open reading frame

PBS	primer binding site
PCR	polymerase chain reaction
pol	polymerase region
PPT	polypurine tract
PR	protease
pro	protease region
RAV	Rous-associated virus
RNA	ribonucleic acid
RSV	Rous sarcoma virus
RT	reverse transcriptase
sag	superantigen
SFV	Simian foamy virus
SINE	short interspersed element
SIV	Simian immunodeficiency virus
SNV	Spleen necrosis virus
SQL	structured query language
SU	surface unit protein
TE	transposable element
TM	trans-membrane protein
TRIM	tripartite motif containing
tRNA	transfer RNA
TSD	target site duplication
U3	untranslated 3` region
U5	untranslated 5` region
WDSV	Walleye dermal sarcoma virus
WEHV	Walleye epidermal hyperplasia virus

Chapter 1: Introduction

1.1: *Crocodylus porosus*

1.1.1: Introduction to Order Crocodylia

Crocodylia and its sister taxa Aves represent the two surviving orders of Archosauria, and form the basal taxa of Class Reptilia. There are currently 23 recognised species within the Order Crocodylia, belonging to three families: *Alligatoridae*, *Crocodylidae*, and *Gavialidae*. *Alligatoridae* consists of the genera *Alligator*, *Caiman*, *Paleosuchus* and *Melanosuchus*, *Crocodylidae* consists of *Crocodylus*, *Osteolaemus*, and *Mecistops*, and *Gavialidae* consists of *Tomistoma* and *Gavialis* (Li et al., 2007, Roos et al., 2007). The species that make up each of the genera are shown below, in Table 1.1.

Crocodilians have a wide global range with populations across much of the tropical and subtropical regions, although this distribution is more restricted at the family level. *Alligatoridae* is largely confined to the Americas, with the exception of the critically endangered *A. sinensis* which is restricted to a small region along the lower Yangtze River and surrounding provinces in China. *Crocodylidae* has a much wider geographic distribution, ranging from Australia, Melanesia, and South East Asia, to Africa, and Central America. *Gavialidae* is restricted to areas of South East Asia, India, and Nepal (Figure 1.1) (Molnar, 1993).

There is an ongoing debate as to the correct terms and groupings for species within this order (Brochu, 2003). To remove confusion for the purpose of this thesis, the term ‘crocodilian’ will be used to refer to the order Crocodylia and all species within it.

Table 1.1: Classification of crocodilian species.

Family *Alligatoridae*		
Alligator		
	A. mississippiensis	American alligator
	A. sinensis	Chinese alligator
Caiman		
	C. crocodylus	Spectacled caiman
	C. latirostris	Broad-snouted caiman
	C. yacare	Yacare caiman
Melanosuchus		
	M. niger	Black caiman
Paleosuchus		
	P. palpebrosus	Cuvier's dwarf caiman
	P. trigonatus	Schneider's dwarf caiman

Family *Crocodylidae*		
Crocodylus		
	C. acutus	American crocodile
	C. intermedius	Orinoco crocodile
	C. johnstoni	Australian freshwater crocodile
	C. mindorensis	Philippine crocodile
	C. moreletii	Morelet's crocodile
	C. niloticus	Nile crocodile
	C. novaeguineae	New Guinea crocodile
	C. palustris	Mugger crocodile
	C. porosus	Saltwater crocodile
	C. rhombifer	Cuban crocodile
	C. siamensis	Siamese crocodile
Mecistops		
	M. cataphractus	Slender-snouted crocodile
Osteolaemus		
	O. tetraspis	African dwarf crocodile

Family *Gavialidae*		
Gavialis		
	G. gangeticus	Indian gharial
Tomistoma		
	T. schlegelii	False gharial

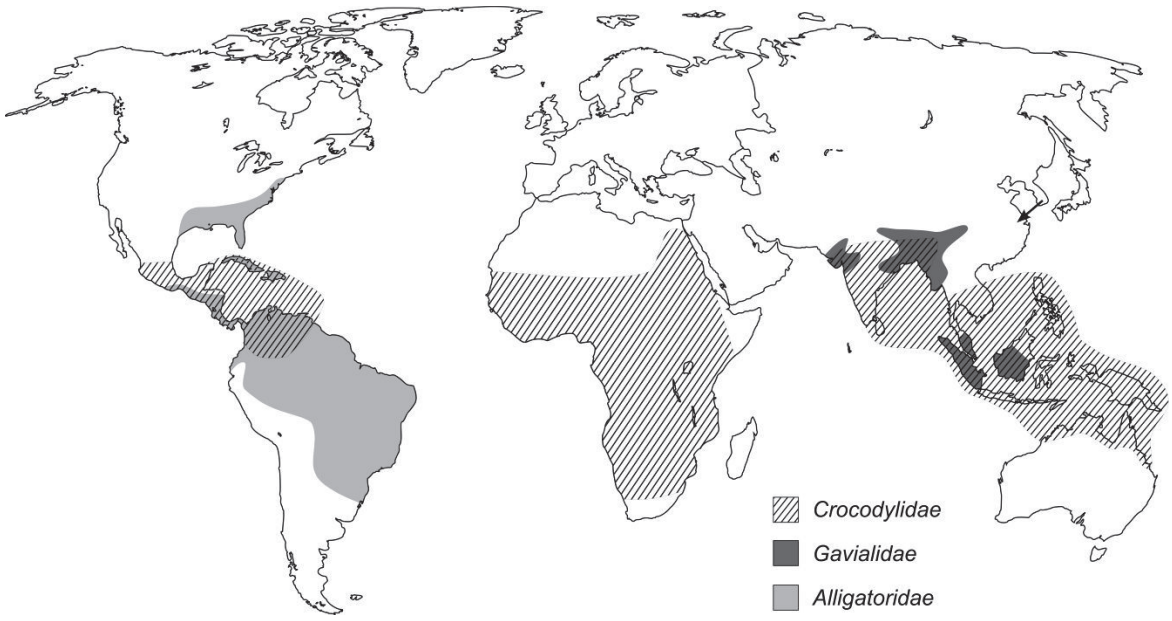


Figure 1.1: Global distribution of the extant crocodilians. Broad-scale distributions are shaded as indicated. The arrow indicates the region in China where *A. sinensis* is found. Adapted from Molnar (1993) and the IUCN Red List data for *A. sinensis* (Crocodile Specialist Group, 1996a).

1.1.2: Distinguishing between species

Crocodilians share a very similar gross morphology, and aside from the general sizes of the different species, most variation and taxonomically informative characteristics are found either in the scaling patterns of the skin, or in the cranial regions (Grigg and Gans, 1993, Molnar, 1993). Of the three crocodilian families, *Alligatoridae* tend to have broader, rounder snouts, while *Crocodylidae* tend to be more pointed. This varies from species to species within *Crocodylidae*, with some species (e.g. *C. porosus*) having broader snouts, while others (e.g. *C. johnstoni*) have much narrower snouts (Grigg and Gans, 1993). The gavialids (*G. gangeticus* and *T. schlegelii*) are unusual among *Crocodylidae* with exaggerated long, narrow snouts and it is thought that morphologically these species demonstrate a reversion to an ancestral form (Figure 1.2). This atavism is likely the cause for the traditional morphological placement of gavialids as the basal taxa of Crocodylia (Gatesy et al., 2003).

The other major distinguishing characteristics of crocodilian species are their scaling patterns. This is a particularly important feature for determining the species of origin for crocodilian skins (Brazaitis, 1987, Brazaitis, 1989). Notable features are the presence and location of osteoderms (particularly the nuchal crests), or bony plates within the scales, integumentary

sense organs, and scaling patterns (Grigg and Gans, 1993). Integumentary sense organs are present on the ventral scales of species in *Crocodylidae* and *Gavlialidae*, but are absent in *Alligatoridae*. Osteoderms are present on the ventral scales of many crocodilian species, and their pattern, shape and size are used for identification. Notably, osteoderms are absent on ventral scales of *C. porosus*. Similarly, the number, shape and arrangement of scales on the belly and head can be used to differentiate between crocodilian species (Brazaitis, 1987, Brazaitis, 1989).

1.1.3: Taxonomy of *Crocodylidae*

The evolutionary history of the Order Crocodylia remains a hotly contested subject, despite being one of the oldest groups of terrestrial vertebrates. While it is widely accepted that crocodilians and avians comprise the basal clade of modern day Reptilia, the relative positions of the crocodilian genera are still under debate. Depending on whether morphological or genetic data are used, and the types of markers employed for this purpose, *Gavialidae* is either placed basal to the two other families, or as a basal lineage within *Crocodylidae* (Densmore and Owen, 1989, Hass et al., 1992, Norell, 1989). This later classification is consistent with the genetic data to date, and there is growing evidence to support the position of *Gavialidae* as a separate family (Densmore and Owen, 1989, Densmore and White, 1991, Harshman et al., 2003, Hass et al., 1992, Janke et al., 2005, Man et al., 2011, Oaks, 2011).

Protosuchia is the earliest known crocodilian, and based on fossil records, existed during the Late Triassic period. Alligators and crocodiles are thought to have diverged during the late Cretaceous period, while the crocodile/gharial split occurred during the Early Tertiary period (Hass et al., 1992). Subsequent DNA analyses have placed the Crocodylia/Aves divergence at 254 MYA (Janke and Arnason, 1997), the divergence of alligators and crocodiles at 97–103 MYA, and the crocodile/gharial divergence at approximately 47–49 MYA (Hugall et al., 2007, Roos et al., 2007).

Within *Crocodylidae*, the order of species divergence is also contentious although a growing amount of genetic sequence data support the current taxonomy as shown in Figure 1.2. Under this taxonomic scheme, *O. tetraspis* and *M. cataphractus* are distinct from *Crocodylus*, and

are likely sister taxa to each other (Gatesy et al., 2004, McAliley et al., 2006, Oaks, 2011, Poe, 1996). *Crocodylus* is regarded as a monophyletic clade within *Crocodylidae*, although until recently, the grouping of taxa within this clade has remained unclear (Densmore and White, 1991, Man et al., 2011, Oaks, 2011).

The species within *Crocodylus* can largely be divided into three groups based on geographic locations (Meredith et al., 2011, Oaks, 2011). The Australasian group consists of *C. johnstoni*, *C. novaeguineae*, *C. mindorensis*, *C. porosus*, *C. palustris* and *C. siamensis*. The African species include two proposed species of *C. niloticus* (depicted as east and west in Figure 1.2) (Hekkala et al., 2011, Schmitz et al., 2003), and the New World species from the Americas consist of *C. rhombifer*, *C. moreletii*, *C. acutus*, and *C. intermedius*.

Genetic sequencing of mitochondrial and nucleic DNA, and the development of species or sub-group specific microsatellite markers has facilitated the identification of subspecies and cryptic species within currently defined species units, such as within *C. niloticus* (Schmitz et al., 2003) and *O. tetraspis* (Eaton et al., 2009). Genetic studies have also shown that hybridisation can occur between many of the species within *Crocodylus*, both in the wild, and in captivity (Fitzsimmons et al., 2002, Milián-García et al., 2011, Ray et al., 2004), raising concerns for the preservation of wild populations, and identification of suitable captive individuals for re-introduction into wild populations.

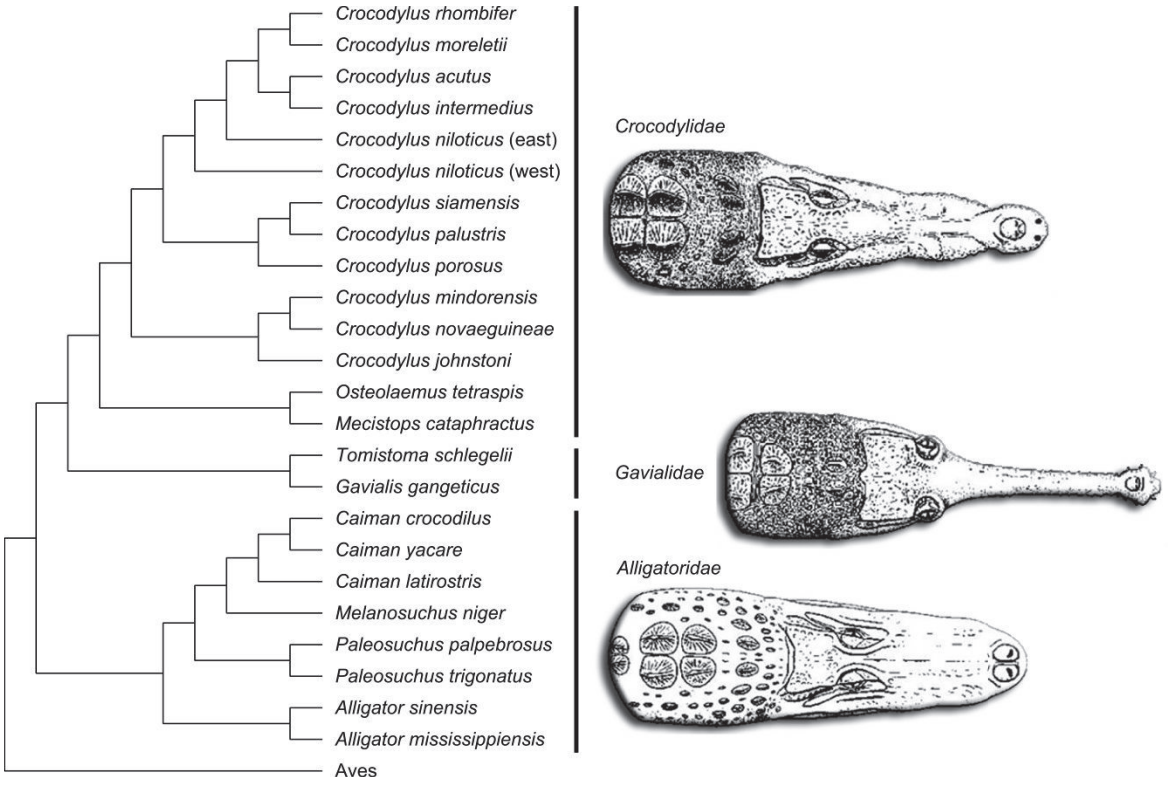


Figure 1.2: Currently accepted crocodilian taxonomy and gross cranial morphology of key crocodilian species. Adapted from Oaks et al. (2011) and Hekkala et al. (2011). Crocodilian images courtesy of Debbie McBride (Mississippi State University, Institute for Genomics, Biocomputing and Biotechnology).

1.1.4: *Crocodylus porosus*

There are two species of crocodilians found in Australia, *C. porosus*, and *C. johnstoni*. *C. porosus* has the broadest geographical distribution of all crocodilian species with populations in Australia, the Indo-Pacific region, South-East Asia, and up to India (Figure 1.1) (Crocodile Specialist Group, 1996c, Russello et al., 2007). In Australia, its range extends from Rockhampton in South-East Queensland, through to Broome in Western Australia. *C. johnstoni* is endemic to Australia, and shares a similar distribution to *C. porosus*, although its range extends further inland (Figure 1.3) (Cogger, 1992, DEC, 2009, Leach et al., 2009, QEPA, 2007).

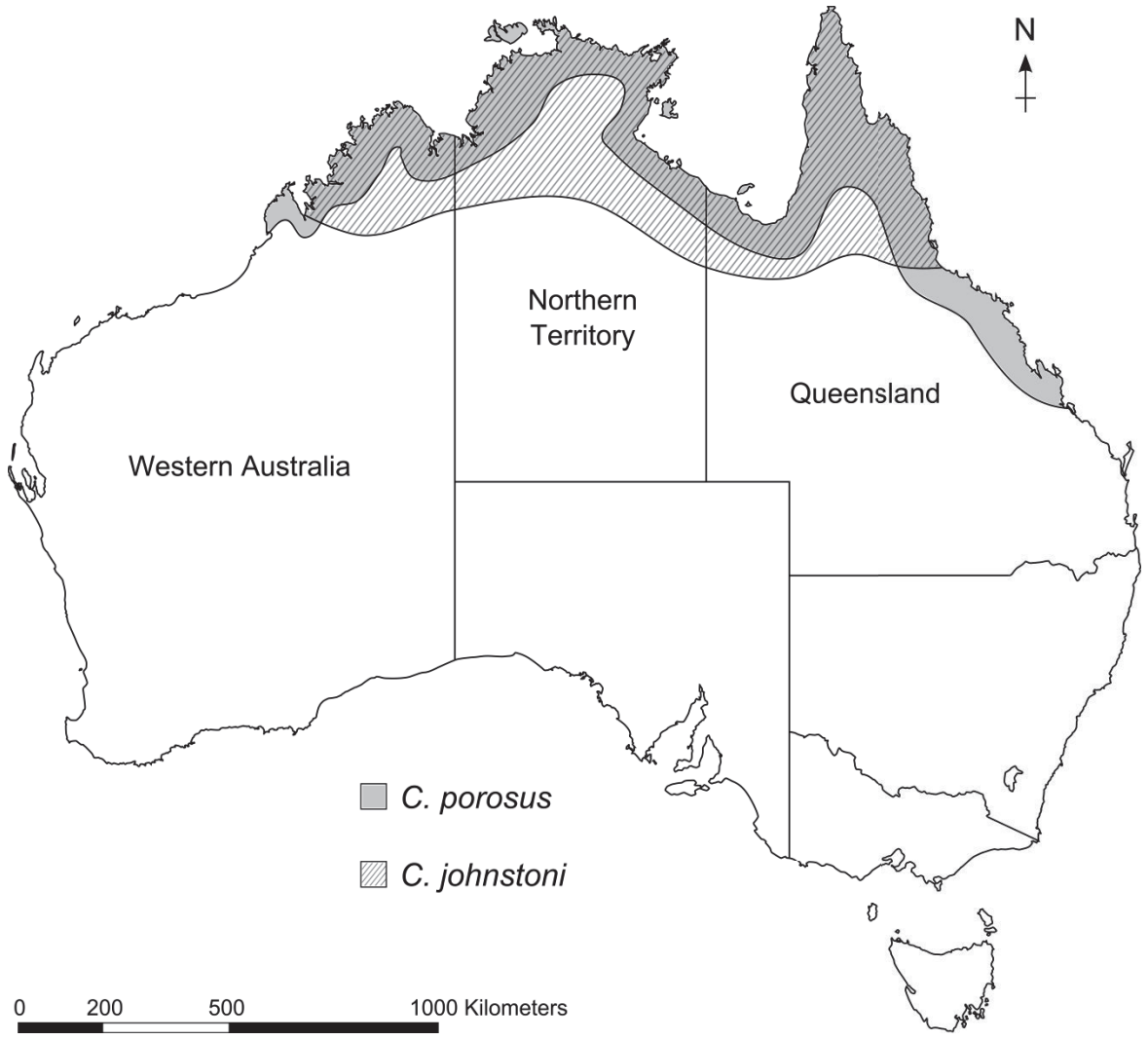


Figure 1.3: Distribution of *C. porosus* and *C. johnstoni* within Australia. Adapted from Cogger (1992).

C. porosus is one of the most adaptable crocodilian species, and can be found in a diverse range of habitats, from tidal rivers and channels on coastal floodplains, to billabongs, swamps, and some freshwater river channels (Webb et al., 1987). This species is particularly unique among crocodilians as it is able to live and breed successfully in a wide range of habitats from freshwater to highly saline conditions (Grigg et al., 1980, Taplin and Grigg, 1989). Of particular note is the capacity for these animals to travel long distances by sea, although this is noted to be atypical behaviour for the species as a whole (Allen, 1974).

The abundance and density of *C. porosus* populations vary depending on the quality of the habitats, with vegetation, rainfall, and the presence of suitable nesting areas having the

greatest effects (Fukuda et al., 2007, Webb et al., 1987). Habitat degradation has also been cited as a problem affecting crocodile abundance in some countries, although hunting has been the largest cause of population decline (Webb et al., 1987).

Commercial hunting in Australia in the mid-20th century decimated crocodile populations, leading to the implementation of a protection program for the species (Webb et al., 1987). This has included the implementation of strict licensing conditions for the regulation of ranching (egg collection and subsequent raising in captivity), removal of crocodiles from highly populated areas, and commercial hunting. Subsequently, this has ensured economic incentives for landowners to conserve important crocodile habitats on their properties and provided some compensation for resultant stock losses (DEC, 2009, Leach et al., 2009, QEPA, 2007). These programs have had a significant contribution to the recovery of the species in Australia, with many habitats now thought to be reaching carrying capacity (Leach et al., 2009).

Pre-protection estimates of populations in the Northern Territory suggested that there were between 3000 and 5000 individuals remaining, including hatchlings (Webb et al., 2000). Since then, numbers have recovered to approximately pre-hunting levels, with the population in 1998 estimated at between 70,000 and 75,000 non-hatchling individuals (Webb et al., 2000). While *C. porosus* is still considered a protected species, it is no longer listed as threatened (Crocodile Specialist Group, 1996c). Internationally, the species is listed as CITES Appendix I, with the exception of populations in Australia, Indonesia and Papua New Guinea which are listed as Appendix II (Crocodile Specialist Group, 1996c).

1.1.5: Farming of *C. porosus*

The farming of *C. porosus* is worth over AU$8 million annually in skin and meat sales. There are currently 14 commercial farms operating across northern Australia, including five in the Northern Territory, six in Queensland, and three in Western Australia (Foster, 2009). Currently, crocodile farming consists mainly of a combination of captive breeding and ranching, where eggs are harvested from wild nests (Leach et al., 2009, Webb et al., 1987). Some animals also enter the commercial system through the management programs for "problem" crocodiles. These are animals captured and removed from areas where they are

perceived to cause a danger to public safety or loss of livestock (DEC, 2009, Leach et al., 2009, QEPA, 2007, Webb et al., 1987).

The main product of the crocodile industry are the skins which account for approximately 80--90% of the market value of the animal (Shim-Prydon and Camancho, 2007). Skin production in Australia totalled about 26,990 skins in 2009, and originated from a combination of captive bred and ranched animals (Caldwell, 2011). The target market for the crocodile skins are the exclusive fashion houses of Europe, such as Hermes. These well-reputed brands produce high quality handbags and shoes. The remaining market value consists of meat sales and novelty items such as heads, skulls, teeth, feet and tail tips (Shim-Prydon and Camancho, 2007).

1.1.6: Diseases of farmed crocodiles

Crocodilians are susceptible to a number of viral, bacterial and fungal diseases. While these do not appear to pose any particular threat to wild populations, the captive production of animals may initiate or exacerbate these disease threats. From, the total number of mortalities (13.87%) on Darwin Crocodile Farm between 2005–2007, 12% were attributable to disease (Isberg et al., 2009). By far the largest cause of on-farm mortality is runtism (49%), followed by deaths from unknown aetiologies (23%). Stress-related deaths were also a considerable problem accounting for 7% of the total.

Management-related stress is presumed to be the main contributing factor for many disease outbreaks in farmed crocodilians. Animals exposed to high stress or less than optimal conditions have decreased immune responses (Morici et al., 1997) making them more susceptible to pathogens. This can lead to infection and the generation of disease by organisms that are not normally pathogenic. Disease as a result of acute stress from short term procedures such as handling and moving of animals can easily be identified and treatment obtained (Isberg et al., 2009). Chronic stress, on the other hand, is the result of prolonged or persistent exposure to external or internal stressors, and can be caused by suboptimal environmental conditions, competition for resources, or prolonged exposure to disease agents (Elsey et al., 1990, Huchzermeyer, 2003). This makes it more difficult to identify the resulting disease incidences or remedy the cause.

A number of viral diseases have been identified in crocodilians, including pox viruses, adenoviruses, herpesviruses, flaviviruses and influenza (Huchzermeyer, 2003, Melville et al., 2012). Various strains of pox viruses have been isolated from crocodilians, and are frequently associated with the development of lesions, both internally and externally (Buenviaje et al., 1998, Wellehan and Johnson, 2005). Adenoviruses have been associated with hepatitis in infected animals, and have been implicated in the incidence of runting in farmed *C. niloticus* (Buenviaje et al., 1994, Revol, 1995, Wellehan and Johnson, 2005). Herpesviruses have been identified in *C. porosus*, *C. johnstoni*, and *A. mississippiensis*, and are associated with a variety of clinical symptoms including lesions and ulceration of mucosal tissues (Govett et al., 2005, Melville et al., 2012). West Nile virus has been associated with disease in captive *A. mississippiensis*, and has been isolated from *C. niloticus* with no apparent clinical symptoms. Influenza viruses have also been isolated from captive individuals (Huchzermeyer, 2003, Miller et al., 2003, Wellehan and Johnson, 2005).

Bacterial pathogens are also a significant cause of disease in farmed crocodiles. These diseases include bacterial septicaemia, which has been attributed to a variety of bacterial pathogens, and chlamydiosis (Ladds and Sims, 1990, Revol, 1995). Septicaemia, in particular, is caused by a range of bacteria commonly found in the environment, such as *Providencia rettgeri*, *Morganella morganii* and *Edwarsiella tarda* (Foggin, 1992, Ladds et al., 1996). These potentially pathogenic bacteria have been isolated from apparently healthy individuals, suggesting that other external stressors may predispose animals to opportunistic infections (Foggin, 1992). Given that outbreaks of bacterial disease occur after management-related stress (Foggin, 1992), or exposure to suboptimal environmental conditions (Ladds et al., 1996), it is probable that a depressed immune response as a result of these stressors is a major contributing factor. Clostridial bacteria also have been associated with diseases in captive animals (Buenviaje et al., 1997, Buenviaje et al., 1998).

Likewise, a number of parasites such as Coccidia and Metaprotozoa have been identified in crocodilians (Ladds and Sims, 1990). Coccidial infections have been observed in a number of diseased animals in New Guinea, and were determined to be the primary cause in a large proportion of cases reported (Ladds and Sims, 1990). Metaprotozoan parasites such as Helminths and Pentastomids have also been observed though these have usually occurred in conjunction with other pathogens, and as such, the primary causative factor for disease is unclear (Ladds and Sims, 1990).

Opportunistic fungal infections have also been noted, and are likely to be a result of poor management and high stress leading to immunosuppression in the host animal (Buenviaje et al., 1994). In particular, *Mycobacteria* and *Dermatophilus sp.* have been associated with lesions in a number of species. Systemic fungal infections appear to be associated with high stress conditions such as low environmental temperatures (Buenviaje et al., 1994).

Non-transmissible diseases in captive crocodilians include congenital deformities, miscellaneous management-related disorders and those with no known aetiology. Common congenital deformities include tail deformities, absence of tails, cleft palate, weak yolk scar sutures, jaw deformities and syndactyly or polydactyly (Huchzermeyer, 2003, Isberg et al., 2009). Management-related disorders can include the effects of inappropriate temperatures (Turton et al., 1997), stocking densities (Elsey et al., 1990, Morpurgo et al., 1992), and nutritional deficiencies (Buenviaje et al., 1994, Huchzermeyer, 2003). These disorders, whilst not necessarily causing disease themselves, can predispose the animals to stress allowing the above-mentioned diseases to occur. Of particular concern are the deaths with no known aetiology (23%). Generally, these animals return negative bacterial pathology findings lending to the theories that either the animal has succumbed to prolonged stress or an underlying viral infection may be the cause (Isberg et al., 2009).

Runting is a condition characterised by continued retardation of growth and development, where affected individuals fail to thrive. This condition has only been noted in captive hatchlings (Buenviaje et al., 1994, Huchzermeyer, 2003, Isberg et al., 2009, Ladds and Sims, 1990), and it is unclear whether it occurs in wild populations to the same degree. As such, further investigation is needed to determine if the additional environmental control of artificial incubation allows these animals to fully develop and hatch, which would not occur in wild populations, or whether affected hatchings just do not survive. Runting in farmed *C. porosus* hatchlings is the major cause of juvenile mortality, making up approximately 50% of total mortalities (Isberg et al., 2009). This is similar to reported rates in *C. niloticus*, where runting is estimated to cause deaths in 5–15% of hatchlings, decreasing to 0–2% in animals 0.7 m and above (Revol, 1995).

Recent studies to elucidate the cause of this disease in *C. porosus* hatchlings have indicated that affected animals show increased levels of corticosterone, a reduction in the size and prevalence of lymphoid tissue, and increased vacuolation of adrenocortical cells, indicative of

chronic stress (Isberg et al., 2009). Stress is also suggested to play a role in this condition in *C. niloticus* (Huchzermeyer, 2003). However, it is unlikely that stress is the sole causative factor, as animals do not recover after removal of stressors or mediation of the environment. Bacteriological and parasitological factors are also unlikely to be responsible, although it has been suggested that immune suppression from chronic stress may make runts more susceptible to disease (Huchzermeyer, 2003, Isberg et al., 2009, Morici et al., 1997).

While early studies have attributed runting as a failure to adapt to captive conditions, it has more recently been acknowledged that there may be a genetic component to this disease (Huchzermeyer, 2003, Isberg et al., 2009, Peucker and Mayer, 1995). In particular, Isberg et al. (2009) noted that area and clutch related factors may have contributed to the incidence of runting among hatchlings. Analysis of clutches of known parentage also found the effect of parental pairs to be a significant factor, increasing the interest in identifying the underlying genetic causes.

1.1.7: Genetic studies of *C. porosus*

Due to its broad geographic range and economic significance, *C. porosus* is one of the better studied crocodilian species, both in terms of genetic diversity as well as phylogeographically. Analysis of *C. porosus* mitochondrial DNA (mtDNA) haplotypes across the range of this species has demonstrated low levels of diversity and a lack of deep phylogeographic structures, although it has been noted that mtDNA haplotypes do appear to show some geographic localisation (Gratten, 2003, Luck et al., 2012, Russello et al., 2007). This lack of population structure of *C. porosus* in the Northern Territory, Australia is further corroborated by studies of the major histocompatibility complex (MHC). Additionally, these data suggest that at these loci, levels of genetic diversity are comparable with that of other reptilian species (Jaratlerdsiri et al., 2012).

The development of the crocodile farming industry in Australia has also resulted in increased research into desirable traits for implementation of breeding programs. This research has included the development of microsatellite markers for use in parentage testing for farm bred crocodiles (Isberg et al., 2004a) and the development of a genetic improvement program (Isberg et al., 2004b). This program includes the development of heritability measures and

quantitative analyses of economically important traits, such as reproductive success, growth rates, juvenile survival, and skin traits (Isberg et al., 2009, Isberg et al., 2005a, Isberg et al., 2005b, Isberg et al., 2006a, Isberg et al., 2006b, Miles et al., 2009, Miles et al., 2010). Despite this, our understanding of the underlying mechanics of the crocodile genome is still rudimentary, with limited knowledge of the genes responsible for these traits.

1.2: The *Retroviridae*

1.2.1: Retroviruses

Retroviruses (family *Retroviridae*) comprise a large and diverse family of single-stranded, enveloped RNA viruses (Bishop, 1978, Coffin, 1992). They require reverse transcription and a DNA intermediate for viral replication, making them unique among vertebrate viruses (Bishop, 1978). Other similar viruses include the *Ty-Copia* (*Pseudoviridae*) and *Gypsy* transposons (*Metaviridae*) (Fauquet et al., 2005). Retroviral insertions are ubiquitous in the genomes of all vertebrate species studied to date, and may be from exogenous retroviral infections, or be inherited as endogenous retroviruses (ERVs).

Retroviral virions are spherical particles, and the virion core contains an RNA dimer as well as a collection of retroviral enzymes required for integration into the host cell. The virion envelope is made up of a combination of the host cell membrane and retroviral proteins (Burmeister, 2001, Coffin, 1992). The RNA dimer is made up of two identical copies of plus-strand RNA that comprise the retroviral genome (D'Souza and Summers, 2005, Fauquet et al., 2005). This genome ranges in size from 7 to 13 kb (thousand base pairs) in size and encodes the three major coding domains, and depending on the retroviral genus, a number of accessory domains (Fauquet et al., 2005).

1.2.2: Retroviral classification

There are currently seven recognised genera within *Retroviridae* named *Alpharetovirus*, *Betaretrovirus*, *Gammaretrovirus*, *Deltaretrovirus*, *Epsilonretrovirus*, *Lentivirus* and *Spumavirus* (Fauquet et al., 2005). Historically, retroviruses were classified into Type A, B, C and D particles by gross virion morphology such as the location and shape of the virion core and the characteristics of the surface glycoproteins, and the mode of virion assembly (Coffin, 1992). While these Type classes correspond to the *Alpha-*, *Beta-* and *Gammaretroviruses*, this system of nomenclature is incomplete, and sequence similarities and genome structure are now generally used instead (Table 1.2) (Fauquet et al., 2005).

Table 1.2: Summary of virion morphology and genomic arrangement. Adapted from Coffin (1992), Vogt (1997), Burmeister (2001), Fauquet et al. (2005), and Jern et al. (2005).

Retroviral genera	Type species	Virion morphology[a]	ERV Class[b]	Genome arrangement
Alpharetrovirus	ALV	Avian Type C Spherical, centred core Core is assembled in the plasma membrane		Simple
Betaretrovirus	MMTV[c]	Type B Spherical, eccentric core	Class II	Simple
	MPMV	Type D Rod shaped core Immature Type A particles are assembled in the cytoplasm before budding as mature Type B or Type D particles		
Gammaretrovirus	MLV	Mammalian Type C Spherical, centred core Core is assembled in the plasma membrane	Class I	Simple
Deltaretrovirus	HTLV	Spherical, centred core		Complex
Epsilonretrovirus	WDSV	Unknown		Complex
Lentivirus	HIV-1	Rod shaped or Conical core		Complex
Spumavirus	SFV	Spherical, centred core	Class III	Complex

[a] Morphology types are based on early electron microscopy studies.
[b] ERV classification will be discussed in Section 1.3.2.
[c] MMTV encodes an additional accessory gene, *sag*, which allows the virus to vary interactions with T-cells (Coffin et al., 1997).

1.2.3: The retroviral genome

The basic retroviral genome consists of three coding gene regions flanked by 5` and 3` long terminal repeats (LTRs). Immediately outside this, within the host genomic DNA, are short 4-6 nucleotide repeats, or target site duplications (TSDs) that are generated during retroviral integration (Vogt, 1997). The major gene regions are the group specific antigens (*gag*), the protease-reverse transcriptase region (*pro-pol*), and the envelope genes (*env*) (Figure 1.4). These regions generally encode five structural proteins as well as the retroviral enzymes,

although the size and number of these vary between retroviral genera and occasionally between viral strains (Bishop, 1978).

Proteins encoded by the *gag* domain encapsulate the core of the retrovirus. Within this shell is the retroviral genome, reverse transcriptase, and virus encoded RNA binding proteins (Bishop, 1978). These proteins are translated as large polyproteins and are cleaved into the smaller 'mature' proteins by the retroviral protease, which is also encoded within these structures (Katz and Skalka, 1994, Temin, 1992).

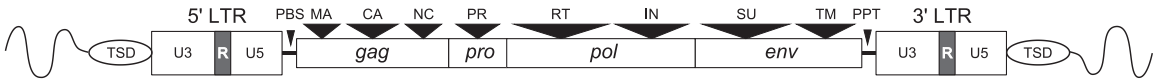


Figure 1.4: Schematic representation of the basic retroviral proviral genome. Boxes indicate retroviral domains, shaded triangles represent the retroviral proteins and promoter regions, and ovals indicate the location of the TSDs within the host genomic DNA.

The retroviral LTRs are made up of unique U3 and U5 regions separated by an R segment that is repeated at each end of the genomic RNA (Coffin, 1992). The R region is made up of a series of short, direct repeats at either end of the proviral genome, and ensures the correct ordering of the viral DNA during reverse transcription (Coffin, 1992). The 5` LTR contains an untranslated leader region that encodes the signal for packaging of the RNA genome (Coffin, 1992).

The binding sites for different cellular transcription factors are encoded by short sequences immediately between the LTRs and the retroviral coding domains. These regions enhance and promote proviral transcription, and include the primer binding site (PBS) and the polypurine tract (PPT). The PBS is a sequence of 18 bases at the 5` end of the retroviral genome that is complementary to the 3` end of the tRNA primer used to initiate synthesis of the viral DNA. At the 3` end of the genome, the PPT is an AG rich sequence that acts as a primer for the synthesis of plus strand viral DNA during reverse transcription (Coffin, 1992).

The *gag* region is the first of the retroviral gene coding domains, and encodes the retroviral structural proteins. These are the membrane associated proteins (MA), nucleocapsid proteins

(NC), and the capsid proteins (CA) (Burmeister, 2001, Coffin, 1992). The MA proteins create the inner shell of the mature virion envelope, and interact with the host cell membrane during retroviral budding. CA proteins form the shell of the virion core, and the NC proteins interact with the RNA duplex that makes up the retroviral genome (Coffin, 1992).

The retroviral enzymes are encoded by the *pro-pol* gene region. These include the retroviral protease (PR), reverse transcriptase (RT), and integrase (IN). The protease enzyme is responsible for cleaving the retroviral polyproteins to form the smaller mature proteins (Coffin, 1992). Reverse transcriptase has three functions. It first copies the plus strand RNA to create a minus strand DNA, then removes the RNA template before synthesising the plus strand DNA (Katz and Skalka, 1994). Finally, the retroviral integrase is responsible for integration of retroviral DNA into the host genome by trimming the 3` end of the retroviral provirus and the target DNA followed by ligation of the two genomes (Coffin, 1992).

The *env* domain encodes two proteins that are responsible for interaction with, and entry into, the host cell. The surface unit proteins (SU) interact with host cell receptors, while the trans-membrane proteins (TM) interact with the cell membrane, initiating viral entry into the cell (Benit et al., 2001, Coffin, 1992). Unlike the proteins encoded by the *gag* and *pro-pol* domains, these proteins are processed by cellular enzymes, and are usually translated as a separate polyprotein (Coffin, 1992).

Many retroviruses also contain additional accessory genes (Figure 1.5). These genes may alter transcription, regulate splicing of the retroviral transcripts, or otherwise affect the infectivity or behaviour of the retrovirus (Burmeister, 2001). In the case of *Spumaviruses* and *Epsilonretroviruses*, the functions of many of these accessory genes and additional open reading frames (ORFs) are still unknown (Vogt, 1997). An addition to this, some retroviruses have been found to contain oncogenes (*onc*) in place of viral domains. These ORFs generally contain DNA copies of the exons of cellular genes, or mRNA incorporated by recombination during retroviral integration. The subsequent deletion of coding domains leaves the resulting chimeric virus dependent on the presence and activity of helper viruses within the cell (Coffin, 1992).

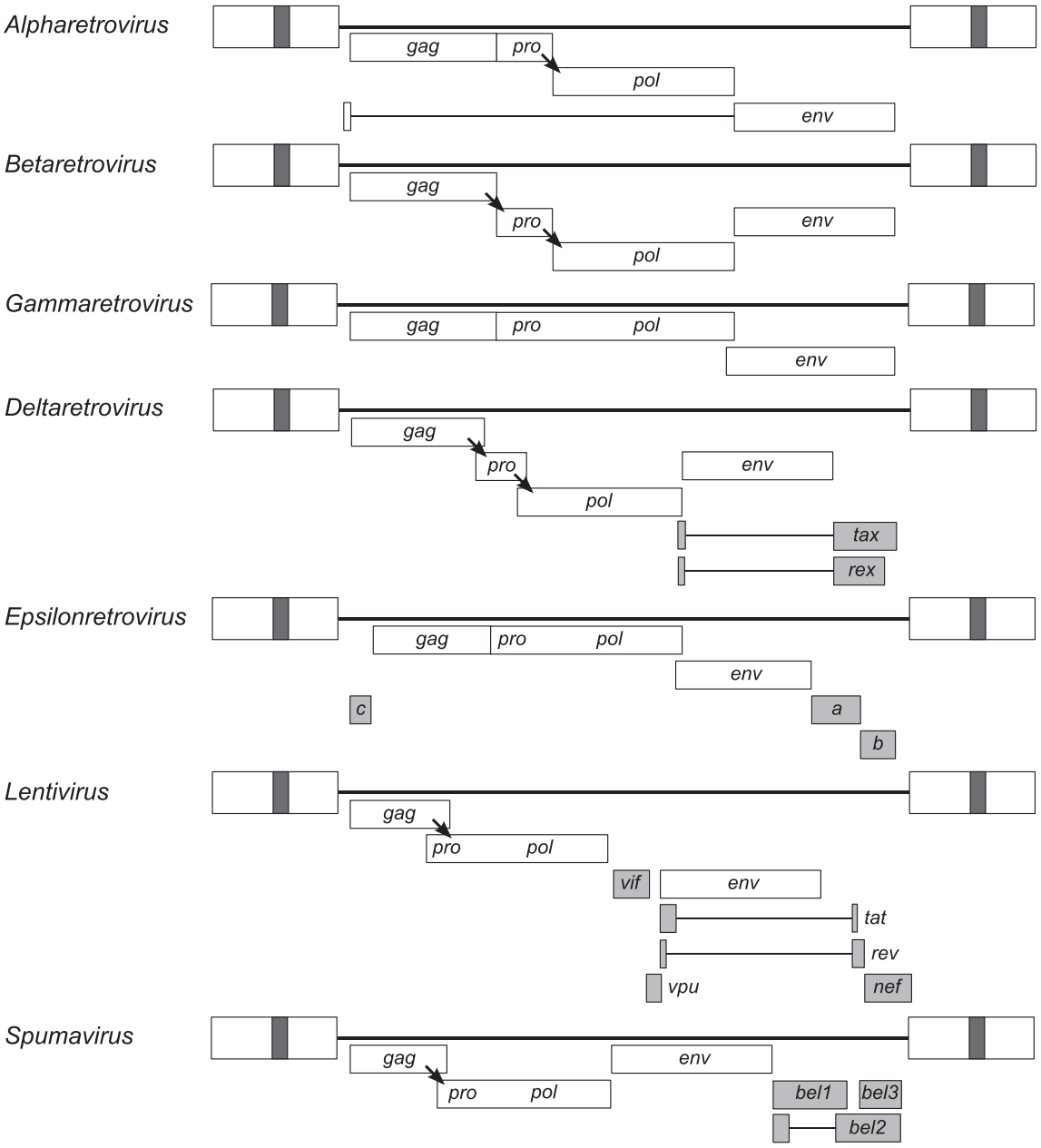


Figure 1.5: Schematic representation of the general retroviral genomes for each genus. White labelled boxes indicate the ORFs for each of the retroviral domains. Light grey boxes represent accessory genes. Boxes connected by lines indicate that alternate splicing may be required for transcription of these genes. Arrows indicate instances where ribosomal frameshifting is utilised for translation. Adapted from Coffin (1992) and Vogt (1997).

1.2.4: Retroviral integration and replication

The RNA to DNA transformation of cells by retroviruses was first proposed by Howard Temin (Temin, 1976). The process of retroviral integration begins when the virion attaches to host cell receptors and the retroviral RNA genome is released into the cell. This is then reverse transcribed to form the viral DNA. Viral DNA is then integrated into the cell genome in a stable manner, forming the provirus which can then be transcribed (Bishop, 1978, Varmus and Brown, 1989). The resulting viral RNA can reinfect the host cell or be processed to form infections viral particles that can then infect new host cells and host organisms (Bishop, 1978, Varmus and Brown, 1989).

Viral entry into the host cell is enabled by the attachment of the virion to the transmembrane receptors of a host cell. This initiates the release of the virion core into the cell followed by creation of linear double stranded viral DNA by reverse transcriptase prior to integration into the host genome (Bishop, 1978, Craigie, 2002, Varmus and Brown, 1989). The mechanisms which release the retroviral particles into the cell are still unknown (Craigie, 2002). Integration of the viral DNA into the host cell DNA is mediated by the retroviral integrase (Figure 1.6) (Katz and Skalka, 1990).

Once integrated into the genome, the retroviral provirus will be replicated along with host cell DNA during gene expression and cell division, effectively ensuring inheritance by all daughter cells of the infected host cell (Varmus and Brown, 1989). There is evidence to suggest that unintegrated viral DNA can also be used for transcription of viral RNA, but at a much lower efficiency than integrated DNA (Iyer et al., 2009, Wu, 2004, Wu, 2008). If this integration event occurs in a germline cell, the provirus may be passed on to the offspring of the infected host organism in a typical Mendelian pattern of inheritance. Proviruses inherited in this manner are termed endogenous retroviruses (ERVs) and are discussed in Section 1.3.

Transcription of viral RNA is regulated by host transcription factors and utilises the host RNA polymerase enzyme, RNA polymerase II (Bishop, 1978, Katz and Skalka, 1990). From here, some of the RNA will be processed into mRNA and translated to form the retroviral protein structures while other full length transcripts are assembled to form the RNA duplex that is packaged into the new virion (Bishop, 1978, Craigie, 2002, Varmus and Brown, 1989).

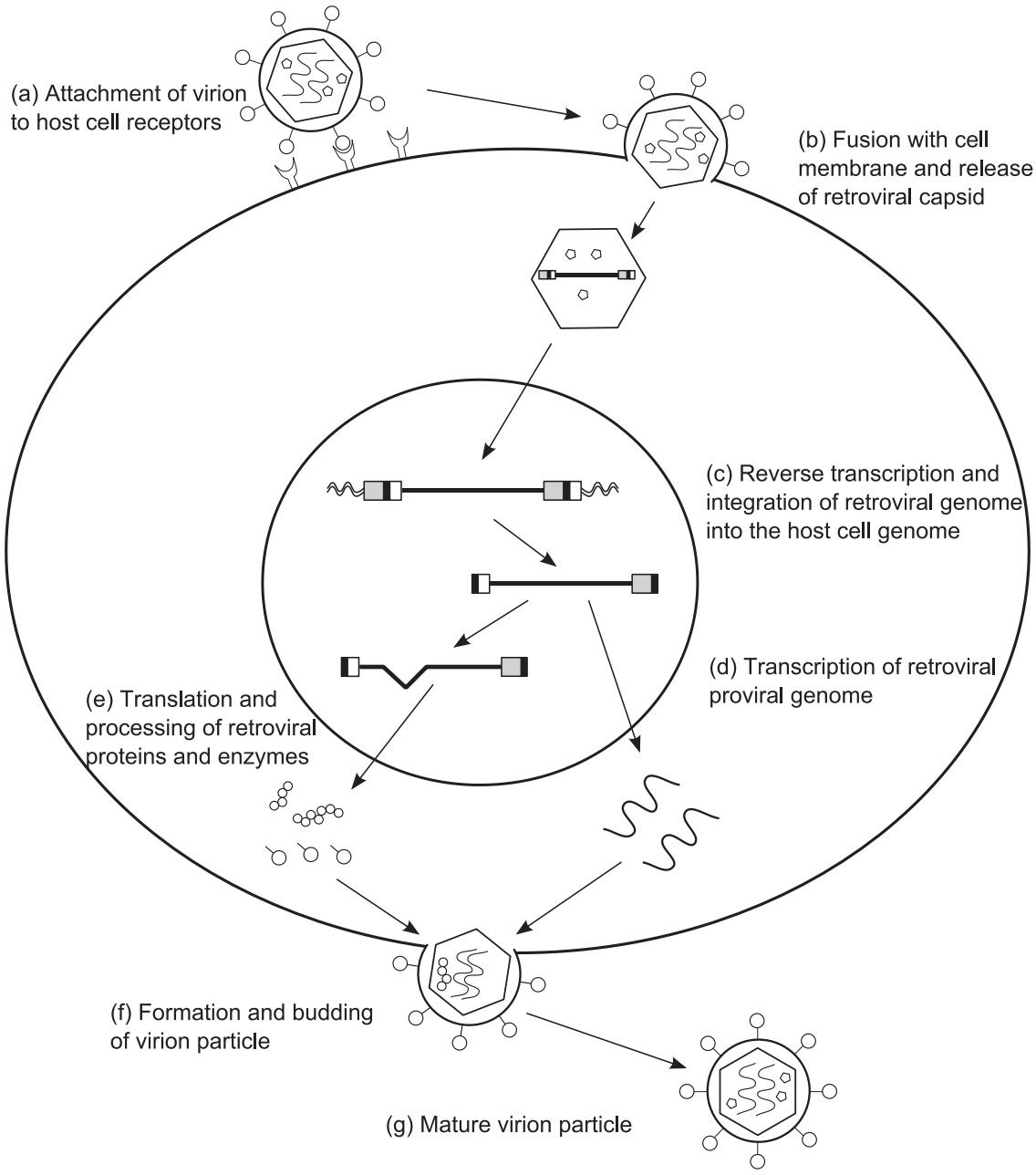


Figure 1.6: Generalised replication cycle of a retrovirus from infection of the host cell to production of the mature virion. Adapted from Vogt (1997).

Retroviral *gag*, *pro* and *pol* polyproteins are generally translated by by-passing encoded stop codons through ribosomal frameshifts, or by read-through mechanisms at the boundaries of these gene regions (Figure 1.5) (Katz and Skalka, 1994). Retroviral particle formation appears to depend on the correct ratio of spliced and un-spliced RNA, as well as cleaved and uncleaved polyprotien structures. Cleavage of these polyproteins is done by the retroviral protease that is encoded by these same polyproteins (Katz and Skalka, 1994).

1.2.5: Generation of retroviral sequence diversity

Selection pressures and host-pathogen interactions are the primary driving factors for retroviral evolution and mutation. Retroviruses must evolve to evade the host immune system while retaining the ability to interact with host cellular receptors and replicate to produce infectious virions (Coffin et al., 1997, Katz and Skalka, 1990). This in turn leads to the evolution of a wide range of retroviral strains that are adapted to infect a specific host range, and often, a specific tissue type (Coffin et al., 1997). This diversity may be generated through a variety of different methods, such as recombination and nucleotide misincorporation by the various polymerases involved in retroviral replication (Katz and Skalka, 1990).

Retroviruses generally have a much higher mutation rate than DNA viruses or their host organisms. This is due to a the lack of proofreading capabilities in both the retroviral and host derived reverse transcriptase enzymes (Katz and Skalka, 1994). While proviral mutation rates are dependent on the mutation rate of the surrounding DNA in the host genome (Temin, 1992), nucleotide misincorporation during synthesis of viral RNA effectively increases the mutation rate of retroviral strains.

Mutation rates also differ between the different retroviral coding regions. The *pro-pol* region is the most conserved of the three major regions due to the essential nature of the enzymes it encodes (McClure et al., 1988). On the other hand, the *env* region is the most variable of the three, due to the role of these proteins in viral-cell interactions (Benit et al., 2001, McClure et al., 1988). *Env* proteins are located on the outside of the virion particle, and thus must evolve to evade the host immune responses while maintaining the ability to interact with the host cell receptors (Benit et al., 2001, Coffin, 1992).

Classification of retroviruses at the sequence level can make use of these differences in mutation rate. For example, the conserved nature of the *pro-pol* region make it ideal for broad-scale classification or grouping of retroviruses from different species, but can limit the usefulness of this region to determine lineage differentiation within a host species. On the other hand, faster evolving regions may contain too many mutations for comparisons between species, and may be more useful for differentiation of lineages within a host.

Recombination with ERVs or other exogenous strains within a co-infected cell is another major source of retroviral diversity (Boeke and Stoye, 1997, Gifford and Tristem, 2003). This

can lead to the development of novel retroviral strains, and occasionally the reactivation of silenced or non-functional ERVs. For example, a comparison of *env* and *pol* phylogenies from HERV sequences has revealed evidence of recombination events in both endogenous and infectious retroviruses (Benit et al., 2001).

1.3: Endogenous Retroviruses

1.3.1: Establishment of ERVs in the host genome

Endogenous retroviruses (ERVs) are inherited copies or remnants of exogenous retroviruses that have been integrated into a host genome and passed on to subsequent generations. Once integrated into the germline cell in a stable manner, ERVs may be inherited by offspring in a normal Mendelian pattern as mentioned previously (Figure 1.7) (Temin, 1985). ERVs in infectious or proviral forms have been found in all vertebrates studied to date, and have been identified from all retroviral genera except the *Deltaretroviruses* (Herniou et al., 1998, Katzourakis et al., 2007, Martin et al., 1999).

The endogenisation of a retrovirus is dependent on a number of factors, including low virulency in the host organism, the ability to infect germ-line cells, survival of the infected cells, and the ability of offspring to survive without loss of fitness (Blikstad et al., 2008). Despite this, newly endogenised ERVs may retain a considerable level pathogenicity (Blikstad et al., 2008), which can affect the length of time that the insertion is maintained in the genome.

ERVs may be found in the genome both as recently "endogenised" exogenous viruses or remnants of ancestral retroelements (Lower et al., 1996). Variations in pathogenicity and impacts on host survival mean that species will generally have a few groups of retroviruses that have undergone independent replication. Furthermore, recombination between the LTR regions of ERV insertions is a common occurrence. Thus, full length proviruses may be present in small numbers while solo LTRs may be present in much larger numbers (often ten to a hundred times more) (Stoye, 2001).

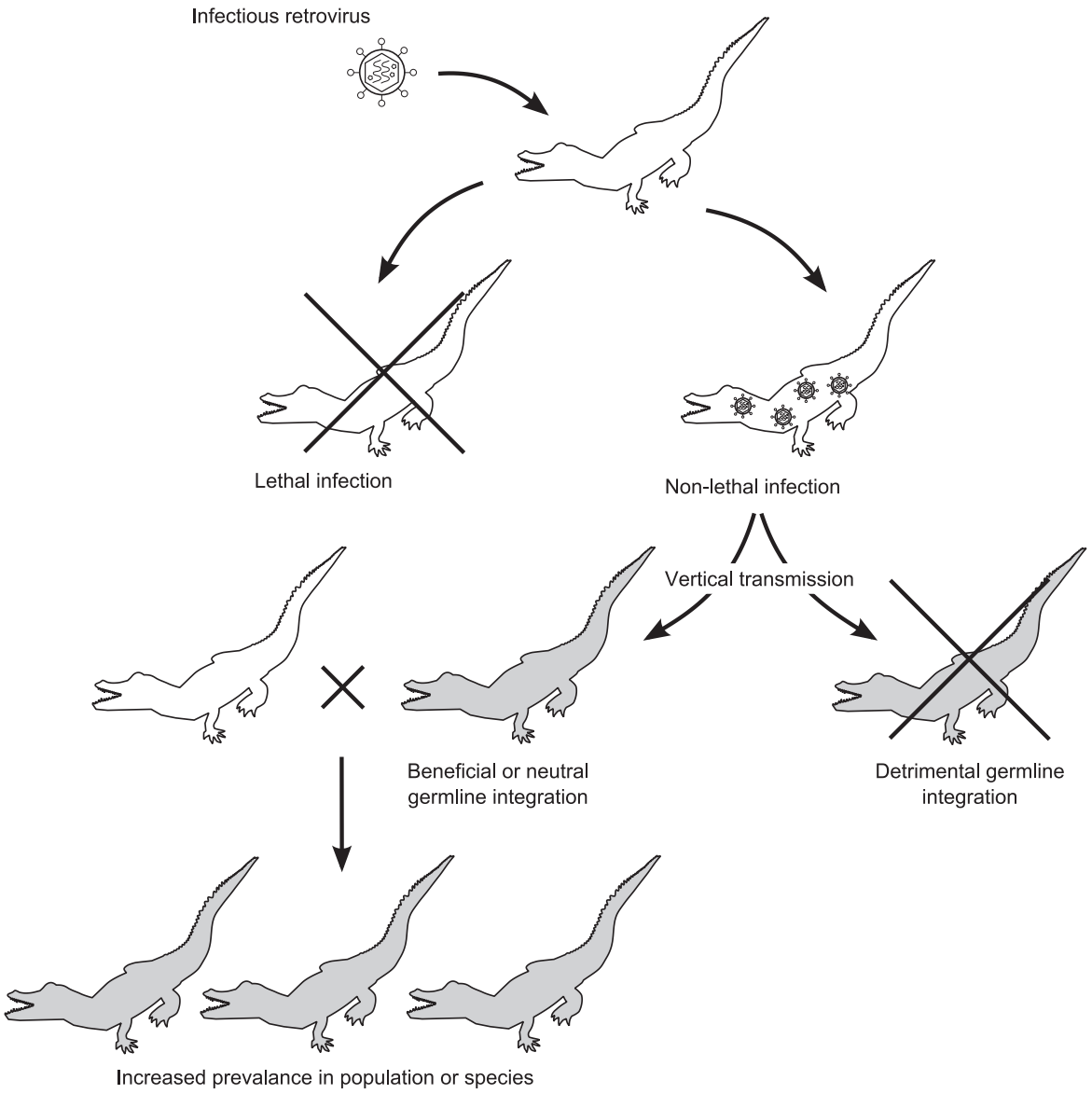


Figure 1.7: Diagrammatic representation of the endogenisation process, showing how ERVs may increase in prevalence within a population or species. Detrimental infections and integration events are selected against and removed from the gene pool.

1.3.2: ERV classification and distribution

The commonly found ERV lineages have historically been classed into three major groups, Class I, II and III, although this classification has been shown to be incomplete with the discovery of endogenous sequences related to *Lentiviruses* and *Spumaviruses* (Blomberg et al., 2009). Thus, classification is more commonly based on similarity to exogenous sequences from the seven major retroviral genera. Using this nomenclature, Class I ERVs (ERV1) are generally similar to *Gammaretroviruses* and *Epsilonretroviruses*, Class II ERVs (ERV2) are similar to the *Betaretroviruses*, and Class III ERVs (ERV3) are similar to *Spumaviruses* (Table 1.2 in Section 1.2.2) (Jern et al., 2005). Despite this, classification of ERVs is still complicated, due to the degradation of proviral sequences by selection against functional proviruses, mutation, and recombination (Barrio et al., 2011).

While individual retroviral strains may be confined to specific taxa, there is a large amount of variation in host specificity within the retroviral genera. This is echoed in the distribution of ERVs across vertebrate species. For example, the ERVs similar to *Alpharetroviruses* are confined to avian species, while *Gammaretrovirus* and *Spumavirus* related ERVs have a much wider host range, including mammals, avians, non-avian reptiles, and amphibians (Gifford and Tristem, 2003, Herniou et al., 1998, Martin et al., 1997). *Betaretrovirus* related ERVs are predominantly found in mammals and avians, with the exception of two, more divergent sequences identified from boid snakes, possibly indicating a limited host range within non-mammalian vertebrates (Gifford et al., 2005, Gifford and Tristem, 2003, Huder et al., 2002).

A number of retroviral genera have been identified only in exogenous or endogenous forms within the host species, suggesting that there are numerous factors affecting the endogenisation process. *Deltraretroviruses*, for example, have only been identified in exogenous form, whilst *Lentiviruses* have only recently been recovered in their endogenous forms and from a limited number of taxa (Gifford et al., 2008, Gilbert et al., 2009, Katzourakis et al., 2007). Exogenous *Epsilonretroviruses* have only been identified in salmonid fish (Holzschu et al., 1995), although endogenous sequences showing similarity to these exogenous viruses have been identified in other basal vertebrates, including amphibians, crocodilians, and other non-avian reptiles (Gifford and Tristem, 2003, Herniou et al., 1998, Kambol et al., 2003).

Due to this host specificity, ERVs generally co-diverge with their host species, and indeed there is a high correlation between ERV phylogenies and those of their hosts (Andersen et al., 1979, Martin et al., 1999). Despite this, instances of cross-species transmission have been documented. Studies using ERV phylogenies have detected a number of instances within the Class I ERVs. Two retroviral strains appear to have crossed from mammals into avian, the first being the progenitor of the avian Spleen necrosis virus (SNV) (Martin et al., 1999). The second of these, the Murine leukaemia virus (MLV)-related ChiRV1, appears to be the source of two instances of trans-species transmission (Borysenko et al., 2008). The Koala retrovirus (KoRV), is the result of transmission between placental mammals and marsupials, and has resulted in the development of a pathogenic ERV in *Phascolarctos cinereus* (koala) (Hanger et al., 2000, Martin et al., 1999).

1.3.3: ERV evolution in host genomes

Although most ERV insertions are inactive or unable to replicate autonomously, they still have the potential for reactivation through interactions with exogenous retroviruses (Bishop, 1978). This can occur through recombination with the infecting exogenous virus, or through the utilisation of the exogenous retroviral proteins and enzymes to compensate for inactivated or missing genes. In addition to this, recently acquired ERVs may retain the ability for reinfection, either within the host cell or to infect other cells of host organisms (Jern and Coffin, 2008).

Most species will also carry a small number of proviruses that are capable of replication (Stoye, 2001). Some ERVs may be capable of some level of expression and replication even after tens of millions of years after integration (Gifford and Tristem, 2003). Positive selection, reinfection, and complementation by other retroviruses and ERVs can all serve to extend the length of time that an ERV may remain active in the genome (Gifford and Tristem, 2003).

Once in the host genome, ERVs are able to replicate using three mechanisms: reinfection, retrotransposition, and complementation (Figure 1.8). Replication by reinfection requires that the ERV retains functionality of all promoter regions and coding domains (Bannert and Kurth, 2006, Belshaw et al., 2004, Katzourakis et al., 2005). A functional e*nv* domain allows

the movement of retroviral particles out of the host cell, and consequently, infection of other cells within the host organism, or transmission between organisms. ERV lineages with this capacity for replication would likely be under strong negative selection due to detrimental effects on the host genome (Belshaw et al., 2004), and are therefore more likely to represent recently endogenised lineages.

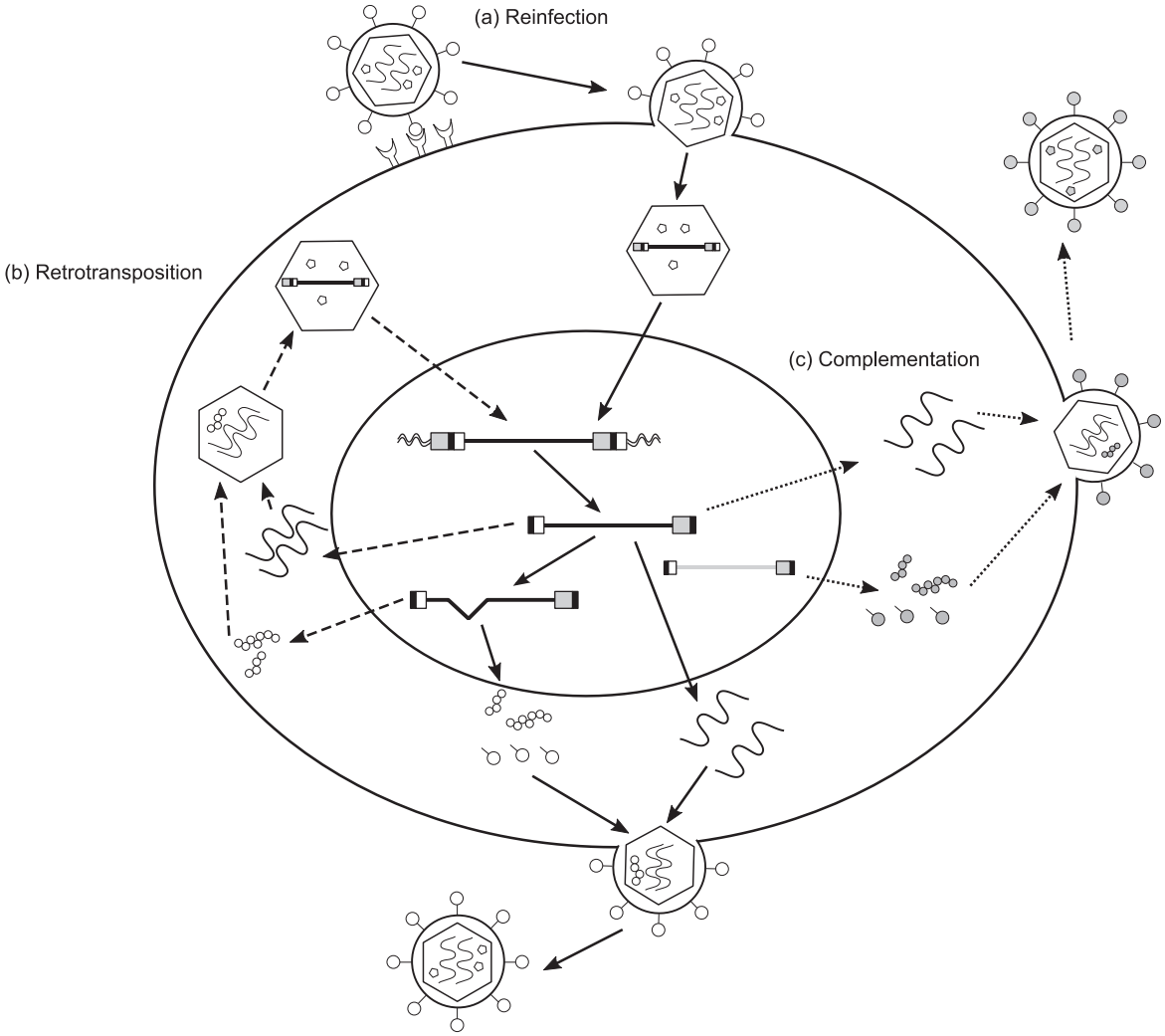


Figure 1.8: Mechanisms of ERV replication. Solid arrows show the progress of retroviral infection and reinfection (a). Dashed arrows represent retrotransposition (b), and dotted arrows represent complementation (c).

ERVs with defects in one of more coding regions may still be able to replicate within hosts through retrotransposition or complementation. Replication through retrotransposition requires functional promoters, *gag*, and *pro-pol* domains (Bannert and Kurth, 2006,

Katzourakis et al., 2005). Using this mechanism, ERVs are able replicate intracellularly, but are unable to infect other cells (Belshaw et al., 2005b). Complementation requires that only the promoter regions, such as the LTRs, PBS, and PPT, are functional (Bannert and Kurth, 2006, Gifford and Tristem, 2003). However, replication using this mechanism would appear to be rare, since it is dependent on the retroviral proteins and enzymes being supplied by other functional or partially functional viruses within the cell (Gifford and Tristem, 2003, Katzourakis et al., 2005).

ERV integration has a number of effects on the host genome, regardless of whether it is the result of recent infection or replication, or an ancestral insertion. These effects include interactions with exogenous retroviruses, disruption of gene expression and regulation, interaction with the host immune system, and provision of sites for recombination. Most of these will have detrimental effects on the host and are therefore subject to selection against these insertions (Barr et al., 2005, Gifford and Tristem, 2003). Deleterious effects of ERVs include modification of transcription or RNA processing, chromosomal rearrangements through homologous recombination, the provision of novel control sequences for cellular genes and insertional mutagenesis or activation of oncogenes (Gifford and Tristem, 2003, Katzourakis et al., 2005, Stoye, 2001, Temin, 1985).

ERVs with little or no selective effects on the host may be maintained or removed by genetic drift. However, the presence of these ERVs may still have impacts on the genome through the provision of sites for recombination between similar ERV sequences. This can lead to small scale recombination or rearrangement, such as in the case of the MHC which is known to contain a large number of ERV insertions which are closely associated with duplication breakpoints (Doxiadis et al., 2008); or large scale chromosomal rearrangements such as those that have been observed in studies comparing human and primate genomes (Hughes and Coffin, 2001).

In addition to this, a small number of insertions may have beneficial effects. These insertions are likely to be subject to positive selection, leading to an increase in prevalence. Such insertions may eventually become fixed in the genome of the host species, whereby they are found in all members of that species (Figure 1.7) (Gifford and Tristem, 2003, Jern and Coffin, 2008).

Over time, a number of ERV insertions have also been co-opted to regulate gene expression, or for additional physiological functions (Maksakova et al., 2008). Examples of this are the primate syncytin genes, which are a co-opted ERV *env* protein. Similar genes are present in Carnivora, higher ruminants, rabbits, and mice, although these insertions are present in a different genomic location and are likely to have been acquired independently (Cornelis et al., 2012, Cornelis et al., 2013, Heidmann et al., 2009, Jern and Coffin, 2008, Nakaya et al., 2013).

1.3.4: ERV transcription

Transcription of proviral ERV DNA primarily occurs in undifferentiated tissues, such as germline cells and early stage embryos (Maksakova et al., 2008). These insertions are then likely silenced in differentiated tissue types due to epigenetic mechanisms, regulator genes and control of the various gene expression pathways (Crittenden et al., 1974, Maksakova et al., 2008). Thus, re-integration into somatic cells is rare, although it may occur if transcription is activated (Maksakova et al., 2008).

Activation of ERV transcription may be induced by changing environmental signals and activation of the immune system, such as by hypomethylation as a result of stress signals (Cho et al., 2008, Perl, 2003). While this may have an impact on the host response to injury and infection, it is still unclear as to whether this reactivation is harmful or beneficial (Cho et al., 2008).

Expression of ERVs has been associated with a number of diseases in humans, such as cancer, neurodegenerative disorders, and schizophrenia (Jern and Coffin, 2008). KoRV has been linked with neoplastic disease and chlamydiosis in infected *P. cinereus* (Tarlinton et al., 2005). In these cases, ERVs have been proposed as a significant factor for the genesis of disease. However, while a correlation between ERV transcripts and presence of disease has been observed, the causative factors for these diseases are still unproven. ERVs have also been associated with autoimmune diseases, through alteration of antigen presentation and cell apoptosis pathways (Perl, 2003). However, this is also controversial as the effects of ERVs involved have not been sufficiently quantified (Gifford and Tristem, 2003).

On the other hand, ERVs may provide protection against disease caused by infection from related exogenous retroviruses (Arnaud et al., 2007a). For example, ERV derived *env* proteins can block the entry of exogenous retroviruses by competition for receptor binding sites. The presence of ERV transcripts within a cell can also interfere with the replication of exogenous retroviruses (Arnaud et al., 2008). Beneficial effects such as these have been observed in sheep, where endogenous forms of the Jaagsiekte sheep retrovirus (JSRV) have been observed to either block viral entry to the cell, or prevent the release of viral particles (Arnaud et al., 2007b, Arnaud et al., 2008). Such ERV insertions will likely come under some degree of positive selection, thus increasing the chances of becoming fixed in the host genome.

1.3.5: ERVs in non-mammalian vertebrates

It has been suggested that characterisation of the distribution and diversity of retroviruses in non-mammalian vertebrates may help increase understanding of retroviral evolution and lead to discovery of novel genes involved in pathogenesis (Hart et al., 1996, Tristem et al., 1996). Despite this, there are currently very few studies on the diversity and distribution of ERVs within these taxa, with the exception of the *Gallus gallus* (chicken) (Tristem et al., 1996).

Studies that have investigated ERVs in these taxa have uncovered many retroviral sequences that show similarity with the major exogenous retroviral genera, as well as a number of unusual elements. For example PyT2RV and PyERV, which have been sequenced from two pythonid species (*Python molurus* and *Python curtus*) with inclusion body disease, appear to be intermediates between *Betaretroviruses* and *Gammaretroviruses*. Analyses of the *pol* region have placed it with the *Betaretroviruses* but *env* analyses place it closer to *Gammaretroviruses*. As a result, these are still currently unclassed, although it is speculated that the retrovirus involved could be a recombinant virus (Huder et al., 2002). Xen1, a very large endogenous retrovirus (>10kb) that has been characterised in *Xenopus laevis* (the African clawed frog), shows similarity to the exogenous *Epsilonretroviruses* WDSV and WEHV, but contains an additional two ORFs of unknown function (Kambol et al., 2003).

Other divergent ERV sequences have been identified from reptilian species. SpeV is a divergent ERV from a tuatara (Genus *Sphenodon*, species unknown), that shows very little

similarity to known retroviruses (Tristem et al., 1995). Similarly, divergent ERVs have been identified in species of Crocodylia (Jaratlerdsiri et al., 2009, Martin et al., 2002). In both instances, the closest retroviral genus was the *Spumaviruses*, although the low sequence similarity between these and characterised *Spumaviruses* suggest that they may form separate lineages.

The reasons for this apparent divergence and diversity among non-mammalian ERVs are not well documented. The prevalence of reptilian and amphibian ERVs that are not easily classified into one of the currently recognised ERV families suggests that these ERVs may reflect the ancestral forms of current retroviral families. It has also been hypothesised that some of these instances may represent hitherto uncharacterised retroviral lineages, suggesting that reptilian, amphibian and piscine species may harbour additional retroviral diversity, and potentially, novel pathogens (Tristem et al., 1996, Tristem et al., 1995).

It is also possible that the divergence observed in reptilian and amphibian ERVs is due to co-evolution between these ERVs and the immune systems of these taxa. Although the basic immune responses are similar across vertebrate taxa, the specifics of reptile immunology are not well understood (Origgi, 2007, Warr et al., 2003). In particular, our knowledge of the adaptive immune system in reptiles, their cellular receptors, and the subsequent co-evolution of these factors and viral pathogens is limited.

1.3.6: ERVs as Transposable Elements

Transposable elements (TEs) are repetitive DNA sequences that are able to move or copy themselves within a genome. Once integrated in to a host genome, ERVs may also be considered transposable elements due to their ability to replicate within the host without the need for an exogenous phase. TEs make up a large portion of the genomes of all eukaryotes, accounting for nearly half of mammalian genomes and up to 90% of some plant genomes (Kazazian, 2004). These elements constitute a major source of variation within genomes, either through the insertions themselves or by facilitating recombination and chromosomal rearrangements (Buzdin et al., 2003, Doxiadis et al., 2008, Hughes and Coffin, 2001, Moran et al., 1999).

TEs can broadly be classed into two groups: DNA transposons, and retrotransposons (Craig, 2002, Kazazian, 2004). DNA transposons move around the genome through a cut-paste mechanism whereby the transposon sequence is excised from the genome and re-integrated into another location. Retrotransposons, on the other hand, use a copy-paste mechanism involving transcription into the RNA element, followed by reverse transcription and integration into a new site within the genome (Figure 1.9) (Craig, 2002).

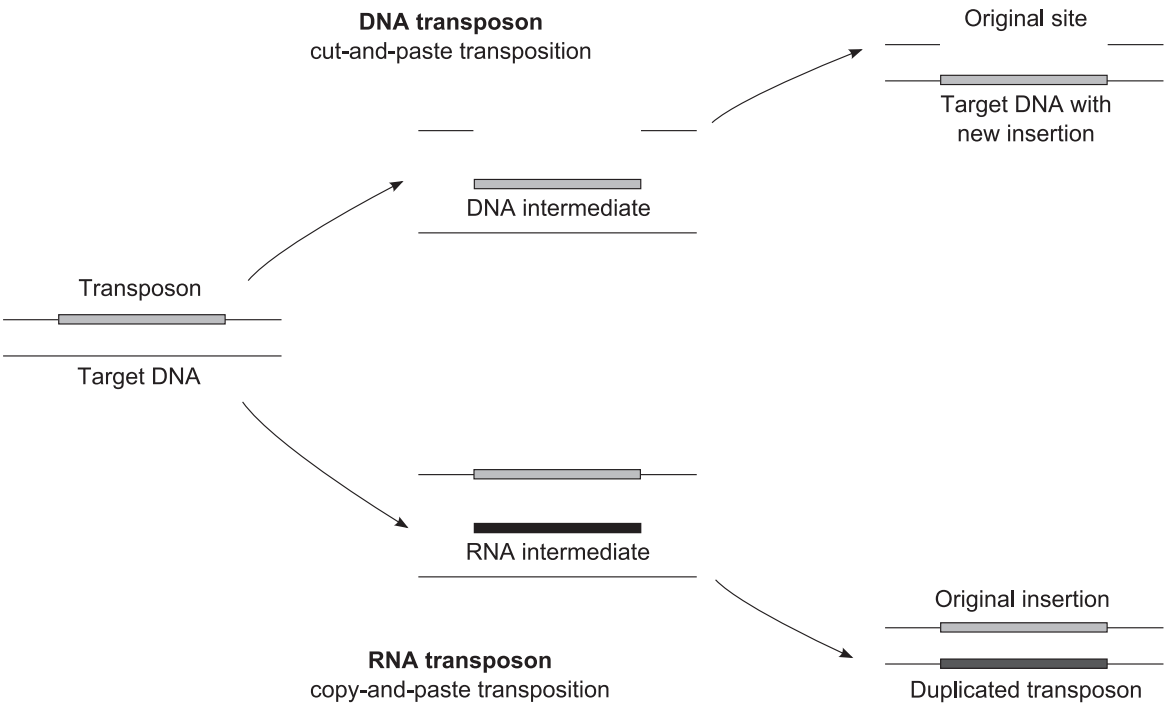


Figure 1.9: A comparison of methods of transposition for DNA transposons and RNA transposons.

1.3.7: ERVs and retrotransposons

Retrotransposons can be further divided into LTR transposons and non-LTR transposons (Kazazian, 2004). LTR transposons include ERVs as well as the *Gypsy* and *Ty* elements, and contain, at minimum, *gag*, *pro,* and *pol* domains flanked by LTRs. Most ERVs also contain an *env* domain, although not all lineages do (Craig, 2002, Magiorkinis et al., 2012). LTR transposons tend to be present in the vertebrate genome in relatively low copy numbers compared with non-LTR elements, although they can by much more prevalent in other eukaryote genomes (Lower et al., 1996, Wicker et al., 2007). Retrotransposons also contain

various promoter regions that can affect the regulation of gene expression if integration occurs near coding regions (Boeke and Stoye, 1997).

Non-LTR elements are made up of long interspersed elements (LINEs) and short interspersed elements (SINEs). As their name suggests, these do not contain LTR regions although they still utilise an RNA intermediate for transposition. LINEs are relatively long elements of several kb in length, and encode reverse transcriptase and a nuclease. Their 3` region is made up of either a poly-A tail, tandem repeats, or an A rich region. SINEs on the other hand are much shorter elements, of about 80 to 500 bp in length (Wicker et al., 2007). Unlike other retrotransposons, SINEs are not coding elements, and are instead believed to be the result of accidental retrotransposition of Polymerase-III transcripts. The 3` end of a SINE element may either be an A or AT rich region, or a poly-T tail (Wicker et al., 2007). Both LINEs and SINEs have a very high copy number in genomes, with many thousands of copies present (Lower et al., 1996).

1.4: Transitioning from genetics to genomics

1.4.1: Next generation sequencing technologies

Next generation sequencing, also called 2nd and 3rd generation sequencing, offers the capacity to sequence large regions of DNA, such as bacterial artificial chromosome (BAC) inserts up to whole genomes, for a much lesser cost than traditional Sanger sequencing. The so-called 2nd generation sequencing methods still currently require amplification of the target template, while 3rd generation sequencing methods are being developed that allow sequencing from single strands of DNA (Glenn, 2011). A number of technologies and platforms are available for this, including 454 pyrosequencing (Roche), Illumina/Solexa (Illumina), SOLiD (Applied Biosystems), and PacBio (Pacific Biosciences), all of which are commercially available. Of these, the 454 and Illumina technologies are the most readily accessible and cost effective. Most of these technologies produce much shorter read lengths than traditional Sanger sequencing (approx 700 bp) but are capable of producing a far greater number of reads at a time.

The Illimina HiSeq systems are most commonly implemented by commercial sequencing services, and are capable of producing upwards of 60 billion bases (Gb) from a single run (Illumina). While read lengths for this system are relatively short (100–150 bp), the large number of raw reads produced makes it the most cost effective method for generation of raw data (Glenn, 2011). However, this large amount of data can pose problems, with high-performance computing (HPC) facilities required for assembly (Glenn, 2011).

In contrast, 454 sequencing produces a much smaller amount of data overall, with 450–700 million bases (Mb) produced from a single run. This is offset by much longer read lengths of 600–1000 bp for current technologies (Roche). Unlike the Illumina technology, this allows for most assemblies to be completed on upper-end desktop computing resources, eliminating the need for access to HPC facilities (Glenn, 2011). Furthermore, the additional length of these reads allows for the use of additional DNA barcoding tags or primers and subsequently an additional degree of parallelisation of samples within a single cell of sequencing (Binladen et al., 2007, ten Bosch and Grody, 2008). This allows for multiplexing of samples within a sequencing run, thereby increasing the number of samples that can be sequenced while minimising costs.

1.4.2: Crocodilian genomics

With the exception of research into commercially farmed crocodilian species, such as *C. porosus*, much of the research into crocodilian genetics relates to untangling the phylogenetic relationships between species, or estimating diversity within a species. As evident from Section 1.1.3, much of the research carried out between crocodilian species focuses on elucidating the phylogenetic relationships between and within the various crocodilian genera. This research has largely focussed on mtDNA and highly conserved nuclear genes. While these resources are adequate for small-scale analyses of divergence and diversity, such resources can offer limited insights into the broad-scale evolution of crocodilian genomes.

As with all uncharacterised genomes, much of the gene discovery and characterisation is carried out through comparisons between the species of interest and closely related taxa. Such research has provided insights into the functionality of specific genes and gene families of interest, such as the keratin genes (Alibardi and Toni, 2007, Ye et al., 2010), and the MHC (Jaratlerdsiri et al., 2012), but has limited potential for the identification of the regulatory pathways responsible for the expression of these genes, or the development of complex traits.

The advent of next generation sequencing technologies and subsequent advances in these technologies has led to a proliferation of new genomes being sequenced. While previous projects have focussed on model species and livestock, the decrease in the cost of whole genome sequencing means that more genomes can now be sequenced. Consequently, the sequencing of three crocodilian genomes has been carried out, with the intention of elucidating the evolutionary biology of these taxa (St John et al., 2012). The specifics of this project will be outlined in Chapter 6. This latest advancement will allow for much more detailed comparisons to be made between the major crocodilian lineages, as well as providing a reference for future studies of crocodilians.

1.4.3: Genetic and genomic characterisation of ERVs

There are various methods available for the detection and characterisation of ERVs. Early genetic studies have relied on hybridisation using radioactive probes derived from known retroviral sequences or synthetic probes directed at the PBS. These methods have been

targeted at DNA fragments, such as from restriction enzyme digests, or against genomic libraries. Positive fragments identified in this manner may be further characterised by restriction enzyme fragment length analyses or more recently, through sequencing (Andersen et al., 1979, Boyce-Jacino et al., 1992, Gifford and Tristem, 2003). These methods are time consuming, but have the benefit of being able to obtain complete ERV sequences if appropriate probes are used (Gifford and Tristem, 2003).

For faster methods of detection, polymerase chain reaction (PCR)-based typing using primers directed against conserved retroviral motifs is frequently used. PCR-based methods usually do not obtain the full length ERV sequence, but will provide sufficient data for most phylogenetic analyses (Gifford and Tristem, 2003). PCR-based screening is also not suitable for detailed searches for novel integrations or disease association as these studies frequently require knowledge of flanking sequence (Stoye, 2001). It is possible to build out from known sequences using a primer walking method to obtain full length proviruses (Kambol et al., 2003), or to amplify full length sequences using long range PCR of the LTR sequences are known (Gifford and Tristem, 2003). However, these methods can be complicated by the presence of multiple copies of related proviral sequences, and may fail to detect insertions that are more ancient, or where the primer binding sites are degraded.

Next generation sequencing technologies have provided a wealth of additional resources that may be used for ERV detection. The large volumes of sequence data that can be produced using these technologies have made it possible to sequence large regions of DNA with minimal knowledge of the target sites or ERV sequence. Furthermore, the increasing number of whole genome sequences now becoming available provides a wealth of new data for the study of the evolution of ERVs and their interactions with their hosts.

In silico (computer-based) detection of ERVs allows the extraction of both ancient and modern retroviral sequences as well as their flanking regions (Gifford and Tristem, 2003). However, this technology also comes with its challenges as screening and detection methods can be computationally expensive, and the quantity of data produced may be many times greater than that produced by PCR screening and sequencing.

Methods of detection using *in silico* screening usually involve comparing genomic sequences against databases of known retroviral or TE sequences (Blomberg et al., 2009). This method

works well for known classes and relatively modern insertions, but may not be able to detect novel sequences and lineages. As a consequence, a number of other approaches have been designed around the detection and interactions of unique retroviral signatures (Blomberg et al., 2009, Gifford and Tristem, 2003). These programs and the methods involved will be discussed further in Chapter 5.

1.5: Rationale for this project

Despite the growing literature surrounding ERV evolution, there have been very few studies examining ERVs of non-mammalian vertebrates in any detail. In addition to this, the apparent genetic predisposition to runting observed in *C. porosus* suggests a link between rearing environment and inherited factors. Given the potential for ERVs to modify gene regulation in the genome, or to cause disease themselves, it is possible that these elements are involved. Although the methods implemented herein have the potential to detect exogenous retroviral insertions as well as endogenous insertions, a lack of evidence for exogenous retroviral infections (see Section 1.1.6) would suggest that most, if not all, detected proviruses are of endogenous origin. This project seeks to provide a significant contribution to the literature in this area and provide the basis from which future investigations into the possible links between ERVs and crocodilian disease can be carried out.

The overall aim of this project is to characterise crocodilian ERVs, with a specific focus on ERVs in *C. porosus*, to examine their evolution and potential impacts on the genomes of these species. This project encompasses the characterisation of crocodilian ERVs, evolutionary studies of these ERVs across key crocodilian species, and explores the replicative potential of these ERVs in their host genomes. Initial studies presented in Chapters 2 and 3 establish the diversity of ERVs across a population of *C. porosus*, and within a *C. johnstoni* individual using a highly conserved region of the ERV genome. Chapter 4 describes the screening and sequencing of *C. porosus* DNA likely to contain full length ERV sequences in order to characterise potentially functional ERVs and provide an estimate of the ERV content of a crocodilian genome. Finally, the sequenced genomes of three key crocodilian species (*A. mississppiensis*, *C. porosus*, and *G. gangeticus*) are interrogated using bioinformatics tools (Chapters 5, 6, and 7) to gain insights into the evolution of ERVs in these species.

It should be noted that the nomenclature for crocodilian ERVs is refined throughout the thesis, resulting in changes to naming conventions between chapters. Chapter 2 uses previously published nomenclature (CERV1 and CERV2) to define ERV clades. This was subsequently found to conflict with ERV lineages previously described in other species, prompting a more specific nomenclature system as outlined in later chapters.

Chapter 2: Strong purifying selection in endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia

Author contributions for published material

The work presented in this chapter has been published as:

Chong, A. Y. Y., Atkinson, S. J., Isberg, S. & Gongora, J. 2012. Strong purifying selection in endogenous retroviruses in the saltwater crocodile (*Crocodylus porosus*) in the Northern Territory of Australia. *Mobile DNA,* 3**,** 20.

I confirm that, as the primary author, I contributed to the design of the study, generated the data, performed all analyses, interpreted the data, and wrote the manuscript.

Ms Sarah J. Atkinson contributed some of the sequence data, performed preliminary analyses, and assisted with early drafts of the manuscript as part of her fourth year honours project. This work was conducted under my guidance and with my assistance.

Dr Sally Isberg collected the samples used in this study and provided advice on the final manuscript.

Dr Jaime Gongora conceived, guided the design of the study, and provided advice regarding analysis, data interpretation, and finalising manuscript.

2.1: Background

Endogenous retroviruses (ERVs) are a group of retrotransposons derived from germ-line integrations of exogenous retroviruses and are found in the genomes of most vertebrate taxa (Lower et al., 1996). The ERV complement of mammalian taxa has been studied in detail, particularly in humans, primates, model organisms, and to a lesser extent, domestic species (Barrio et al., 2011, Belshaw et al., 2004, Garcia-Etxebarria and Jugo, 2010). However, there is very little information regarding diversity and distribution of retroviruses in non-mammalian vertebrates, with the exception of those of the chicken (Tristem et al., 1996). Research into the diversity of ERVs within these taxa has focussed more on specific elements or the distribution of the various ERV classes across species, rather than detailed studies into the ERV complement of a specific species (Chandra et al., 2001, Clark et al., 1979, Gifford et al., 2005, Herniou et al., 1998, Jacobson et al., 2001, Jaratlerdsiri et al., 2009, Martin et al., 1999, Martin et al., 1997, Martin et al., 2002). Thus, there is little data on evolution and diversity of ERVs within individual non-mammalian vertebrate species including crocodilians. To address this, we have investigated the distribution and evolution of these retroelements in *Crocodylus porosus* (saltwater crocodile).

Once integrated into a host genome, ERVs quickly become defective due to selection against the functional retroviruses (Gifford and Tristem, 2003, Stoye, 2001). While these ERVs are mostly non-functional, degenerate ERVs may also retain the capacity to replicate if the necessary regulatory sequences are present and the proteins required for replication are provided by other functional ERVs or exogenous retroviruses (Gifford and Tristem, 2003). Movement and proliferation of ERVs throughout the genome is one of the processes by which multiple related ERV lineages may occur. These lineages may evolve independently within the host genome, to the point that a single genome may contain many thousands of copies of a provirus from a single infection (Stoye, 2001, Tristem, 2000).

ERV replication within the genome can occur through a number of mechanisms, such as reinfection, retrotransposition, and complementation. The likelihood of each of these occurring is dependent on the functionality of the various retroviral domains. For example, reinfection requires that all retroviral genes are functional, and is the method by which retroviruses may infect other host cells (Bannert and Kurth, 2006). Replication within host cells may occur through retrotransposition or complementation. Retrotransposition occurs

when the ERV utilises its own encoded domains to integrate proviral copies into new locations in the cellular genome. Complementation is where the proteins required for replication are supplied by other ERVs or exogenous retroviruses (Bannert and Kurth, 2006, Belshaw et al., 2004).

Exogenous retroviruses and their endogenous counterparts comprise a large and diverse family that can be divided into seven genera: A*lpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, Lentivirus* and *Spumavirus*. ERV classification into these genera is generally based on similarity to classified exogenous retroviruses (Jern et al., 2005). The discovery of a divergent clade of endogenous retroviruses in the Order Crocodylia (families *Alligatoridae*, *Crocodylidae*, and *Gavialidae*) (Martin et al., 2002) suggests that these taxa may harbour hitherto unseen retroviral diversity, and potentially functional novel elements. Subsequent research has identified two clades of crocodilian ERVs (CERVs) (Jaratlerdsiri et al., 2009). One of these groups, termed CERV1, falls within the *Gammaretrovirus* related ERVs and has only been isolated from species within *Crocodylidae*, while the other, CERV2, forms a separate cluster distinct from other ERVs. This second clade of ERVs has been identified in a number of species within both *Crocodylidae* and *Alligatoridae*. This evidence for recent and ancient ERV insertions in these taxa makes it an ideal candidate for the exploration of ERV evolution and the diversification and differentiation of ERVs at species level.

There are 23 recognised species within the Order Crocodylia, belonging to nine genera. *Alligatoridae* consists of the genera *Alligator*, *Caiman*, *Paleosuchus* and *Melanosuchus*, while *Crocodylidae* consists of *Crocodylus*, *Osteolaemus*, and *Mecistops*, and *Gavialidae* consists of *Tomistoma* and *Gavialis* (Li et al., 2007, Roos et al., 2007). *C. porosus* has the broadest geographical distribution of all crocodilian species with populations in Australia, the Indo-Pacific region, South-East Asia, and up to India (Crocodile Specialist Group, 1996c, Russello et al., 2007). It is one of two crocodilian species found in Australia, and the only farmed crocodilian species in this country.

Given the current knowledge of the distribution of ERVs in crocodilians, it would be expected that the majority of ERV sequences will be the result of ancient reinfections and retrotransposition. However, sufficient sequence data are not available to assess the evolutionary processes associated with retroviral proliferation within these species.

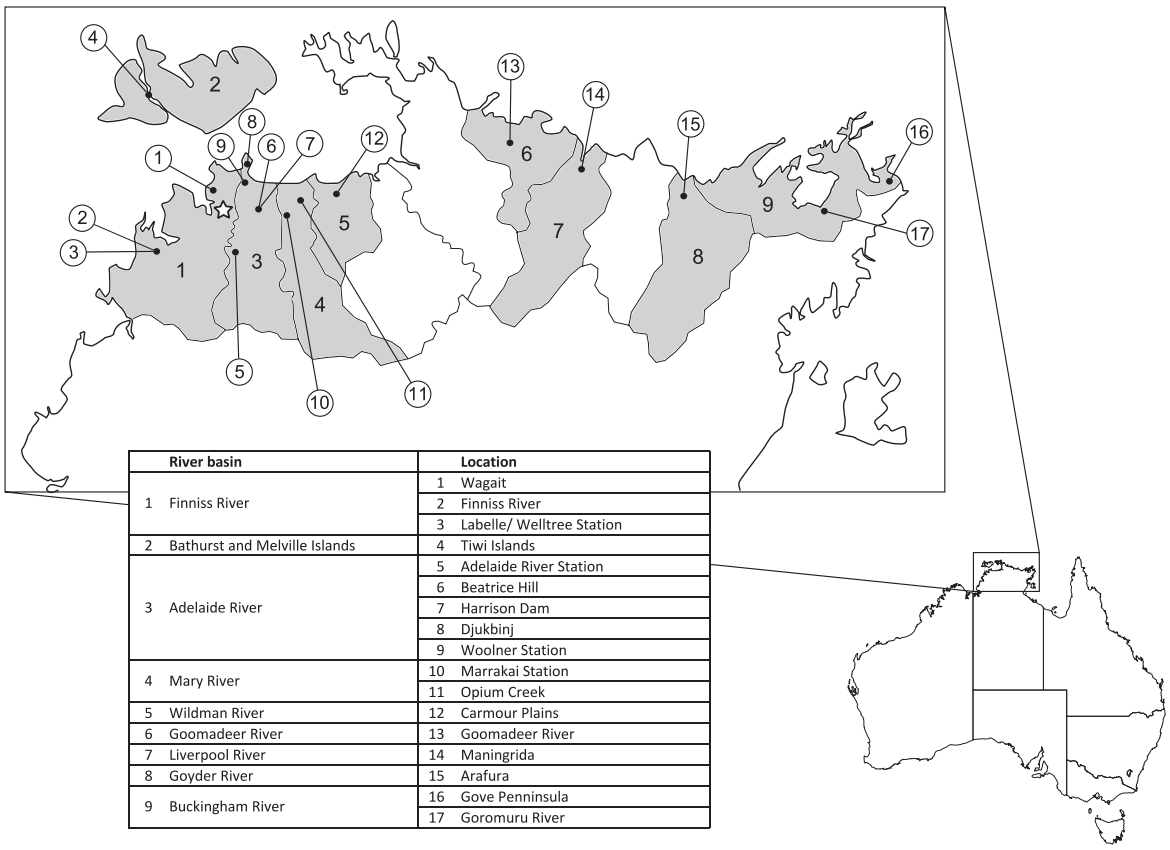
Crocodilian genomes display a significantly lower mutation rate than other vertebrate species (Eo and DeWoody, 2010, Hugall et al., 2007, Lynch, 1997, Ray et al., 2004), providing a good opportunity to study the dynamics of rapidly evolving DNA such as ERVs in a slow mutation rate genome. *C. porosus* is an ideal candidate species for these studies given the interest in sequencing its genome (St John et al., 2012) and ready access to samples from specimens hatched in commercial farms.

Here we present the results of a survey into the ERV complement of *C. porosus* based on analysis of the *pro-pol* gene region. The study focuses on the genetic diversity and potential functionality of these ERV fragments from animals across the Northern Territory of Australia (Figure 2.1). This is one of the first in-depth studies into the diversity of ERVs within a single reptilian species; and will encompass a large number of individuals across a large portion of the range of *C. porosus*.

2.2: Materials and Methods

2.2.1: Sampling

Blood samples were collected from *C. porosus* hatchlings from nests across 17 locations, representing nine river basins in the Northern Territory, Australia (Figure 2.1). The animals sampled were from eggs collected under the Northern Territory Government's ranching program. One to two individuals per clutch were sampled, from a total of 45 clutches. Blood samples were collected from the cervical sinus as described by Lloyd and Morris (1999). DNA was extracted using the QIAamp DNA Mini kit (Qiagen).



River basin	Location
1 Finniss River	1 Wagait
	2 Finniss River
	3 Labelle/ Welltree Station
2 Bathurst and Melville Islands	4 Tiwi Islands
3 Adelaide River	5 Adelaide River Station
	6 Beatrice Hill
	7 Harrison Dam
	8 Djukbinj
	9 Woolner Station
4 Mary River	10 Marrakai Station
	11 Opium Creek
5 Wildman River	12 Carmour Plains
6 Goomadeer River	13 Goomadeer River
7 Liverpool River	14 Maningrida
8 Goyder River	15 Arafura
9 Buckingham River	16 Gove Penninsula
	17 Goromuru River

Figure 2.1: Sampling locations in the Northern Territory, Australia. Map of the northern end of the Northern Territory, Australia, showing the river basins and locations of sampling sites. Shaded regions indicate the river basins included in this study. Numbers within shaded regions correspond to the basin names in the first column of Table 1.1, while numbers in circles correspond to the locations listed in the second column. The star indicates the location of Darwin, the largest city in this area. Image adapted from the Australian Bureau of Meteorology (http://www.bom.gov.au).

2.2.2: PCR amplification and sequencing

PCR was used to amplify a 700–1000 bp region of the retroviral *pro-pol* gene region using universal primers (Tristem, 1996). Amplicons were gel purified and cloned using the pGEM-T Easy Vector and JM 109 *Escherichia coli* cells (Promega, Madison, WI, USA) according to manufacturer's instructions. To ensure that the correct inserts were present, clones were verified by PCR as described above, and by *Eco*RI enzyme digests after purification. Positive clones were purified and sequenced using Sanger sequencing at the Australian Genome Research Facility (AGRF, Brisbane, QLD, Australia).

2.2.3: Sequence alignment and analysis

Nucleotide sequences were aligned using CLUSTALW (Thompson et al., 1994) as implemented in the program package MEGA5 (Tamura et al., 2011). Representatives of the major sequence groups identified here were compared with previously identified ERV sequences in the GenBank and RepBase databases using BLASTX (Altschul et al., 1997) and Censor (Kohany et al., 2006) respectively. Unique haplotypes from this study were identified using FaBox (Villesen, 2007) and re-aligned against other similar sequences generated in this study using the program MACSE (Ranwez et al., 2011). The resulting alignments were translated in MEGA5 (Tamura et al., 2011) using the standard vertebrate genetic code tables, and putative amino acid sequences were aligned in CLUSTALW using the BLOSUM matrix with residue specific and hydrophilic penalties, and high gap penalties as described by Xiong and Eickbush (Xiong and Eickbush, 1988). The presence of conserved retroviral motifs and domains was assessed based on similarity to motifs defined by Sperber et al. (Sperber et al., 2007). Genetic distances were calculated using the Jukes-Cantor model for the nucleotide alignments and the JTT model for amino acid alignments. The presence of recombinant sequences was evaluated using the program RDP3 (Martin et al., 2010) with default program settings.

2.2.4: Phylogenetic analysis

Phylogenetic analyses were used to detect evidence of sub-lineages within each of the major clades. For both major clades, Neighbour Joining and Maximum Likelihood analyses were carried out with 1000 bootstrap replicates and representative sequences from the respective retroviral genera as outgroups (HERV-E for CERV1 and HERV-L for CERV2). Neighbour Joining trees were created in MEGA5 (Tamura et al., 2011) using the Jukes-Cantor and Poisson corrections to account for multiple substitutions. The best fit model of substitution (CERV1: HKY, JTT; CERV2: GTR, JTT for nucleotide and amino acids respectively) was determined using ModelGenerator (Keane et al., 2006) and Maximum Likelihood phylogenies were generated in PhyML (Guindon et al., 2010).

Additional phylogenetic analyses were performed to assess the evolutionary relationship of the novel *C. porosus* sequences with other published ERV sequences. This dataset comprised of five representative novel sequences from this study, 55 published sequences from other

species within Crocodylia, and 113 published sequences from other species (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Jern et al., 2005, Martin et al., 1999, Martin et al., 2002) (See Appendix I, Table S2.1 for accession numbers and further details). Due to the highly diverse nature of the sequences from the various species, sequences were aligned using the program MAFFT, and the E-INS-i algorithm (Katoh et al., 2005). Phylogenetic trees were created as described above.

2.2.5: Tests for selection

Codon based Z-tests were carried out in MEGA5 (Tamura et al., 2011) to investigate overall selective forces acting on the two major CERV clades in *C. porosus*. Datasets were analysed using the Nei and Gojobori method with the Jukes-Cantor correction to account for multiple substitutions (Nei and Gojobori, 1986). Tests were conducted to test for non-neutrality, positive, and purifying selection. Synonymous and non-synonymous ratios were also calculated for these datasets in PAML v4.4 (Yang, 2007) using a likelihood ratio test (LRT) to assess significance of the detected selection signatures.

Further details on the amplification conditions, RDP program settings, selection criteria for representative sequences, and the PAML model comparisons are available in Appendix III.

2.3: Results

2.3.1: Sequence overview

A PCR survey of ERVs in 47 individuals yielded a total of 227 clones, which were subsequently sequenced. These sequences represented 176 novel DNA haplotypes and 126 novel amino acid haplotypes [GenBank accession numbers: JX157669 to JX157844]. Sequences ranged in length from 665 to 957 nucleotides. Up to 12 unique sequences were identified per individual, with all individuals yielding at least one positive clone, although very few sequences were recovered from more than one clone per individual. All sequences except for two could be assigned to the two CERV clades previously described (Jaratlerdsiri et al., 2009) based on visual inspection and genetic similarity values. A total of 45 haplotypes belonged to clade CERV1, and 129 haplotypes were assigned to clade CERV2. Overall,

CERV2 clones were more prevalent across the 17 sampling locations (Figure 2.1). The proportion of sequences from each of the CERV clades did not appear to vary from the overall average proportion of sequences recovered across all locations (Table 2.1). Although recent genetic studies have revealed some level of diversity among animals from the same or similar sampling locations (Jaratlerdsiri et al., 2012, Luck et al., 2012), a comparison with the current dataset was not possible as different regions of the genome (mtDNA and gene coding regions) were used in these studies.

BLAST searches and comparisons with sequences in Repbase suggested that one of the outlier sequences, haplotype 58, showed similarity to the exogenous *Epsilonretrovirus*, Walleye dermal sarcoma virus (WDSV). Haplotype 119 appeared more similar to a gypsy retrotransposon. A third sequence, haplotype 107, also appeared to be divergent from other crocodilian sequences but phylogenetic analysis grouped this within clade CERV2. In addition to this, 18 sequences belonging to clade CERV1 were found to encode intact ORFs (haplotypes 1, 2, 14, 19, 25, 26, 27, 30, 46, 77, 87, 88, 90, 140, 148, 157, 175 and 176). No intact ORFs were identified from CERV2 related ERV fragments. There was no apparent prevalence of intact ORFs from any particular river basin (data not shown).

Table 2.1: Sequences obtained from each sampling location and the assigned clades.

River basin	Sampling location	Number of clutches	Number of individuals	Number of sequences	CERV1[a]	CERV2[a]	Other
Finniss River	Wagait	1	1	3	-	3	-
	Finniss River	2	2	3	-	3	-
	Labelle/ Welltree Station	4	4	9	4	5	-
Bathurst and Melville Islands	Tiwi Islands	3	3	10	3	7	-
Adelaide River	Adelaide River Station	2	2	14	1	13	-
	Beatrice Hill	1	2	9	1	7	1
	Harrison Dam	4	4	19	4	15	-
	Djukbinj	4	4	26	6	19	1
	Woolner Station	3	4	12	5	7	-
Mary River	Marrakai Station	5	5	23	4	19	-
	Opium Creek	1	1	2	-	2	-
Wildman River	Carmour Plains	1	1	1	1	-	-
Goomadeer River	Goomadeer River	3	3	21	3	18	-
Liverpool River	Maningrida	2	2	12	3	9	-
Goyder River	Arafura	2	2	4	1	3	-
Buckingham River	Gove Penninsula	3	3	11	3	8	-
	Goromuru River	4	4	19	6	13	-
Total number of samples		45	47	198	45	151	2

[a] Note that number of sequences does not equal number of haplotypes as some haplotypes were recovered from more than one individual

The YXDD retroviral reverse transcriptase motif was highly conserved in both CERV clades. CERV1 sequences also contained a number of gammaretroviral motifs from the protease and reverse transcriptase domains. CERV2 sequences had regions showing some similarity to spumaviral domains, though only three of the possible six domains were detected (Table 2.2). Haplotype 58 was shown to contain an epsilonretroviral motif as well as a number of motifs shared between *Gamma*- and *Epsilonretroviruses*.

Table 2.2: Conserved retroviral *pro-pol* motifs in crocodilian ERV sequences.

CERV clade	**Motif**	**Genera**	**Motif sequence**[a]
CERV1	PR2	*Gammaretrovirus*	(A/V)L(V/L)DTG(A/S)TFSM
	PR3	*Gammaretrovirus*	LLG(Q/R)DLLTKL
	RT1	*Gammaretrovirus*	YN(S/T)PILGV(L/P)K(A/V)
	RT2	*Gammaretrovirus*	SVLDLKDAFFSI(P/S(L
	RT3	*Gammaretrovirus*	(Q/R)LMWTVLPQGF(I/V)(A/V)AP
	RT4	*Gammaretrovirus*	LL(H/Q)YVDD(I/L)L
Haplotype 58	PR2	*Gammaretrovirus*	VLLDGTATMSM
	PR3	*Gammaretrovirus*	LLGRDLLCK
	RT1	*Gammaretrovirus*	CNTPVLPVRKP
	RT2	*Epsilonretrovirus*	TVIDLCAAFFPIPV
	RT3	*Gammaretrovirus*	HTLNTQLPQGYTKSP
	RT4	*Gammaretrovirus*	LVQYVDDIL
CERV2	RT2	*Spumavirus*	(A/T)AID(L/P)K(D/E)MF(C/Y)(H/Q)IPL
	RT3	*Spumavirus*	F(E/K)G(C/H/R)VY(E/K)WKVC(P/S)(E/Q)GYKNSP
	RT4	*Spumavirus*	(L/N)SYVDD(I/L)L

[a] Residues presented here are those that occurred in more than one sequence. Motifs are based on those defined by Sperber et al. (2007). For the complete alignments, see Appendix II, Figure S2.1.

Sequences within each of the CERV clades were highly conserved, with pairwise genetic distances of 0.058 and 0.039 between nucleotide sequences, and 0.071 and 0.084 for amino acid sequences within CERV1 and CERV2 respectively. Visual inspection of the sequence alignments for each of the two clades did not reveal any distinct grouping within CERV1 but suggested that an additional two groups exist within CERV2 (Appendix II, Figure 2.1). Within CERV2, genetic distances decreased further when calculated within each of these

groups, with distance values of 0.008/0.017 (CERV2a), 0.008/0.011 (CERV2b) and 0.015/0.029 (CERV2c) for nucleotide and amino acid alignments respectively.

The distribution of stop codons and frameshift mutations differed between the two clades. Sequences within CERV1 contained very few stop codons or frameshifts that were shared between sequences, and those that were, tended to be present in only a small number of sequences. In contrast, stop codons and frameshifts within CERV2 sequences were mostly present in all of the sequences within a group.

Recombination analyses detected five recombinant sequences within the two major CERV clades, and two possible recombinants where only trace evidence of a recombination event was detected. Within CERV1, the recombinant sequences were haplotype 1, and Csi_IV, with Cni_I also suspected to be recombinant. Haplotypes 60, 81, and Ami_II were detected from clade CERV2, and Ami_I was also suspected to be a recombinant sequence (Appendix I, Table S2.2). Of these, only haplotypes 1 and 81 within *C. porosus* show strong evidence of recombination, being reported as having a significant *P*-value using all methods implemented.

The expected parental sequences for recombinant sequences were isolated from different individuals and in some cases from different species. While this reduces the likelihood that the observed recombination occurred during amplification, the possibility cannot be ruled out, since we do not know the full extent of the ERV complement of individuals used in this study. The sites involved in recombination were different in all recombinant sequences detected and did not appear to correspond to specific regions of the *pro-pol* domain.

2.3.2: Selection

Tests for selection across the major clades gave consistent results both across species in Crocodylia, and within *C. porosus*. Codon based Z-tests suggest that purifying selection is occurring across crocodilian species (CERV1: $Z = 7.496$, $P < 0.001$, CERV2: $Z = 2.224$, $P = 0.014$), and within *C. porosus* (CERV1: $Z = 7.060$, $P < 0.001$, CERV2: $Z = 2.633$, $P = 0.005$). Comparisons of d_N/d_S across the different ERV clades gave average d_N/d_S ratios under the one ω model between 0.2696 and 0.5539 (Table 2.3). In all cases, allowing sites to

evolve under positive selection produced a better fit in the resulting phylogenies, although the overall d_N/d_S ratios strongly supported purifying selection acting on these elements (Appendix I, Table S2.3). Positive selection was detected at a small number of sites in both clades, but these sites do not appear to correspond to retroviral motifs.

Table 2.3: Average d_N/d_S for each of the selection scenarios tested.

Hn	Model[a]	Average d_N/d_S	
		CERV1	CERV2
H0	M0: One ratio	0.4904	0.5539
H1	M3: Discrete	0.5187	0.6898
H2	M1a: Nearly neutral	0.4344	0.3057
H3	M2a: Positive selection	0.5234	0.5864
H4	M7: Beta	0.4217	0.3349
H5	M8: Beta & ω	0.4973	0.6217

[a] Analysis was conducted using PAML (Yang, 2007). Model names are those defined in the program.

2.3.3: Sequence clustering and phylogenetic analysis

Nucleotide and amino acid trees created using Neighbour Joining and Maximum Likelihood methods present very similar topologies with little phylogenetic differentiation within each clade. Neither Neighbour Joining nor Maximum Likelihood methods provided any better resolution of the phylogenetic relationships between the sequences. Overall, the tree topology was similar within both clades, with very short internal and terminal branches (Appendix II, Figure S2.2). The lack of phylogenetic resolution is most notable within the clade CERV1. No highly supported groups or lineages were identifiable within *C. porosus*, or when these sequences were compared with those of other crocodilian species. Interestingly, we observed a tendency for sequences encoding intact ORFs to cluster within one clade of the CERV1 phylogeny. Within CERV2, phylogenetic trees supported the presence of three groups of sequences within the clade, with moderate bootstrap support, consistent with what was observed with sequence genetic distances.

Neighbour Joining and Maximum Likelihood analyses incorporating retroviral sequences from non-crocodilian taxa consistently placed the CERV1 sequences with the

Gammaretrovirus related ERVs. CERV2 related sequences consistently clustered with the *Spumaviruses*. Haplotype 58 clustered with the *Epsilonretrovirus* related ERVs, while haplotype 119 was placed within the *Spumaviruses* but separate from the CERV2 sequences (Figure 2.2). While there appears to be no host species-related sorting among CERV1 sequences, groupings within CERV2 suggest some degree of lineage specific evolution at the level of host family. A *Crocodylidae* specific group was observed, consisting of sequences from *C. porosus*, *C. niloticus*, and *C. mindorensis* (Philippine crocodile). Haplotype 107 was placed midway between this lineage and the majority of known CERV2 sequences, which consist mostly of those isolated from the alligators and caimans. Notably, within clade CERV2, the majority of sequences from *Crocodylidae* cluster together, while those from *Alligatoridae* appear to be more divergent from each other.

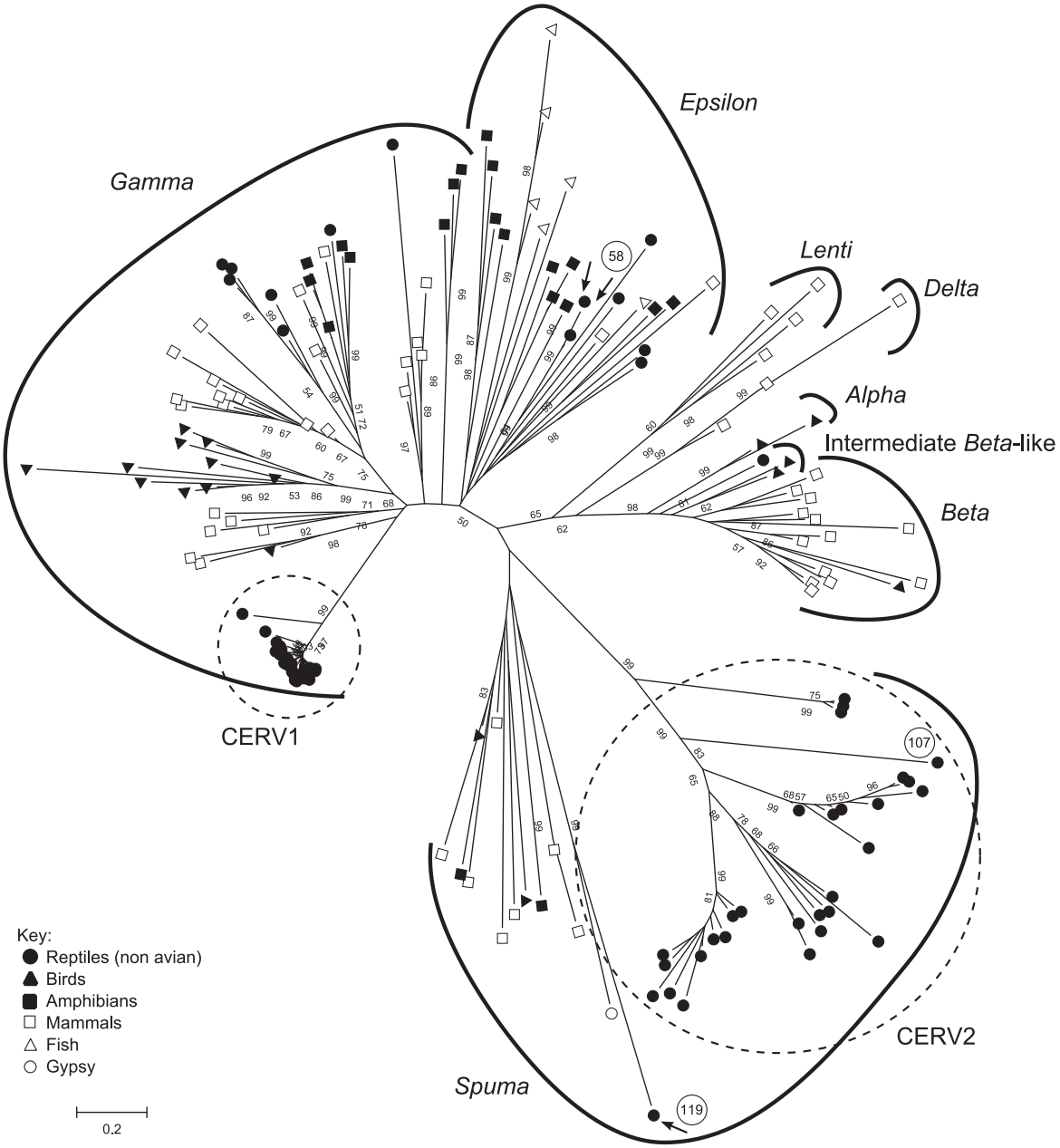


Figure 2.2: Phylogenetic clustering of crocodilian ERVs (CERVs). Neighbour Joining tree based on aligned amino acid sequences from the retroviral *pro-pol* gene region. The final alignment length was 799 characters including gaps. The general host species taxa are indicated by symbols. The two major clades of ERVs found in crocodilians are indicated by dashed circles. Circled numbers refer to haplotypes 58, 107, and 119. Arrows indicate additional crocodilian ERV sequences. Numbers near branches indicate bootstrap support values.

2.4: Discussion

The data presented in this study suggest that there are high levels of sequence diversity in *C. porosus* ERV sequences. The sequences isolated in this study correspond largely with the two major CERV clades previously identified. In addition, we have identified novel sequences that appear to be related to other retroviral genera. Within the clades, we have found evidence of strong purifying selection acting across both of the major clades which is suggestive of recent integration or transposition. We have found preliminary evidence to propose the presence of sublineages within clade CERV2. While it is unlikely that this study encompasses the full extent of retroviral diversity in *C. porosus*, the data generated here provide a comprehensive insight into the process by which ERVs have populated the genomes of crocodilians, and their evolution within the genome of this species.

2.4.1: High diversity present in CERV clades

A large number of novel sequences were generated, with very few haplotypes being recovered more than once. This high diversity of ERVs haplotypes within *C. porosus* can be explained by several possible scenarios: i) several recent and independent infection events by exogenous retroviruses from the various retroviral genera that have resulted in the current ERV diversity; ii) a single infection event by an exogenous retrovirus from each of the represented retroviral genera followed by repeated reinfection by the same ERV lineages; iii) a single infection event by an exogenous retrovirus from each of the represented retroviral genera followed by replication of ERVs within the genome, either by retrotransposition or complementation.

The similarity of sequences within each of the two major clades within *C. porosus* would make the first of these scenarios unlikely. Of the remaining scenarios, the second scenario would appear to explain the pattern of evolution seen in clade CERV1, while the presence of shared stop codons in CERV2 suggests proliferation through retrotransposition or complementation (Belshaw et al., 2004). Further support for each of these methods of replication will be discussed in the following sections. Both of these scenarios result in many lineages of related retroviruses that can replicate and mutate independently (Tristem, 2000). This leads to a collection of proviruses that show high levels of nucleotide and amino acid

diversity while at the same time retaining the original sequence characteristics, as seen in this study. This level of sequence diversity is not uncommon for ERVs, and is comparable to what has been observed in mammalian ERVs (Klymiuk et al., 2003, Nascimento et al., 2011).

A low frequency of recombinant sequences was detected among the crocodilian ERV sequences, suggesting that recombination does not play a large role in the generation of ERV diversity in *C. porosus*. The detection of recombinant sequences where the predicted parental sequences were isolated from other species is indicative of ancestral recombination events. Given the rarity of cross species transmission (Martin et al., 1999) and the distinct distributions of crocodilian species, it is unlikely that these lineages arose from cross species transmission of ERVs between crocodilian species.

2.4.2: Potential for autonomous replication in CERV clades

It is plausible that CERV1 may still be active within the genome of *C. porosus*, replicating through reinfection of host cells. Reinfection restores the fitness of the replicating ERV, thereby increasing preservation of the lineage (Bannert and Kurth, 2006, Herniou et al., 1998) and allowing for further proliferation within the host genome. Based on the high number of sequence variants and high level of sequence similarity, we propose that this clade represents the product of a fairly recent integration event. The majority of stop codons present within this clade occupied unique positions within each sequence, indicating that it is unlikely that these ERVs arose through retrotransposition or by complementation with other related ERVs (Belshaw et al., 2004). This is further supported by the presence of sequences with intact ORFs, and the strong purifying selection that has been observed within this clade (Bannert and Kurth, 2006, Herniou et al., 1998).

The observed clustering of sequences encoding intact ORFs suggests that there may be a particularly active strain of ERV that has largely managed to escape inactivation within *C. porosus*. This clade also includes a number of available sequences from other species within *Crocodylidae*, suggesting that there may also be active ERVs present in these species. While most ERV insertions are inactivated by mutation shortly after integration, it is plausible that active lineages may still retain their capacity to replicate by reinfection well after species divergence (Benit et al., 1999, Katzourakis and Gifford, 2010).

On the other hand, shared stop codons in sequences from clade CERV2 mean that replication by reinfection is unlikely, lending support to retrotransposition or complementation as possible means of replication. Replication by either of these methods does not require that all genes are functional. Retrotransposition, for example, does not require a functional *env* domain (Bannert and Kurth, 2006, Herniou et al., 1998). Complementation on the other hand does not require functional proteins within the provirus, providing that the required regulatory regions within the LTRs are intact, and that missing functional proteins are supplied by an exogenous retrovirus or partially intact ERV (Bannert and Kurth, 2006, Belshaw et al., 2004). The strong levels of purifying selection detected in this clade suggest that this clade has recently been active, although sequence data from the other retroviral domains are needed to determine the likely method of replication.

2.4.3: Low levels of phylogenetic resolution

The rapid proliferation of ERVs within a host genome can also confound attempts to differentiate ERVs by phylogenetic analyses. This is especially in the case of recent integration events where not enough evolutionary time has passed to allow insertions to develop distinguishing or phylogenetically informative mutations. In addition, the mutation rates of the *pro-pol* domain are, comparatively, the lowest of the various retroviral domains (McClure et al., 1988). While this characteristic makes this region ideal for studies of ERV proliferation across taxa, it could be argued that regions with typically higher mutation rates such as *gag* or *env* may be more appropriate for generating phylogenies within a species (Jern et al., 2005, McClure et al., 1988).

Furthermore, studies into the nucleotide substitution rate of crocodilian nuclear and mitochondrial sequences suggest that this is much lower in crocodilians than in most other vertebrates (Eo and DeWoody, 2010, Hugall et al., 2007, Lynch, 1997, Ray et al., 2004). Thus, degenerate ERV sequences are likely to accumulate changes at a slower rate in crocodilians than most other vertebrate species, leading to a low level of host lineage specific evolution, as well as low levels of lineage differentiation. This has the effect of reducing our ability to detect host species specific lineages based on these data alone.

For this reason, it could also be argued that other more quickly evolving retroviral domains should be considered to provide resolution between host specific lineages. The characterisation of the remaining ERV domains will also provide further insights into the methods of replication, and potential for reinfection. As such, future studies into the diversity of ERVs within crocodilians should now be oriented towards characterising the entire length of proviral insertions rather than individual domains.

2.4.4: Estimated infection times of the ERV clades

Strong purifying selection and low levels of phylogenetic resolution on both ERV clades suggests a recent population expansion may have occurred. Regardless of the method by which this is achieved, replication of elements results in the expansion of the population and the creation of autonomous, but related lineages that are capable of replicating and evolving independently (Tristem, 2000). In relatively recent population expansions, therefore, it would be expected that sequences would still share high levels of sequence similarity. This is corroborated by short internal branch lengths, and the lack of phylogenetic resolution seen across the CERV alignments.

In the absence of LTRs or knowledge of the founding retroviral sequence, the ages of the initial integration events of the crocodilian ERVs can only be estimated from what is known about the species phylogenies and the assumed presence or absence of the ERVs in each of these species. Based on nucleotide and amino acid similarity to previously classified ERVs, and the presence of conserved retroviral motifs, we propose that the ERV complement of *C. porosus* came about through infection by three related lineages of retroviruses belonging to the gamma-, epsilon-, and spumaviral genera. The first of these infections would be that leading to the CERV2 clade, as sequences have been identified in species representing all families within Crocodylia. This integration is likely to pre-date the Alligator-Crocodile split, approximately 90 million years ago (MYA) (Oaks, 2011). The infection that gave rise to the CERV1 clade is likely to have occurred after this time period. The presence of nearly identical CERV1 sequences in other species within *Crocodylidae* would indicate that this integration could have occurred prior to diversification of the various crocodile species, at least 20−30 MYA (Oaks, 2011).

Sequence divergence and phylogenetic evidence supports the presence of three sublineages within CERV2. These three groups are characterised by a number of diagnostic positions including shared frameshift mutations and stop codons within each group, and is moderately supported by bootstrap analyses on the phylogenetic trees of the CERV2 clade. This evidence furthers the notion that this clade represents an older integration event that has been present in the genome for a sufficient amount of time for the differentiation of distinct sequence lineages.

In the case of the *Epsilonretrovirus* related sequence, haplotype 58, the full extent of its proliferation within crocodilians is not known, as only one other similar sequence has been isolated – from *G. gangeticus* (gharial) (Herniou et al., 1998). Thus, it cannot yet be determined whether this lineage is also present among crocodilians, or if it is the result of cross species retroviral infection involving a limited number of crocodilian species. The apparent rarity of this sequence within the genome also raises questions as to why it has not multiplied further, or why it is so much less likely to be detected. Deeper or different sampling strategies may be required to understand the reasons behind this.

2.4.5: Reclassifying CERV2

Contrary to the data presented by Jaratlerdsiri et al. (2009), CERV2 related sequences were grouped within the *Spumaviruses* rather than forming a separate distinct clade. This can mainly be attributed to the use of different alignment algorithms between studies. The conventional strategy of high gap penalties within global alignments can be problematic when aligning highly divergent sequences, such as ERV sequences, where the use of high gap penalties can result in the forced alignment of non-homologous sequence regions (Huang et al., 2006). Instead, we elected to use the alignment program MAFFT, which implements algorithms specifically designed for the alignment of highly divergent sequences. These algorithms result in alignments based around a series of local alignments of conserved regions, and allows for long lengths of un-alignable sequence between the conserved domains (Katoh et al., 2005). In studies such as this, where ERV discovery is, more or less, *de novo*, we believe that this may be a more effective method for sequence comparison, as novel sequences may share only a low level of similarity with known sequences, making them difficult to align and potentially reducing the power of downstream analyses.

2.5: Conclusions

We propose that the ERV complement of *C. porosus* has come about through a combination of recent infections and replication of ancestral ERVs. Two major clades are present as a result of infection by gammaretroviral and spumaviral lineages. Strong purifying selection acting on these clades suggests that this activity is recent or still occurring in the genome of this species. We have uncovered a large amount of sequence variation within both of the major clades of ERV present in *C. porosus*, as well as the presence of an additional lineage that appears to be present in the genome to a much lesser degree. While no host taxa dependent clustering was observed, there is evidence for the divergence of sub-lineages within the more ancient ERVs in *C. porosus*. The discovery of elements encoding an intact ORF, and therefore the potential for autonomous replication is an interesting development that warrants further investigation.

Chapter 3: Lineage specific evolution of endogenous retroviruses in the freshwater crocodile (*Crocodylus johnstoni*)

Author contributions for submitted material

The work presented in this chapter has been to submitted to *Journal of Herpetology* in June, 2013 as:

Chong, A. Y., Kjeldsen, S. & Gongora, J. Lineage specific evolution of endogenous retroviruses in the Freshwater Crocodile (*Crocodylus johnstoni*).

I confirm that, as the primary author, I contributed to the design of the study, generated the data, performed all analyses, interpreted the data, and wrote the manuscript.

Ms Shannon Kjeldsen contributed some of the sequence data, performed preliminary analyses, and assisted with early drafts of the manuscript as part of her fourth year honours project. This work was conducted under my guidance and with my assistance.

Dr Jaime Gongora conceived, guided the design of the study, and provided advice regarding analysis, data interpretation, and finalising manuscript.

3.1: Introduction

Endogenous retroviruses (ERVs) are a diverse group of vertebrate transposable elements derived from germline infections by exogenous retroviruses (Lower et al., 1996). Crocodilians (Crocodylia) have previously been shown to harbour a particularly diverse range of ERVs, some of which are only distantly related to the currently recognised retroviral genera (Chapter 2) (Jaratlerdsiri et al., 2009, Martin et al., 1999, Martin et al., 2002). However, these studies have focussed on examining diversity across species and within a population. Given the high degree of ERV sequence variation that may be present within a population, studies from a singular genome may provide a better understanding of the micro-evolutionary processes to which ERV insertions are subjected. To address this, we provide insight into the process of replication and diversification of ERVs from a single specimen of the *C. johnstoni* (freshwater crocodile) and compare this with the insights offered by broader studies of ERV diversity.

Retroviruses (family *Retroviridae*) consist of exogenous and endogenous RNA viruses. These are classed into seven major genera based on DNA and amino acid sequence similarity, and the structure of the viral genomes (Blomberg et al., 2009, Jern et al., 2005). These genera are: *Alpharetroviruses*, *Betaretroviruses*, *Gammaretroviruses*, *Deltaretroviruses*, *Epsilonretroviruses*, *Lentiviruses*, and *Spumaviruses* (Fauquet et al., 2005). Their endogenous counterparts, ERVs, are generally found as degraded copies of ancient retroviruses. This degradation is a consequence of genetic mutations, such as base substitutions, insertions, deletions, and recombination, and can make it difficult to classify older ERV sequences within these genera (Barrio et al., 2011, Lower et al., 1996). Consequently, ERV classification may reflect the likely exogenous retroviral genera, eg *Gammaretrovirus*-like, or be classed into three major groups: ERV1, ERV2, and ERV3. ERV1 loosely encompasses the *Gamma-* and *Epsilonretroviruses*, ERV2, the *Alpha-* and *Betaretroviruses*, and ERV3 the *Spumaviruses*. (Blomberg et al., 2009, Jern et al., 2005).

The basic proviral ERV genome consists of three major coding domains flanked by long terminal repeats (LTRs). The coding regions encode the viral capsid protiens (*gag*), retroviral enzymes (*pro-pol*), and the retroviral envelope proteins (*env*). The LTRs are not translated, but contain the promoter regions responsible for initiating transcription and reverse transcription at various stages of the retroviral replication cycle (Coffin, 1992).

Once integrated into a host genome, ERVs may retain some capacity to replicate, forming numerous lineages, some of which will also be capable of replication. This can potentially lead to thousands of related insertions of varying ages dispersed throughout the host genome (Stoye, 2001, Tristem, 2000). ERV replication within a host genome can largely be explained by three major mechanisms. The first of these, reinfection, mimics the retroviral replication cycle whereby elements insert into the genome, are replicated, and mature viral particles are budded into the extracellular matrix where they can infect naïve host cells. This method of replication among ERVs is rare, as it requires that all retroviral genes and domains are intact (Belshaw et al., 2004). The second of these mechanisms, retrotransposition in *cis*, requires functional promoter regions and *gag* and *pol* domains. Alternatively, the third mechanism, complementation in *trans,* requires only that the promoter regions are intact, with some or all retroviral enzymes and proteins provided by other endogenous or co-infecting exogenous retroviruses (Bannert and Kurth, 2006, Belshaw et al., 2004). The method of replication by which ERVs have proliferated within a genome can be determined by examining sequence similarities and phylogenetic relationships between insertions.

Following the initial integration into the genome, most ERV insertions will be removed from the host species by purifying selection against those that reduce the genetic fitness of the host. Providing that the effect of the insertion is not immediately lethal, these recently endogenised ERVs may retain the capacity to cause disease through active reinfection. This is of particular interest in crocodilians as previous studies have suggested that at least one lineage of ERVs may retain the capacity to replicate by reinfection (Chapter 2). The association of ERVs with novel diseases has been observed in other taxa. One such example of this is the KoRV in *P. cinereus*, which has been associated with neoplasia and increased susceptibility to chlamydiosis. The presence of viral transcripts in blood plasma, and variation in the number of proviral copies suggests that KoRV is still active within the koala genome (Tarlinton et al., 2005). While this virus does not appear to be causing disease directly, it has been suggested that specific viral variants or insertion locations may play a role in determining susceptibility or progression of associated diseases (Rosenberg and Jolicoeur, 1997, Tarlinton et al., 2005).

On the other hand, it should also be noted that ERVs can cause disease without active reinfection of host cells. Silenced or dormant ERVs may also have the capacity to cause disease, either by reactivation, or by interfering with normal cellular functions. Dormant ERVs may also be reactivated through interaction or recombination with exogenous

retroviruses (Bishop, 1978), or released from transcriptional suppression by activation of the immune system or other stressors (Cho et al., 2008, Perl, 2003). ERVs that are capable of replicating within a host cell by retrotransposition have the capacity to cause disease by integrating into new genomic locations, there they can disrupt gene expression by insertion within a coding region, or by altering the expression of genes surrounding the insertion site (Gifford and Tristem, 2003, Katzourakis et al., 2005, Stoye, 2001).

There are 23 recognised species of crocodilians (Order Crocodylia), making up three families: *Alligatoridae* (alligators and caimans), *Crocodylidae* (crocodiles), and *Gavialidae* (gharials) (Li et al., 2007, Roos et al., 2007). *C. johnstoni* is endemic to Australia and can be found in coastal and inland regions of northern Australia, from northern Queensland, across to northern Western Australia (Cogger, 1992). This species is generally found in freshwater habitats upstream of tidal influence, and in the absence of competition from the *C. porosus*, with which its range overlaps, its range may extend into tidal, saline waters (Webb and Manolis, 2010). Hunting and the introduction of invasive species, such as *Chaunus [Bufo] marinus* (cane toad), greatly reduced *C. johnstoni* populations to the point where it was classified as vulnerable (Crocodile Specialist Group, 1996b, Doody et al., 2009, Letnic et al., 2008, Webb et al., 1984). Subsequent conservation efforts and the increasing focus on ecotourism, have increased population numbers to a level where *C. johnstoni* is now considered a low risk species, although it is still protected (Crocodile Specialist Group, 1996b, Webb and Manolis, 2010).

Retroviral diseases have not previously been observed in crocodilians. However, recently, it has been noted that a number of juvenile freshwater crocodiles were presenting with an unidentified syndrome, typically characterised by cutaneous lesions and extensive lymphoid infiltration of internal organs, possibly indicative of a retroviral infection (Melville et al., 2012). We chose to investigate the potential for replication competent retroviruses from one of these diseased individuals as there has been some suggestion that other crocodilian species may harbour active retroviral elements.

To date, ERV knowledge in crocodilians has focussed on phylogenetic analyses across species of this lineage and only one other crocodilian species, *C. porosus*, has been studied extensively. Studies using the *pro-pol* gene region have shown that crocodilian ERVs cluster into two major groups (Chapter 2) (Jaratlerdsiri et al., 2009). The first of these clusters with

other ERV1 sequences, with one lineage showing similarity to exogenous *Gammaretroviruses*, and another that was more similar to the *Spumaviruses.* Alignments constructed from conserved domains of those genes, demonstrated that the second major group appears to be only distantly related to other recognised ERVs, and it has alternately been placed in a separate clade from recognised ERVs (Jaratlerdsiri et al., 2009), and as a distantly related clade to ERV3 and spumaviral ERVs (Chapter 2). More recently, identification of complete proviral sequences from this clade has shown that it is distinct from other recognised ERV clades, and thus has been given a new classification, termed ERV4 (J. Jurka, pers comm.). Due to the high sequence divergence observed between these proviral sequences and the closest related spumaviral ERVs, this latest classification is likely to represent the actual relationship between this ERV clade and other sequenced ERVs.

To our knowledge, this study represents the first study into ERVs within a single animal, providing the opportunity to examine the micro-evolution of ERV lineages within a genome, rather than their evolution within a population, genus, or family. ERV studies in non-model organisms, such as crocodilians, are more likely to focus on the examining diversity across taxa and inferences of likely points of ERV integration with regards to speciation, rather than a systematic characterisation of the ERV complement of a particular species (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Martin et al., 1999, Martin et al., 1997). Furthermore, the characterisation of ERVs within a species is instead usually restricted to species where the genome sequence has been generated, such as in humans, cattle, and in the dog (Barrio et al., 2011, Belshaw et al., 2004, Garcia-Etxebarria and Jugo, 2010). In the absence of genomic resources for *C. johnstoni*, a targeted study such as this provides an effective, alternative method for ERV characterisation.

Here we present the results of an in-depth survey of ERVs in a single *C. johnstoni* individual by targeted amplification and isolation of the highly conserved *pro-pol* domain. The ability to examine ERVs within a single genome will allow for a closer examination of the microevolution of ERVs at the level of a single individual. In particular it will provide useful insights into the mechanisms by which ERVs may be replicating, and the potential impacts of ERV integration at the gene level. In this current study, we have also investigated the possibility of recovering replication competent ERVs from an individual experiencing immunological stress. In doing this, we have also investigated the possibility of a novel

infection by a crocodilian ERV with the capacity to cause disease, through the investigation of possible preferential distribution of ERV insertions across different tissue types.

3.2: Materials and Methods

3.2.1: Sample collection

Tissue samples for this study were collected from a euthanised *C. johnstoni* specimen at Berrimah Veterinary Laboratories (Berrima, NT) as part of routine necropsy procedures. Samples of lung, liver, kidney, and spleen were used for this study. DNA was extracted using a standard phenol-chloroform extraction (Sambrook and Russell, 2006).

3.2.2: PCR amplification and sequencing

PCR was carried out to amplify fragments of genomic ERV insertions. Universal primers for the retroviral *pro-pol* region were used maximise the number of insertions that could be recovered (Forward sequence: GTK TTI KTI GAY ACI GGI KC, reverse sequence: ATI AGI AKR TCR TCI ACR TA). These primers amplify a 700–1000 bp region between highly conserved, functional, retroviral motifs (Tristem, 1996). PCR was carried out in duplicate, in 25 µL reaction volumes, containing 100 pmol of each primer, 2 mM $MgCl_2$, 0.16 mM dNTPs, PCR buffer and 1 U of high fidelity *Taq* polymerase. Thermocycling conditions were carried out as follows: initial denaturation at 94°C for 2 min, 35 cycles of 45°C (30 s), 72°C (60 s) and 94°C (30 s), followed by a final annealing period of 3 min at 45°C and a final extension period of 10 min at 72°C.

ERVs may be present in the genome in many thousands of copies (Stoye, 2001). Consequently, multiple amplicons were expected in each reaction. To distinguish between each of these, amplicons were gel purified and artificially cloned using the pGEM-T Easy Vector and JM109 *Escherichia coli* cells (Promega). Bacterial cloning allows the separation and amplification of individual DNA fragments by use of a bacterial plasmid vector and transfection competent *E. coli* cells. Positive clones were verified by PCR as described above before sequencing by Sanger sequencing at the Australian Genome Research Facility (AGRF, Brisbane, Australia).

3.2.3: Sequence analysis

Nucleotide sequences were initially aligned using CLUSTALW (Thompson et al., 1994) as implemented in the program package MEGA5 (Tamura et al., 2011). Primer sequences were then trimmed from the alignments as primer mispriming can increase or mask variation at the annealing sites. Unique haplotypes were then identified using FaBox (Villesen, 2007) and putative amino acid translations obtained by aligning the resulting haplotypes using the program MACSE (Ranwez et al., 2011) before translating into amino acids using the standard vertebrate genetic code tables. To confirm that the recovered sequences were from ERV insertions, nucleotide and amino acid haplotypes were compared against published ERV sequences in GenBank using BLASTN (Altschul et al., 1990, Zhang et al., 2000) and BLASTP (Altschul et al., 1997). For analyses across species, we realigned the sequences using the E-INS-i algorithm in the alignment program MAFFT (Katoh et al., 2005) to take into account the presence of multiple conserved regions as well as more variable regions within the ERV sequences.

Sequence identification tags were assigned based on the species of origin (Crjo for *C. johnstoni*), the region from which the sequence originated (*pro-pol*), and the haplotype number, for example "CrjoERV-pro-pol-1". Haplotypes were used over individual sequences as is not possible to ascertain whether the sequences within a haplotype belong to different insertions or are from multiple amplifications of the same insertion.

We calculated genetic distances using the Jukes-Cantor and JTT models for nucleotide and amino acid sequence alignments respectively. These models were selected to account for the possibility of multiple mutations at each site in the alignment. Tests for sequence recombination were carried out using RDP (Martin et al., 2010) with default settings. Sequence clustering and identification of representative sequences was carried out using CD-HIT with the recommended program settings for nucleotide and amino acid alignments (Li and Godzik, 2006). These sequences were then compared with published ERV sequences in GenBank and RepBase using the NCBI BLAST suite (Johnson et al., 2008) and Censor (Kohany et al., 2006) respectively. We also investigated the selective pressures operating on these sequences through the Codon based Z-tests and Tajima's neutrality test in MEGA5 (Tamura et al., 2011). Codon based Z-tests were carried out using the Nei and Gojobori method with the Jukes-Cantor correction.

Phylogenetic analyses using Neighbour Joining and Maximum Likelihood methods were used to determine the evolutionary relationships within the sequences recovered in this study and between sequences recovered from crocodilians (Chapter 2) (Jaratlerdsiri et al., 2009, Martin et al., 2002); as well as the relationships between these insertions and other vertebrate species (Herniou et al., 1998, Jern et al., 2005, Martin et al., 1999). Neighbour Joining trees were created in MEGA5 using the Jukes-Cantor and Poisson corrections to account for multiple substitutions, with 1000 bootstrap replications. Maximum Likelihood analyses were implemented in PhyML using HKY and JTT models for nucleotide and amino acid alignments as determined by ModelGenerator (Guindon et al., 2010, Keane et al., 2006). aLRT values were used to determine statistical support for each of the branches (Anisimova and Gascuel, 2006). Due to the divergence between sequences across all retroviral genera, the phylogeny incorporating sequences from a range of vertebrate taxa were created using Neighbour Joining analyses and the p-distance method with 1000 bootstrap replicates.

3.3: Results

3.3.1: ERV diversity in *C. johnstoni*

A total of 43 nucleotide sequences were obtained from this study, ranging from 739–787 nucleotides in length. These encoded 28 unique nucleic acid haplotypes [GenBank accession numbers: KC790541 to KC790568] which were reduced to 19 after translation into amino acid sequences. All sequences were found to contain stop codons and/or frameshift mutations. The translated amino acid sequences were between 250 and 265 amino acids in length, including stop codons and frameshift mutations.

Sequence haplotypes were made up of between one and 13 sequences recovered from positive clones. Most of the haplotypes recovered were unique to a specific tissue type (Table 3.1), although three haplotypes were shared between at least two tissue types: haplotype CrjoERV-pro-pol-2 was found in all four tissue types, CrjoERV-pro-pol-8 was recovered from the liver and spleen, and CrjoERV-pro-pol-11 was common between liver and lung. No recombination was detected within the sequences generated from this study. Similarities between the sequences recovered in this study and those from previous studies revealed that all haplotypes except for one, CrjoERV-pro-pol-28, could be assigned to the ERV4 clade

(previously CERV2) (Chapter 2) (Jaratlerdsiri et al., 2009). Within these sequences, overall p-distances were 0.028 at the nucleotide level, and 0.062 for amino acid sequences.

Table 3.1: Summary of the clones and nucleic acid sequence haplotypes obtained from each tissue in this study.

Tissue type	No. clones	No. haplotypes[a]	No. unique haplotypes[a]
Liver	17	12	9
Lung	5	5	3
Kidney	14	11	10
Spleen	7	5	3
Total	43		

[a] Note that the number of haplotypes listed for each tissue type does not equal the total number of haplotypes as some haplotypes were recovered from more than one tissue type

Sequence clustering analysis of the *C. johnstoni* haplotypes suggested the presence of four distinct sequence groupings at the 95% similarity level, three of which corresponded to the ERV4 related sequences, and the fourth to haplotype CrjoERV-pro-pol-28 (Figure 3.1). The first of these clusters (Crjo-ERV4c in Figure 3.1) consisted of 23 out of the 28 haplotypes recovered. An additional three sequences (CrjoERV-pro-pol-1, -3, and -11) formed a minor group of sequences, and the remaining two sequences (CrjoERV-pro-pol-24 and CrjoERV-pro-pol-28) were deemed to be distinct from each other as well as the previously identified groups of sequences. While certain haplotypes were found in certain tissues, these sequence clusters did not appear to be more prevalent in any of the tissue types studied.

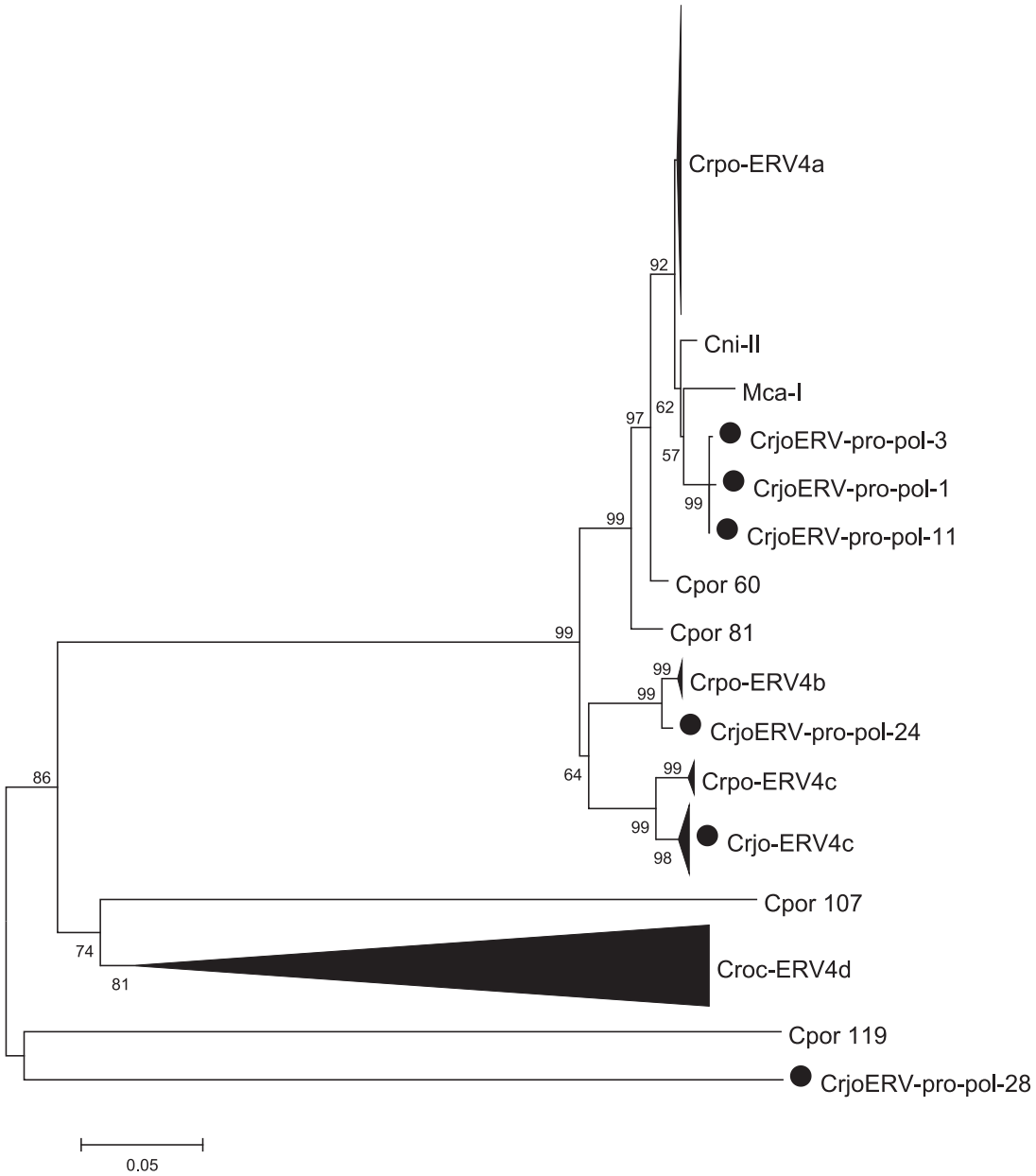


Figure 3.1: Neighbour Joining tree of crocodilian ERV4 *pro-pol* sequences, with the larger, closely related clades collapsed for clarity (indicated by shaded triangles). *C. johnstoni* sequences are indicated by a shaded circle. The final alignment was 1032 positions in length, with sequences ranging from 625 to 787 nucleotides in length. The scale bar at the bottom left indicates inferred genetic distance between sequences. Numbers at the nodes indicate bootstrap support values greater than 50%. 'Crpo' refers to *C. porosus* ERV clades, 'Crjo' refers to those from *C. johnstoni*, and 'Croc' refers to crocodilians in general. Other sequence names have been retained from their respective publications.

CrjoERV-pro-pol-28, showed no similarity to other crocodilian ERVs recovered to date (Chapter 2) (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Martin et al., 1999, Martin et al., 2002). Comparisons with published retroviral sequences in the GenBank and RepBase databases showed that this sequence is similar to the foamy viruses, a diverse lineage of retroviruses within the *Spumavirus* genera, although similarity values were low. The closest related sequences were the Feline foamy viruses (35% similarity at the amino acid level), and the Simian foamy viruses (34% similarity).

3.3.2: Selection on *C. johnstoni* ERVs

Tests for selection within the major sequence cluster in *C. johnstoni* suggested that purifying selection is occurring within this clade. Codon based Z-tests rejected the scenario of evolution under neutral selection ($Z = -2.160$, $P = 0.033$) in favour of evolution under purifying selection ($Z = 2.064$, $P = 0.021$). Tajima's neutrality test supported this ($D = -1.557$).

3.3.3: Phylogenetic analysis

Neighbour Joining and Maximum Likelihood phylogenetic trees also supported the presence of four clades within the *C. johnstoni* sequences recovered from this study (Figure 3.1). However, these analyses did not provide further resolution of the phylogenetic relationships between sequences within the clades. Trees derived from the ERV4 sequences were star-like, with short internal and terminal branches, and a large number of unresolved divisions or polytomies (Appendix II, Figure S3.1), similar to that described in *C. porosus* (Chapter 2).

Subsequent phylogenetic comparisons with other crocodilian ERV sequences revealed that all except five haplotypes clustered with the *C. porosus* ERV4c sub-lineage (Figure 3.1). CrjoERV-pro-pol-24 clustered with sub-lineages ERV4b in *C. porosus*. Within these lineages, sequences from each species clustered together in distinct clades. CrjoERV-pol-1, -3, and -11 clustered within the major *C. porosus* ERV4a sub-lineage, while CrjoERV-propol-28 remained separate from the crocodilian ERV4 sequences. Based on these analyses,

CrjoERV-pro-pol-1, -10 (representing Crjo-ERV4c), -24 and -28 were selected for further interspecies analysis.

Phylogenetic analysis of the haplotypes from this study compared with a selection of ERVs from all vertebrates confirmed that the ERV4 sequences clustered within a crocodilian specific clade. These analyses also confirmed that CrjoERV-pro-pol-28 was more closely related to the foamy viruses than the crocodilian ERVs (Appendix II, Figure S3.2).

3.4: Discussion

3.4.1: Recent activity and species specific evolution in crocodilian ERV4 lineages

This study indicates that the ERV insertions within *C. johnstoni* are still under purifying selection, suggesting that these elements have recently been active and replicating. The high levels of sequence similarity between the isolated fragments, and presence of shared stop codons and frameshift mutations support the view that the observed ERV diversity is a result of recent duplication from existing insertions rather than multiple infection events (Belshaw et al., 2004). Likewise, the numerous stop codons and frameshift mutations make it unlikely that the elements recovered in this study encode intact coding domains within the *pro-pol* region, and are therefore not capable of autonomous replication. Therefore, we propose that the proliferation of ERV sequences within the *C. johnstoni* genome is likely to be a result of retrotransposition or complementation (Bannert and Kurth, 2006, Belshaw et al., 2004, Katzourakis et al., 2005).

The ERV sequence data from this study, combined with what has been observed in *C. porosus* provides the first evidence of host specific clustering within sub-lineages in crocodilian ERV4 sequences. This is likely to be the result of a pre-speciation infection, pre-dating the *Alligatoridae-Crocodylidae* divergence, approximately 90 MYA (Oaks, 2011). We propose that this infection event was then followed by species specific replication and lineage evolution (Figure 3.1, Figure 3.2). Sequences from the sub-clades from both species share a common node within the wider ERV4 clade, suggestive of common ancestry. Furthermore, the high levels of sequence similarity observed in the *C. johnstoni* ERV4 sequences from this study, as well as between these sequences and those isolated form *C. porosus*, lend credence to the theory that these lineages arose from a single infection event.

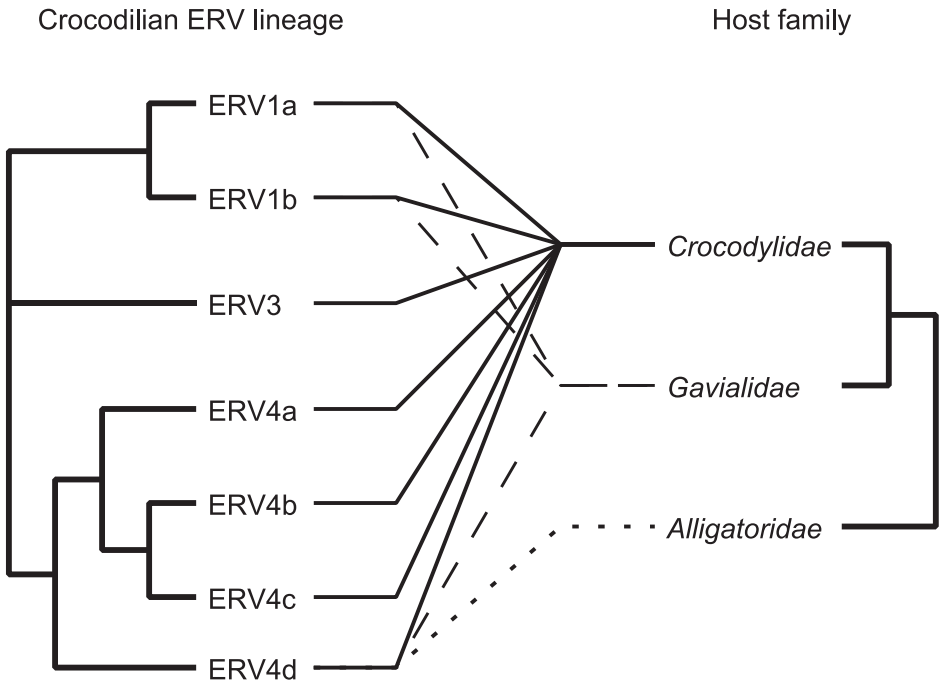


Figure 3.2: Graphical representation of the phylogenetic relationships between the crocodilians ERV clades described to date and their host species within Crocodylia, highlighting the range of ERV lineages recovered from *Crocodylidae* in comparison with the other two families. The tree on the left shows the ERV clades while the tree on the right is the three crocodilian families. The lines in the middle indicate which ERV clades have been isolated from each of the crocodilian families, with solid lines being those from *Crocodylidae*, dashed lines from *Gavialidae*, and the dotted line from *Alligatoridae*.

Species specific evolution of ERV lineages is widespread across vertebrate taxa, although the best studied examples of these are from taxa including model organisms due to the availability of genomic resources. Notable examples of this include the intracisternal A-particles (IAPs) in rodents, and the HERV lineages in primates. IAPs are a *Betaretrovirus*-like group of ERVs that have been actively replicating in rodent genomes despite the lack of an identifiable *env* (Magiorkinis et al., 2012, Vogt, 1997). Similarly, a number of HERV lineages in primates have been shown to display differing levels of lineage proliferation and degradation across primate species (Bannert and Kurth, 2006, Jern et al., 2004).

Furthermore, once integrated into the genome, proviral insertions are subject to the same levels of nucleotide substitution as the surrounding genomic region (Temin, 1992). This means that the length of time in which an ERV lineage has been present in the genomes of host species can also affect the detection of species specific evolution. It has been proposed

that the ERV1 lineages are too recent to allow for the development of host specific lineages. ERV4, on the other hand, is a much older insertion, likely pre-dating the *Alligatoridae-Crocodylidae* split (Chapter 2) (Jaratlerdsiri et al., 2009). Thus where ERV1 did not show any evidence for species specific lineage sorting, the older ERV4 lineages may have had the molecular time to evolve species specific lineages.

Alternatively, species specific expansions such as this may be the result the acquisition of a complementary helper virus, resulting in replication by complementation (Bannert and Kurth, 2006); or a release from transcriptional suppression, allowing the proliferation of a previously inactive ERV (Maksakova et al., 2008). The animal from which the samples were obtained was observed to be suffering from a novel disease or syndrome not observed in *C. porosus* (see Section 3.1), and it is possible that this could provide the necessary trigger. Stressors such as immunological stress from a novel or virulent pathogen can also alter the transcriptional restrictions on the genome of a host cell, allowing a previously silenced ERV to replicate (Cho et al., 2008, Perl, 2003). As such, it will be interesting to obtain the full sequences of these insertions in order to date the likely time of infection and lineage expansion.

We recovered a large number of unique sequences from ERV4a and ERV4c, although there is no evidence for the prevalence of specific lineages in any of the tissue types used in this study. While this could be due to a technical bias as a result of reaction conditions or incomplete sampling of the ERV complement of each tissue type, it may also be due to an undetected somatic cell infection from a closely related ERV4-like strain of exogenous retrovirus. For example, in the event that this additional infection is from a strain that is closely related to those already present in the genome, it is not always possible to distinguish between these somatic cell infections and endogenous insertions. Furthermore, the absence of tissues for cell or viral culture, and knowledge of the likely physical structure of the exogenous counterparts of these viruses means it is not possible to ascertain whether a novel infection is taking place.

3.4.2: Identification of a novel crocodilian ERV lineage

In addition to the ERV4 lineages discussed above, the *C. johnstoni* genome has revealed an additional, novel infection event in crocodilians that appears to be related to the *Spumaviruses*. Unfortunately, only a single sequence (CrjoERV-pro-pol-28) is available for the crocodilian foamy virus-like element, and as such, it is not possible to determine the extent of this proliferation among crocodilian species. However, given that only a single isolate was recovered from this lineage, we hypothesise that it is present in *C. johnstoni* at a much lower copy number than either the ERV1 or ERV4 clades. If remnants of this infection event are present in other crocodilian species, it is likely that the prevalence of insertions in these species will share similar patterns. Given the apparent rarity of these insertions, it is likely that they are the result of infection by an exogenous retroviral strain of low virulency or replicative activity. An alternative scenario is that there are a number of ancient copies present in the genome that have not been detected through PCR-based surveys such as this due to degradation of the primer binding sites as a result of mutation processes.

3.4.3: Absence of ERV1 sequences in the current study

Surprisingly, this study did not recover any sequence data from either of the ERV1 lineages that have previously been identified in crocodilians. There are a number of reasons why this clade may not have been recovered in this study: i) the individual in question did not harbour any instances of this ERV lineage; ii) this lineage is still at a low copy number in the genome and thus less likely to be detected than other linages; or iii) a technical bias due to universal primers and biased reaction conditions.

It is unlikely that these elements are absent from the genome of the individual used in this study, as this clade has previously been identified in this species. Moreover, the *Gammaretrovirus*-like ERV1 lineage has a widespread distribution among species within *Crocodylidae*, and high sequence similarity among the *Crocodylus spp*. This suggests that it is the result of an infection pre-dating speciation of crocodiles. Given the prevalence of this ERV lineage within *C. porosus*, it is very likely that this insertion has become fixed in the genome of *C. porosus*, and would therefore be present in all individuals of that population, and possibly the species. Despite the lack of detailed studies from other crocodilian species, the widespread distribution and level of sequence conservation suggests that it is also highly

probable that a similar trend would be observed in other *Crocodylus spp*. From this, we can extrapolate that these insertions would likely be present in all species within *Crocodylidae*, and therefore would be expected to be present in all individuals of these species.

It is more likely that the failure to recover these sequences in this study is a combination of a lower copy number in the genome, and a technical bias in the methodology. Previous surveys of ERVs in *C. porosus* have suggested that the ERV1 clades were also less likely to be recovered. Technical bias cannot be ruled out as a contributing factor either, as such surveys use a large pool of primer sequences with different nucleotides at specific positions along the sequence. If the specific primer sequences that will anneal to the targeted binding site are not present, or are present at a lower proportion than other primers, it is less likely that that particular insertion will be amplified, and thus it is less likely to be among the amplicons sequenced.

3.4.4: Crocodilian ERVs may stem from at least four infection events

Based on the data from this study, and previous investigations into ERVs in crocodilians, we propose that the crocodilian genomes have been subject to infection by at least four distinct lineages of retrovirus. These lineages include two ERV1 lineages described by Jaratlerdsiri et al. (2009) and in Chapter 2 (the *Gammaretrovirus*-like clade; previously CERV1, and the *Epsilonretrovirus*-like lineage), the lineage leading to the ERV4 clade of crocodilian ERVs (previously CERV2), and the foamy virus-like ERV represented by CrjoERV-pro-pol-28 (Figure 3.2). This is supported by the fact that the sequences within each of the ERV clades identified in crocodilians so far are more similar, and therefore likely to be more closely related to each other than to sequences from the other clades. Furthermore, the inferred phylogenetic relationships between sequences from each of these clades, and shared nodes within each group, support a common origin for each lineage. Mutations in these ERV lineages likely occurred prior to replication within crocodilians, as many of the stop codons and frameshift mutations are shared between species. Such mutations are indicative of proliferation of ancestral infections followed by further evolution post-speciation.

Crocodilians appear to harbour a very different complement of ERVs compared with other studied taxa, with sequences identified from the *Gammaretroviruses*, *Epsilonretroviruses*,

Spumaviruses, and a fourth divergent clade. With the exception of the *Danio rerio* (zebrafish) (Blikstad et al., 2008), this is a greater number of retroviral genera than other characteristed vertebrates. This complement of ERVs differs greatly from both *Gallus gallus* (chicken), where ERVs have been identified from the *Alpha*-, *Beta*-, and *Gammaretroviruses* (Blikstad et al., 2008), and mammalian taxa, which appear to harbour endogenous copies of beta-, gamma-, and spumaviral infections (Barrio et al., 2011, Blikstad et al., 2008, Garcia-Etxebarria and Jugo, 2010).

The biological reasons behind these differing complements are not known, although immunological, cellular, and genomic factors may all play a part. One theory may be that, since these are likely to be signatures of ancestral infections, these insertions reflect infections of host species with naïve or developing anti-viral immune responses. The ERV4 lineage, in particular, may be the remnants of an ancient retroviral lineage that is no longer as prevalent as the modern exogenous retroviral strains recognised today. Alternatively it may have evolved to a state where it no longer shows similarity to the ancient insertions detected in the crocodilians.

3.5: Conclusions

In conclusion, this study has identified a *C. johnstoni* specific lineage of ERVs, and provided evidence for species specific biases in ERV proliferation. The data generated in this study provide a useful platform for future investigations into the behaviour and impact of ERVs in crocodilian species; expanding our knowledge and understanding of ERVs in crocodilians, and providing further evidence for the susceptibility of crocodilians to infection by a broad range of retroviral pathogens. Future investigations should therefore look into the sequencing and characterisation of other retroviral domains to ascertain the likely mechanism of replication for each of the ERV lineages recovered from these species.

While we were not able to identify any ERV insertions capable of autonomous replication, further investigations into the expression of crocodilian ERVs may provide some insight into the current level of ERV activity in these species. Furthermore, investigations into the possible causes of the lesions and lymphoid infiltration in *C. johnstoni* would benefit from investigations into the expression of these ERVs across the affected tissues, and correlations

between presence of ERV insertions and diseased individuals. Despite this, we believe that the data presented here represent an important contribution to the understanding of the genomic environment of crocodilians, and the ways in which this can impact the genesis of disease in these species.

Chapter 4: Screening and sequencing of a crocodilian bacterial artificial chromosome library

4.1: Introduction

In Chapter 2, a number of ERV sequence fragments encoding intact ORFs were identified, suggesting that these ERVS may still retain some capacity for replication. The availability of BAC libraries for a number of key crocodilian species, combined with next generation sequencing technologies allows for the investigation of these insertions at the complete proviral sequence level; a resource not provided by studies of ORF fragments using standard molecular approaches such as those implemented in the previous chapters. As the potential functionality of ERV proviruses is of particular interest for studies into the evolution and population dynamics of crocodilian ERVs, investigations of the crocodilian ERV lineages were expanded in an attempt to obtain sequence data from the four retroviral coding domains, and genomic regions immediately surrounding the ERV insertions.

From a methodological point of view PCR-based methods for the isolation and characterisation of ERV sequences are constrained by the need for known sequences from which to design oligonucleotide primers, and can be easily confounded by the presence of multiple copies of closely related inserts. Sequence data from the previous chapters suggest that there many insertions from many lineages present in *C. porosus*, and that these insertions still share a large amount of similarity within the genome. While the data generated from such studies is sufficient to provide an overview of the ERV diversity, it does not provide sufficient resolution to identify novel lineages and enhance our understanding its possible associations with disease (Gifford and Tristem, 2003, Stoye, 2001). For this reason, additional methods for sequence retrieval were investigated.

The work presented in this chapter overcomes some of the limitations of traditional methods of ERV detection through the use of DNA hybridisation-based detection in combination with next generation sequencing to screen and sequence parts of a *C. porosus* genomic library containing provial ERV sequences. This approach proves to be useful for the identification and quantification of the ERVs present in the library, in particular to retrieve the DNA fragment containing the ERV lineages of interest, and its surrounding regions.

This approach is facilitated by the shot-gun approach used by next generation sequencing technologies which allows the sequencing and subsequent characterisation of ERV DNA fragments without the need for targeted amplification of the region of interest. The nature of this sequencing technology also allows for determination of the surrounding genomic regions, and therefore, the insertion sites of these ERVs. To enable this, we have used recently created crocodilian BAC libraries, representing two of the three major lineages of crocodilians: *C porosus* (Shan et al., 2009), and *G. gangeticus* (Shan, unpublished).

4.1.1: Classification of ERVs based on genomic structures

Detection of ERV insertions, and the subsequent assignment of these to one of the three major ERV classes, or major retroviral genera can be done based on overall sequence similarity and the presence and sequence of conserved structural or functional motifs within the *env*, *gag*, and *pol* domains. As discussed in previous chapters, it is possible to assign ERV fragments to different ERV classes based on sequence conservation and maintenance of conserved domains within the *pro-pol* region. However, the *gag* and *env* domains are much more variable than the *pro-pol* region, making overall sequence conservation a less useful method for classification.

Despite this, there are a number of conserved motifs and overall structural features that may be used. Within the *gag* domain, motifs such as the major homology region can be used as targets for location of *gag* genes. The major homology region is a highly conserved region of 20 amino acids that has been found in all retroviral *gag* domains (Vogt, 1997). Conservation of this region is likely due to functional significance as it is thought to play a role in the late stages of virion maturation and the release of viral genetic material during infection of a host cell (Craven et al., 1995). Another diagnostic feature of the gag domain is the presence of Zinc finger-like motifs. These vary in number between the retroviral classes – Class II retroviruses (*Alpha-*, *Beta-*, *Lenti-* and *Deltaretroviruses*) generally have two of these domains, while this motif appears to be absent in the *Spumavirus gag* domains (Jern et al., 2005, Vogt, 1997).

The *pro* and *pol* domains are much more conserved across retroviral genera due to the retroviral enzymes these regions encode. Xiong and Eickbush (1990) describe a number of

these with particular focus on sequence conservation across a range of retroviruses and related TEs. Indeed, the primer sequences used in previous chapters have been designed from these highly conserved regions. While no reasons for the significance of these regions are given, it is likely that most of these correspond to structurally or functionally significant motifs, such as the YXDD motif which plays an important role in the catalytic functions in the reverse transcriptase enzyme (Katz and Skalka, 1994).

Despite being the most variable of the retroviral coding domains, the *env* region also contains a number of conserved regions that can be used for retroviral classification. Of particular note are the regions encoding endocytic signals, and an immunosuppressive domain, both of which are important for infection and proliferation within host cells (Denner, 1998, Ohno et al., 1997). While the variation within *env* may prohibit the classification and assignation of novel ERV sequences to a particular ERV class, it can allow for further delineation of ERV lineages within classes.

4.1.2: Assessing ERV diversity using BAC resources

BAC libraries contain large inserts (approximately 150 kb) of genomic DNA in an artificial vector, which is then transfected into transfection competent *E. coli*. These inserts can be replicated with high fidelity during bacterial cell division, making them relatively simple to amplify for downstream processing. BAC libraries have been touted as relatively low cost tools for targeted sequencing of genomic regions, as well as tools to facilitate the physical mapping of these regions (Metzker, 2009, Morozova and Marra, 2008, Shan et al., 2009).

BAC clones containing the region of interest can then be detected by screening by PCR or by DNA hybridisation. Hybridisation screening of BAC libraries involves the hybridisation of labelled nucleic acid probes to BAC DNA which has been fixed to a nylon membrane. The probes may be in the form of short, synthetic nucleic acid sequences, such as overgo probes, or longer lengths of DNA fragments derived from the gene or region of interest. By using only these clones in downstream procedures, one can effectively restrict the DNA regions used for PCR-based analysis or sequencing.

BAC libraries have previously been used for the identification and sequencing of ERV proviruses in a wide variety of taxa (Arnaud et al., 2007a, Barbulescu et al., 1999, Rogel-Gaillard et al., 1999, Turner et al., 2001); albeit with a focus on previously characterised ERV lineages, and the detection of insertion polymorphisms within or between closely related taxa. Additionally, studies such as those by Barbulescu et al. (1999) and Turner et al. (2001) demonstrate the potential for these libraries to aid in the recovery of the sequence surrounding the known ERV regions through the use of restriction enzyme digests and inverse PCR to amplify outwards from known ERV sequence. However, the methodologies in these studies rely on knowledge of the proviral sequences of the ERV insertions in question.

Unfortunately, in the case of relatively uncharacterised species, such as crocodilians, these data are frequently limited to sequence fragments. In these situations, a more general approach to sequencing is required, such as the use of shot-gun sequencing on the DNA BAC inserts. Consequently, the use of BACs for isolation of specific genomic regions of interest, coupled with the less specific amplification and sequencing mechanisms offered by next generation sequencing, may prove to be a useful tool for the recovery of complete proviral sequences where only sequence fragments have so far been identified.

4.1.3: Use of next generation sequencing for analysis of TEs

Next generation sequencing enables the sequencing of large regions of DNA, such as BAC inserts up to whole genomes. Despite the continued interest in the evolutionary dynamics of TEs, such as ERVs, and the impacts of these elements on genome function, the use of this technology for the targeted sequencing of TEs is not common. Instead, the analysis of vertebrate TEs, such as the identification and characterisation of ERV sequences is usually carried out in conjunction with whole genome sequencing projects. However, a limited number of studies of TEs in plants have been carried out using these sequencing technologies in combination with Cot-libraries (collections of DNA fragments separated based on DNA renaturation) and BAC libraries (Choulet et al., 2010, Peterson et al., 2002, Wicker et al., 2006), suggesting that these genomic resources, where available, constitute a largely under-utilised asset for wide scale studies of TEs.

The assembly of reads over TEs and gene families containing genes of high similarity pose a particular problem for next generation sequencing technologies, as the assembly algorithms rely on reads mapping to unique positions within the DNA that has been sequenced. (Macas et al., 2007, Wicker et al., 2006). The presence of multiple copies of an ERV or similar sequence in one assembly can result in these reads being aligned to form a single consensus sequence rather than individual copies of the element (Wicker et al., 2006). The algorithms cannot resolve the mismatched ends of the highly repetitive reads, and therefore break the assembly at these sites.

The LTR regions themselves may also cause a problem with the assemblies, given the short read lengths and low coverage. The LTRs at each end of the ERV genome are identical when initially inserted into the genome, and are likely to remain very similar for some time, particularly in a genome such as the crocodilian genomes where there is a low nucleotide substitution rate. If the sequencing reads are not long enough to sequence across the repetitive R region and into the sequence on either side of the LTRs, this may also cause a break in the assembly. The longer reads offered by new sequencing technologies may help to resolve these issues. Alternatively, Sanger sequencing using knowledge of the ERV insertion sites or primer binding sites at the ends of the LTRs may be useful to target specific insertions.

A number of methods to deal with this have been developed, primarily involving clustering of reads or sequences based on sequence similarity (Li et al., 2005, Macas et al., 2007). These methods still have difficulty assembling the more variable TEs, although it has been suggested that the use of technologies that provide longer reads, such as the 454 system, can improve coverage over these repetitive regions, and improve the contiguity of assemblies across TEs (Metzker, 2009).

Four crocodilian BAC libraries have been developed, representing the three major crocodilian lineages. These include: *C. porosus* (Shan et al., 2009), and *G. gangeticus* (Shan, unpublished), *A. mississippiensis* (Yohn et al., 2005), and, *A. sinensis* (He et al., 2012). Of these, the *C. porosus* and *G. gangeticus* libraries are held at the Mississippi Genome Exploration Laboratory (MGEL) at Mississippi State University (MSU), and were readily accessible for this part of the project. The *C. porosus* BAC library consists of 101,760 individual clones, with an average insert size of 102 kb. It is estimated that the library provides 3.7 fold coverage (3.7×, or 3.7 genome equivalents) of the *C. porosus* genome (Shan

et al., 2009). The *G. gangeticus* library contains 156,000 clones at an estimated 5.67× coverage (Shan, unpublished data).

Despite the importance for the development of TE libraries for functional analysis of genomes, very few broad-scale studies have been carried out to characterise these elements *de novo*, and even fewer have been carried out in vertebrate species. This study establishes the use of BAC libraries as an alternative to whole genome sequencing for the characterisation of ERV sequences, and demonstrates the potential applications of such methodologies for the study of ERV dynamics within the genome. In this study, the *C. porosus* BAC library is used to recover the complete proviral genomes of known crocodilian ERVs and demonstrate the use of conserved retroviral motifs detect. This study provides an overview of the genomic structure of crocodilian ERVs, and novel sequence data from other retroviral coding domains, as well as initial insights into ERV evolution in crocodilians from a genomic perspective.

4.2: Materials and Methods

4.2.1: Preparation and hybridisation of BAC library macroarrays

The hybridisation of short DNA fragments to macroarrays containing fixed BAC DNA is a versatile and efficient method for the isolation of BAC clones containing the regions of interest. In order to screen the BAC libraries for presence of ERV fragments, high density macroarrays were created for the *C. porosus* and *G. gangeticus* BAC libraries as described by Shan et al. (2009). Five complete arrays were created for each of the libraries, using a Genetix QPixII Robot and Amersham Hybond N+ membranes (GE Healthcare). Each membrane was made up of 6 fields in a 3 × 2 array. Each field contained BAC clones from eight plates, double stamped in a 4 × 4 pattern. Each plate occupied a unique pair of locations within each grid to allow for easier identification of the corresponding clone ID in downstream analyses (Figure 4.1). The estimated genome coverage for the five filters is approximately 3.35× coverage for *C. porosus*, and 3.40× coverage for *G. gangeticus*.

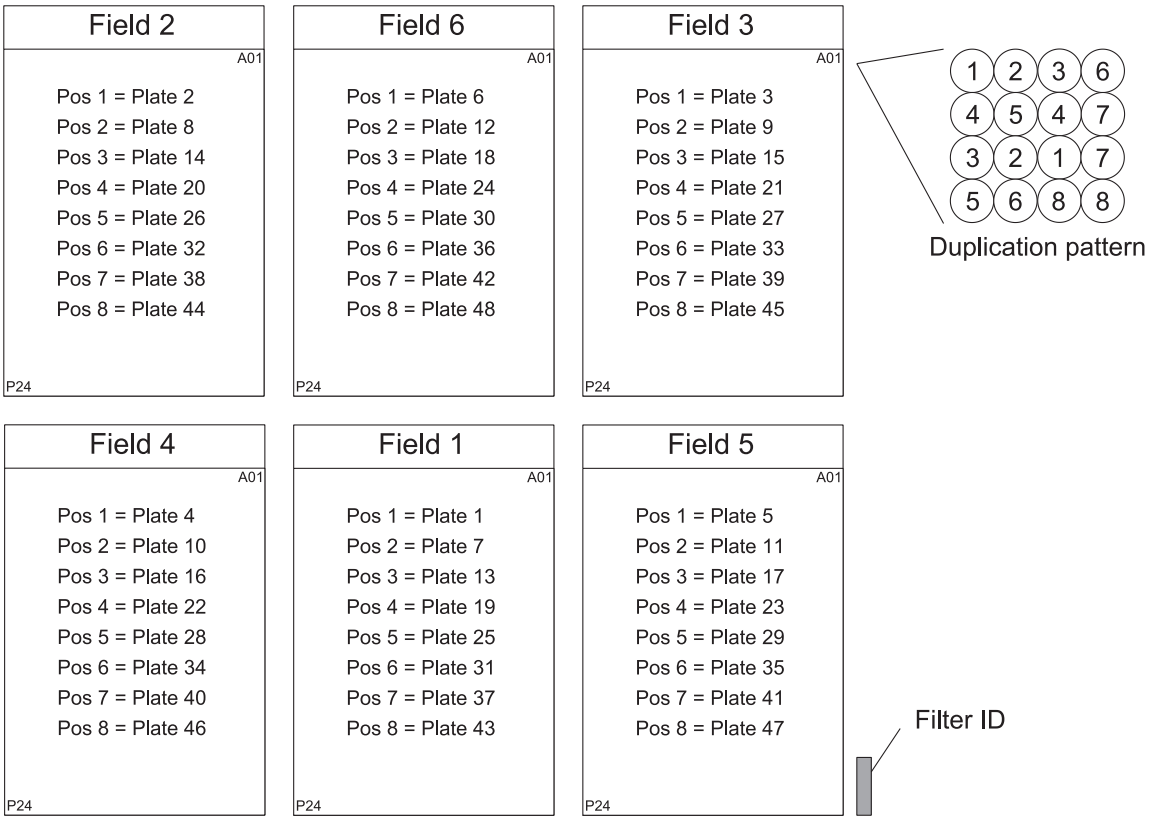


Figure 4.1: Diagrammatic layout of plates on each array, based on a 48 plate array. The six fields are created by the stamping of cells from each plate on the membrane. The positions indicated within the field and the corresponding location within the duplication pattern indicates the location of the colonies for each plate. Well numbers A01 and P24 are indicated in the top right and bottom left corners of each field to show the final orientation of the plate.

DNA fragments encoding *C. porosus* ERV *pro-pol* domains were used to screen the prepared macroarrays for both species to identify BAC clones containing ERV insertions. DNA fragments were used rather than overgo probes as the high levels of sequence variation observed within each ERV lineage made it difficult to identify conserved regions for the design of overgo probes. The DNA probes were selected to represent each of the seven potential lineages of ERVs identified in Chapter 2, and consisted of one sequence encoding an ORF from the ERV1 *Gammaretrovirus*-like clade, four lineages from the ERV4 clade, a more divergent member of this clade, and a single *Epsilonretrovirus* related ERV fragment (sequence haplotypes 13, 58, 77, 80, 96, 107 and 119 from Chapter 2). DNA inserts were excised from the purified plasmids by *Eco*RI enzyme digest and separated by gel electrophoresis. The probe sequences ranged from 665 to 956 nucleotides in length.

Two sets of probes were used for library screening: one consisting of a representative sequence from the *Gammaretrovirus*-like ERV1 fragments (haplotype 77 from Chapter 2), and a second pooled set containing all seven lineages. ERV1 insertions were of particular interest as previous sequence data (Chapter 2) suggested that these were likely to represent a more recent infection, and were therefore more likely to be present as intact proviruses. Haplotype 77 was selected as this fragment encoded an intact ORF. Fifty ng of probe DNA was labelled with radioactive ^{32}P dCTP using the Amersham Megaprime labelling system (GE Healthcare), and excess nucleotides were removed using the QIAquick Nucleotide Removal Kit (Qiagen). The reactions were then halved so that 25 ng of probe DNA was used to hybridise to each of the two libraries. Hybridisation was carried out at 65°C for a high degree of specificity.

Positive clones were identified using the program MacroArray Reader (Philippe Chouvarine, unpublished). This program detects the signals of a BAC clone where the probe has hybridised, and matches this to the duplication pattern used when creating the membranes. This can be manually adjusted to maximise the accuracy of the final output and remove artefacts from strong and weak hybridisation signals. The program then outputs a list of positive wells and the position based on the duplication pattern for each field of the array (see Figure 4.1 for the layout of plates and wells in each field of the array). These can then be matched to the appropriate plates to determine the identity of each positive BAC clone.

Arrays were then analysed for a non-random distribution of hybridisation signals using Holst's "Theorem 2", as it takes into account the non-independence of observations of integration and the possibility that more than one integration event may be present in any given clone. This theorem can be summarised as:

$$mean = Ne^{-npk}; SD = \left(\frac{n^2 N p_k^2}{2}\right)^{0.5}$$

where 'N' is the number of DNA containing clones per array accounting for false positives, 'n' is the expected number of positive clones assuming a random distribution of positively hybridised clones, and 'p_k' is the probability of the probe being in any particular clone. The test value was determined based on the mean number of clones that were expected to lack the region of interest and the number of times that this exceeds the standard deviation of this mean. Statistical significance was determined by comparing the test value to a normal distribution Z-table.

4.2.2: Densitometric analysis of BAC hybridisation

In the absence of a sequenced genome, densitometric analysis was carried out on the hybridised *C. porosus* membranes to provide an estimate of the proportion of the *C. porosus* genome that is made up of ERVs. This is a quantitative method of analysis by which the intensity of a fluorescence signal is calculated relative to other signals within the same image, and related back to a known measurement, in this case, the length of the DNA probe. Once the compete sequence of the ERV insertions has been obtained, this final length can be used to infer the estimated proportion of the genome that is ERV related. Image manipulation and measurements were carried out in ImageJ, and proportions were calculated using a method similar to that described by Magbanua et al. (2011). Briefly, images were cropped to the outer bounds of the stamped area to remove artefacts that may be confused with hybridisation signals. Hybridisation intensities were calculated by subtracting the background intensity from the cropped image and measuring the spot intensity from the perceived centre of each spot. The ImageJ macro is provided in Appendix IV.

The proportion of the genome made up of *pro-pol* containing ERV fragments was calculated based on the assumption that the most intense spots in the membranes were those containing the complete *pro-pol* fragment of approximately 1000 bp. Unlike labelled oligonucleotide systems, the probe labelling system for dsDNA fragments generates shorter lengths of labelled DNA probes from a dsDNA template. As intensity of signal is relative to the number of labelled probe fragments binding to a specific area, more intense signals represent regions were more of the probe DNA has been hybridised. Regions containing the complete probe sequence are more likely to hybridise with more of these probe fragments, and are therefore more likely to show a signal of greater intensity. An assumed target DNA length of 1000bp was selected as this was the median length of ERV *pro-pol* fragments recovered from previous PCR surveys (Chapter 2). The lengths of the hybridised fragments were calculated independently for each membrane to account for the automated background intensity removal. The total proportion of the genome made up of ERVs was then calculated from these values (Table 4.1)

Table 4.1: Calculations for the determination of the ERV fraction of the *C. porosus* genome based in densitometric estimates.

Calculation	Description
A	Spot intensity [a]
B	Assumed length of *pro-pol* fragment
C = A* x B	Intensity calibration factor [b]
D = (A x C)/2	Estimated length of ERV fragment [c]
E = D1+D2+…Dn	Total length of ERVs
F	Estimated genome size
G	Number of clones in library
H	Approximate library genome coverage
I	Number of arrays
J	Number of clones per array
K = (I x J)	Total number of clones in arrays
L = K/G	Fraction of clones in arrays
M = L x H	Array genome coverage
N = M X F	Approximate genome length in arrays
O = E/N	Genome fraction made up of ERVs

[a] Mean intensity value from ImageJ.

[b] Calculated independently for each membrane under the assumption that the highest intensity spot for that membrane represents hybridisation across the entire length of the DNA probe.

[c] Values are divided by 2 to account for the double spotting pattern of the membranes.

4.2.3: Library preparation, sequencing, and assembly

A small number of BAC clones were sequenced to obtain the complete proviral insertions. BAC clones selected for sequencing were those that produced high intensity hybridisation signals using the probe for the ERV1 *Gammaretrovirus*-like sequences. To reduce the chance of signals being false positives, only BAC clones that were positive on both sets of membranes were selected. Positive BACs fitting these criteria were confirmed by PCR using a set of three ERV1 *pro-pol* primers (Appendix I, Table S4.1). PCR was carried out with an annealing temperature of 60°C for all primer pairs. These BACs were then subcultured in LB media overnight before being purified for sequencing. BAC DNA was purified from selected clones using the Qiagen Large-construct kit (Qiagen).

Due to the large number of positive clones and the high probability that clones would contain very similar ERV sequences, two tiers of barcoding were used in order to multiplex the reads.

This involved the ligation of SimpleXT linkers to sheared DNA from each BAC, pooling of these samples, followed ligation of the RL MID tags as the second tier of barcoding (Figure 4.2a).

The purified BAC DNA was sheared to approximately 700 bp using partial enzyme digests. The restriction enzymes *Alu*I, *Mly*I, *Hae*III and *Rsa*I were used as the target cutting sites for these enzymes are known to occur at moderate frequencies throughout most genomes. For this, approximately 1µg of purified BAC DNA was digested for 10 mins at 37°C before the reaction was stopped by heating at 80°C for 20 mins. To enable ligation of the SimpleXT linkers, a single A nucleotide was then ligated to the 3` end of each DNA fragment using 0.25 mM dATP and 2.5 U Klenow Fragment. Ligation was carried out at 37°C for 30 mins, before being stopped by heating to 75°C for 20 mins.

Double-stranded CAG_SXT linkers were created by mixing equal concentrations of the CAG_SimpleXT##_U (upper) linkers with their corresponding CAG_SimpleXT##_Lp (lower) linker with NaCl to a final concentration of 100 mM NaCl. This mix was then heated to 95°C and allowed to cool slowly to room temperature to allow the linkers to anneal. Approximately 0.5 µg of +A-ended DNA from the previous step was ligated with the double-stranded CAG_SXT linkers using T4 DNA ligase (see Appendix I, Table S4.2 for linker sequences, and Appendix I, Table S4.3 for the individual BACs and their corresponding tags). Thermocycling conditions were 22°C for 15 mins, followed by 65°C for 20 mins.

This linker ligated DNA was then amplified and normalised by PCR with the Phos_Pig_CAG primer (/5`Phos/GTTTCAGTCGGGCGTCATCA) in 50 µL reactions. Each reaction contained 5 µL of linker ligated fragments, 1x Phusion HF buffer, 200 µM dNTPs, 25 pmol primer, and 1 U *Phusion* DNA polymerase (Thermo Fisher Scientific). The reaction was carried out by heating to 97°C for 30 s, followed by 18 cycles of: 97°C for 15 s, 55°C for 15 s and 72°C for 30 s. Relative DNA concentrations and fragment sizes were determined by gel electrophoresis on a 1.5% agarose gel containing ethidium bromide.

These reactions were then combined to make five pools of DNA that were used to create the RL MID tagged libraries (see Appendix I, Table S4.2 for the RL MID tag sequences). An 8kb paired-end library consisting of DNA from all sequenced BAC clones was also made. BAC DNA was sequenced at a low coverage on a FLX 454 sequencing platform (Roche). The RL

MID library preparation, paired-end library preparation, and 454 sequencing were at the Georgia Genomics Facility (UGA, Athens, GA).

Sequence reads were assembled using the GS *De Novo* Assembler program from the GS Data Analysis software package (Figure 4.2b) (Roche). All reads were screened to remove sequences from the BAC vector and potential contamination from the *E. coli* strain used in the creation of the libraries. The resulting contiguous sequences (contigs) and scaffolds were then manually checked to minimise the occurrence of assembly artefacts. Sequencing coverage was then estimated for each clone based on the length and number of reads after removal of vector and adapter sequences. Coverage here is defined as the approximate number of times each base in the BAC clone would have been sequenced assuming uniform coverage across the sequenced region (Illumina). Calculations were carried out based on the formula below, where the average BAC insert length was used for the haploid genome length:

$$coverage = \frac{read\ length \times number\ of\ reads}{haploid\ genome\ length}$$

Two sets of sequence assemblies were created: one for each of the BAC clones, and another containing all of the screened reads, including those that could not be assigned to a particular clone. The reads from each of the BAC clones were assembled individually to allow for the detection of the individual ERVs within each clone. Due to multiple levels of tagging, the sequence reads from each of the BAC clones were separated using the program Demuxipy (Brant C. Faircloth; available at https://github.com/faircloth-lab/demuxipy/) prior to the individual assemblies. A combined dataset was also generated to increase the likelihood of reconstructing a complete ERV sequence for further analysis. To confirm the presence of ERV sequences in each of the assemblies, sequence contigs were compared with known ERV *pro-pol* sequences from previous experiments (Chapters 2 and 3) using BLASTN (Altschul et al., 1990, Zhang et al., 2000).

Reads from nine BAC clones sequenced at a high coverage (16× to 102×) for a different project were also assembled and analysed alongside these clones. These clones were of particular interest as they contained sequence data from the MHC region of *C. porosus*. The

MHC region has previously been observed to contain a large number of TEs (Andersson et al., 1998, Edwards et al., 2000, Gasper et al., 2001, Kambhu et al., 1990, Shiina et al., 1999), and consequently, is also a good potential source for ERV proviral sequences.

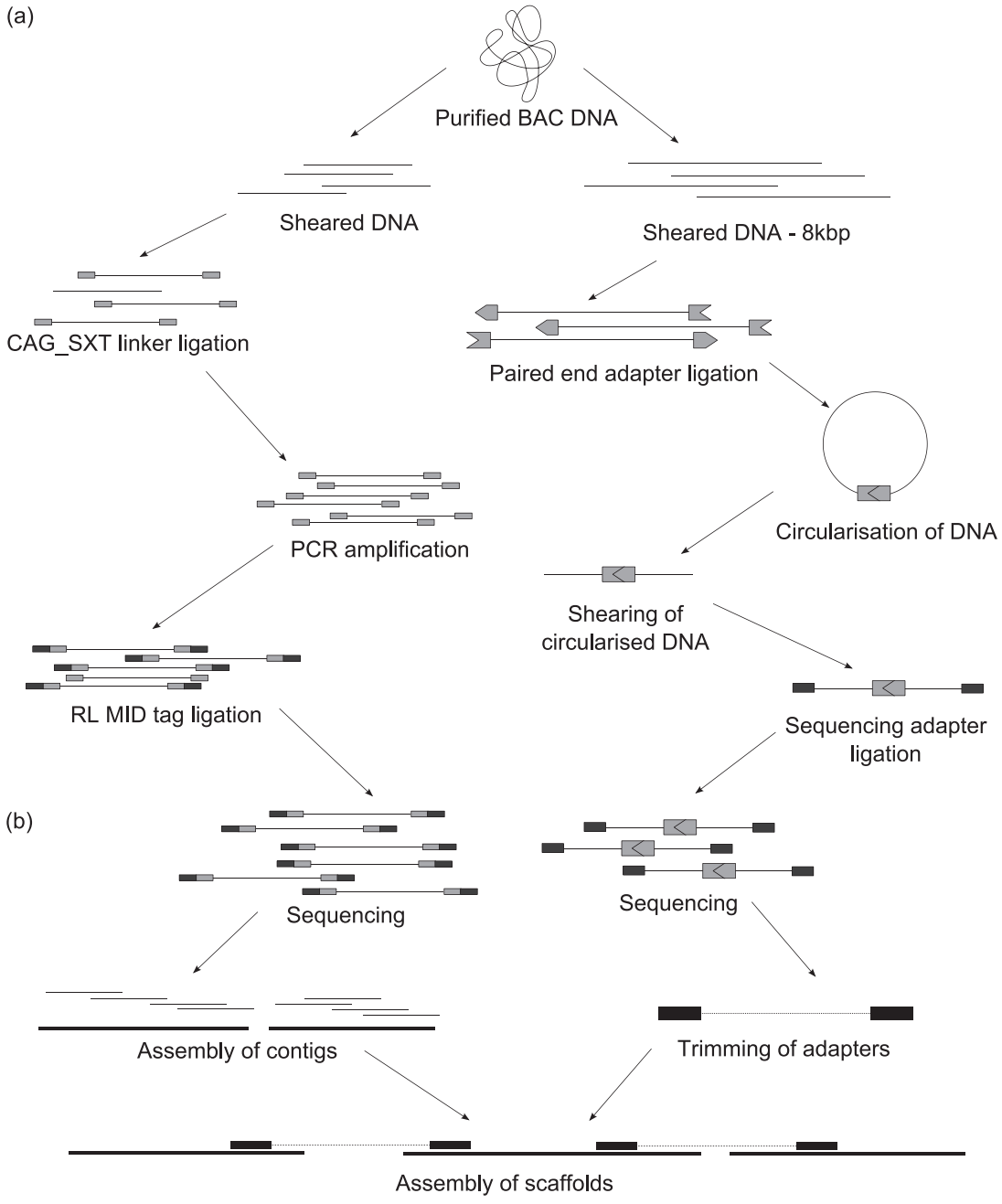


Figure 4.2: Diagrammatic representation of (a) the preparation of libraries for sequencing, and (b) assembly of the resulting sequencing reads.

4.2.4: Characterisation of sequenced ERVs

Characterisation of ERV families

Contigs and scaffolds were then scanned using RetroTector (Sperber et al., 2007) to determine the locations of ERV sequences within the contigs. Briefly, RetroTector is a *de novo* detection program written specifically for the identification of ERVs and similar TEs, that utilises short regions of sequence homology. RetroTector utilises a database of retroviral motifs and attempts to join these into strings or 'putiens' (putative proteins) (Sperber et al., 2007). In addition to identifying the locations of ERVs in the assembled contigs and scaffolds, RetroTector was also used to ascertain the presence of the *gag* and *env* retroviral domains, and the position of the LTRs.

ERV sequences and coding regions detected by RetroTector were extracted from the sequence scaffolds using the sequence locations provided in the program output. Sequences were first aligned using MAFFT (Katoh et al., 2005) as used in previous chapters, then collapsed into haplotypes using FaBox (Villesen, 2007). These haplotypes were then manually checked to ensure that shorter sequences were not truncated copies of longer haplotypes and re-aligned using MAFFT for phylogenetic analysis. This dataset comprised 109 sequences, ranging from 902 to 14472 nucleotides in length.

To investigate the evolutionary relationships between the recovered sequences, sequence clustering and phylogenetic analyses were used. Analysis of sequence clustering was used to identify sequences that were highly similar, and thus likely to be related to each other. The cd-hit-est program in the CD-HIT package (Li and Godzik, 2006) was used to identify sequence clustering and representative sequences at the 95% nucleotide similarity levels using the recommended program settings. Phylogenies were then created to investigate the relationships within and between the sequence clusters. Neighbour Joining and Maximum Likelihood trees were created in MEGA5 (Tamura et al., 2011) as described in previous chapters. As the complete ERV sequences contained both coding domains and LTR regions, comparative analyses on this dataset were carried out on nucleotide alignments only.

A similar analysis was carried out using the *pol* gene of representative sequences from the identified clusters. These alignments and trees were made using the putein sequences identified by RetroTector, a collection of published ERV sequences (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Jern et al., 2005, Martin et al., 1999, Martin et al., 2002), and the

amino acid sequences of representative *pro-pol* fragments from previous chapters. To reduce the number of sequences used in the analysis, the cd-hit program in CD-HIT was used with a 90% similarity cut-off to select representative sequences from the published datasets. This value was chosen as it will collapse very closely related sequences while still maintaining distinct sub-lineages within the ERV clades. The final alignment contained 140 published sequences, 9 sequences from previous chapters, and 49 predicted *pol* sequences from the assembled BAC clones.

Characterisation of ERV genomic structure

Consensus sequences were generated in order to remove spurious variation caused by point mutations and indels. The generation of consensus sequences allows for the reconstruction of what is likely to be the original ERV sequence at the time of replication within the genome. To do this, ERV lineages within the major groups (ERV1, 3, and 4) were identified based on the results of the clustering, strong bootstrap support for the phylogeny (>90%) and visual examination of the alignments for shared regions of sequence. Sequences from the resulting lineages were aligned using MAFFT, and consensus sequences were then created for each of these using BioEdit (Hall, 1999). The resulting sequences were used to identify similar sequences in the original assembly file using BLASTN, and the process was repeated to refine the consensus sequences.

The NCBI ORF-finder tool (http://www.ncbi.nlm.nih.gov/projects/gorf/orfig.cgi) was then used to identify ORFs within the consensus sequences. ORFs longer than 200 nucleotides were compared with the GenBank database using BLASTP (Altschul et al., 1997) to determine if these were retroviral related. Intact retroviral ORFs range from approximately 1000 bp in length (*pro*), through to 4000 bp (epsilonretroviral *env* domain), although the length of these is likely to substantially shorted in ERVs due to the accumulation of premature stop codons. The results of this analysis were further corroborated by padding the resulting sequences with 'neutral' or non-ERV sequence from the assembled contigs and RetroTector was used again to identify the conserved domains, LTRs, and promoter regions within the consensus retroviral sequences.

Investigation of an unusual ERV sequence

An ERV4 lineage containing an additional ORF in addition to the four major retroviral domains was identified by the analyses described above. While ERVs have previously been

shown to capture mRNA transcripts of host genes, this usually occurs at the expense of a viral domain. This lineage is unusual as the major coding domains appear to be intact. To confirm the presence of these sequences, a BLAT alignment (Kent, 2002) was used to locate similar regions in the *C. porosus* genome assembly (available at http://crocgenome.hpc.msstate.edu/gb2/gbrowse/). These genome assemblies became available during the course of experimental work in this chapter, and will be discussed in more detail in the following chapters.

To determine the gene transcript that was captured, the sequence was compared with the GenBank database using BLASTX (Altschul et al., 1997). Genes related to the recovered ORF were then retrieved from the crocodilian genomes and UniProt for phylogenetic analysis to determine whether the transcript was crocodilian in origin. Amino acid translations of the crocodilian sequences were made in using the standard vertebrate codon table for nuclear DNA. Sequence alignments of the resulting amino acid translations were created using MUSCLE (Edgar, 2004), and phylogenetic trees were created using Maximum Likelihood algorithms as described previously. Two datasets were created in this way: the first dataset comprised the sequence from the ERV and four predicted transcripts retrieved from the crocodilian genomes and two homologous sequences from *Gallus gallus* (chicken; NCBI reference ID: NM_205130.1 and NM_001105315.1), while second set included the four crocodilian sequences, the ERV ORF, and 30 sequences from other vertebrates (Table 4.2), with the sequences from *Gasterosteus aculeatus aculeatus* (three-spined stickleback) used as an outgroup.

Table 4.2: UniProt ID, and scientific and common names of additional species used for phylogenetic analysis of the novel ERV ORF.

UniProt ID	Scientific name	Common name
F6SD93 F6SDB0	*Macaca mulatta*	Rhesus macaque
K7D1D2 H2Q6K8	*Pan troglodytes*	Chimpanzee
G1R0E6	*Nomascus leucogenys*	White-cheeked gibbon
P21583	*Homo sapiens*	Human
G3S4K2	*Gorilla gorilla gorilla*	Gorilla
H2NI76	*Pongo abelii*	Sumatran orangutan
Q95MD2	*Equus caballus*	Horse
Q28132	*Bos taurus*	Cow
Q95M19	*Capra hircus*	Goat
Q29030	*Sus scrofa*	Pig
P79169	*Felis catus*	Domestic cat
Q06220	*Canis familiaris*	Domestic dog
Q95N18	*Mustela vison*	American mink
P20826	*Mus musculus*	House mouse
P21581 Q54A14	*Rattus norvegicus*	Brown rat
F7EP53	*Monodelphis domestica*	Opossum
G3VV70 G3VV69	*Sarcophilus harrisii*	Tasmanian devil
F7C5I3	*Ornithorhynchus anatinus*	Platypus
Q09108	*Gallus gallus*	Chicken
Q90314	*Coturnix coturnix japonica*	Japanese quail
K7FR23	*Pelodiscus sinensis*	Chinese softshell turtle
Q28DP4 B4F6K4	*Xenopus tropicalis*	Western clawed frog
Q7ZXV0	*Xenopus laevis*	African clawed frog
A8WC00 A8WC01	*Gasterosteus aculeatus aculeatus*	Three-spined stickleback

4.3: Results

4.3.1: Hybridisation patterns and observations

The ERV DNA probes hybridised successfully in both of the *C. porosus* libraries. Across the five membranes, 4559 positive clones were detected using the pooled probe set, and 3710 positive clones were detected using the probe specific to the *Gammaretrovirus*-like lineage. Forty eight clones produced very strong hybridisation signals with both probe sets and were subsequently sequenced. The pooled probe set also produced positive signals in 223 clones of the *G. gangeticus* library. No significant hybridisation was detected in the *G. gangeticus* library when using the *Gammaretrovirus* specific probe (Figure 4.3).

The genomic distribution of ERVs in *C. porosus* was determined to be highly non-random (Test statistic = 1082.408, $P << 0.001$) with the actual number of clones that were positive for ERV sequence being significantly higher than the expected number of clones given a random distribution of ERVs (See '*D*' and '*n*' in Table 4.3 below). This suggests that there is some bias in the location of the detected ERV insertions.

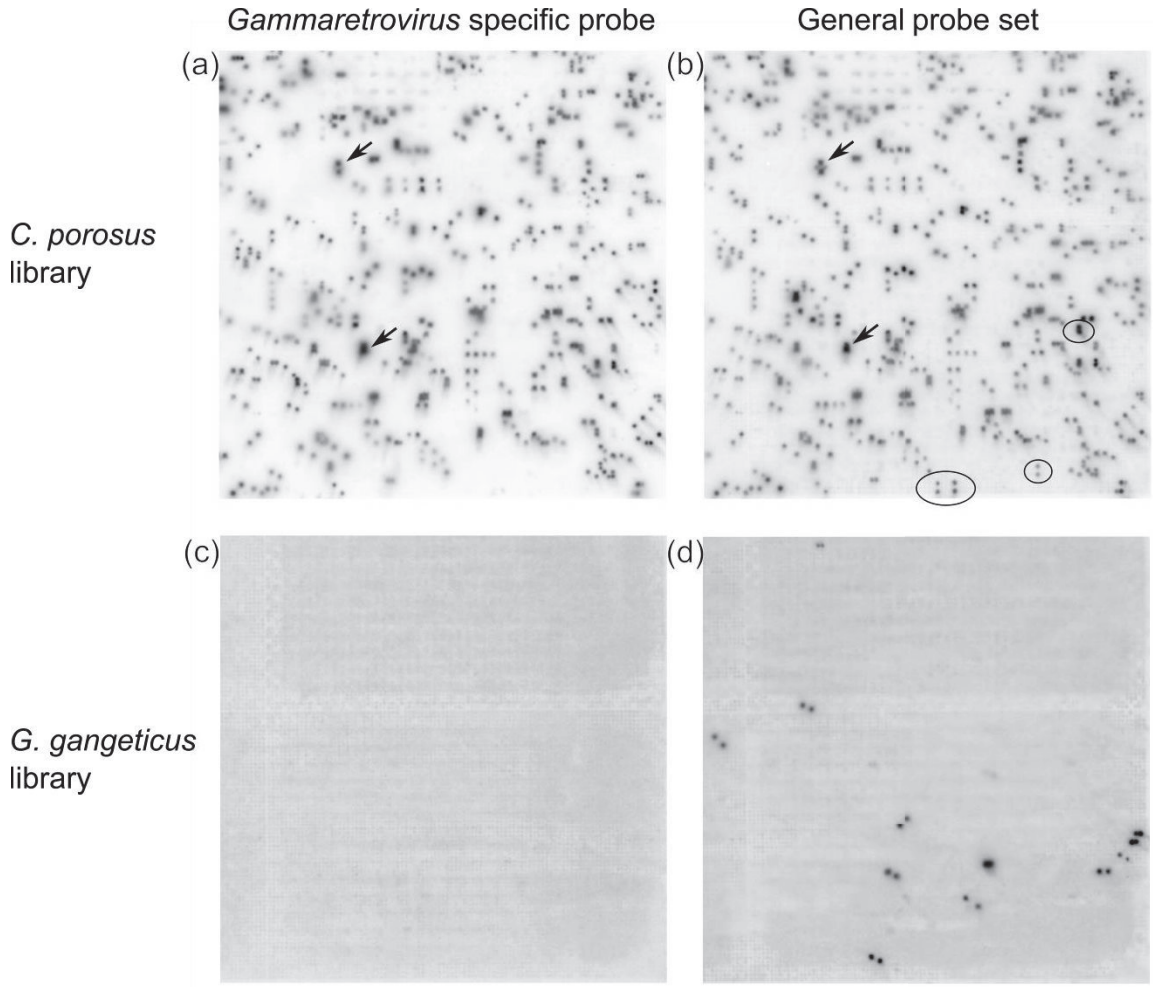


Figure 4.3: Representative fields showing the difference between the probe sets and between species. Parts (a) and (b) above are from the *C. porosus* library. Black circles indicate examples of BAC clones that do not contain the *Gammaretrovirus*-like sequence. The arrows indicate clones that showed strong hybridisation with both probe sets. Parts (c) and (d) show representative regions of the *G. gangeticus* library. Note the lack of hybridisation in (c) and the much smaller number of positive clones in (d) compared to the *C. porosus* library (parts (a) and (b)).

Table 4.3: Values and calculation of the probablility of a non-random ERV distribution using Holst's Theorum 2 as described in Section 4.2.1 above.

	Description	**Calculation**	**Value**
N	DNA containing clones/array	A*B	18,192.38
n	Expected number of positive clones assuming random distribution	C*D	3,045.28
pk	Probability of an element being in any particular clone	1/N	0.00
SD	Standard Deviation		15.96
A	Total number of clones in one array		18,432.00
B	Proportion of DNA containing clones in library		0.99
C	Genome equivalent represented by a single array		0.67
D	Total number of positive clones		4,559.00
E	Average number of positive clones per array	B/Z	911.80
X	Average insert size		102,000
Y	Estimated genome size		2,778,000,000
Z	Number of arrays screened		5

4.3.2: Densitometric estimates of ERV complement

Densitometric estimates suggest that approximately 0.01% of the *C. porosus* genome is related to ERV *pro-pol* sequence (See '*O*' in Table 4.4 below), based on the lineages identified in previous chapters. Assuming that ERVs containing the *pro-pol* gene region are likely to encode other regions of the retroviral genome (for example, the *env* and *gag* gene regions, and 5' and 3' LTRs), these data can be extrapolated to calculate the approximate proportion of the genome that is made up of complete proviral insertions. Based on an average proviral length of 6 to 10 kb, this proportion ranges from 0.07–0.1% of the genome.

Table 4.4: Densitometry calculations from Table 4.4 in Section 4.2.2, extended to show the values obtained from each stage of the calculation

Calculation	Description	Value
A	Spot intensity [a]	Variable
B	Assumed length of *pro-pol* fragment	1000
C = A x B*	Intensity calibration factor [b]	Variable
D = (A x C)/2	Estimated length of ERV fragment [c]	Variable
E = D1+D2+...Dn	Total length of ERVs	1,343,749.439
F	Estimated genome size	2,780,000,000
G	Number of clones in library	101,760
H	Approximate library genome coverage	3.7
I	Number of arrays	5
J	Number of clones per array	18,432
K = (I x J)	Total number of clones in arrays	92,160
L = K/G	Fraction of clones in arrays	0.905660377
M = L x H	Array genome coverage	3.350943396
N = M X F	Approximate genome length in arrays	9,315,622,642
O = E/N	Genome fraction made up of ERVs	0.000144247

[a] Mean intensity value from ImageJ.

[b] Calculated independently for each membrane under the assumption that the highest intensity spot for that membrane represents hybridisation across the entire length of the DNA probe.

[c] Values are divided by 2 to account for the double spotting pattern of the membranes.

4.3.3: Assembly statistics of the sequences BAC clones

Individual assemblies – DNA probe

Each of the BAC clones was sequenced to a reasonable depth, although overall coverage was low. A total of 242,819 reads were generated from the 48 BAC clones, with an average read length of 224.24 bp. Of this, about half of the reads had no recognisable SXT tag, and could not be incorporated into the final individual clone assemblies. The PE library returned 184,982 reads which produced 214,917 single reads after trimming adapter sequence, including 60,349 paired reads. After screening, a total of 406,740 single reads were used in the assemblies (including trimmed paired-end reads) with an average read length of 179.43 bp.

Overall, the sequence coverage of each BAC clone was low, at an average of 4× coverage across the 48 clones sequenced, and 11× coverage if paired-end reads were equally distributed among the clones. One clone, 205-D13 did not assemble due to unknown reasons and was removed from further analysis. Average contig length across the remaining 47 clones was 1,710.19 bp, with an average of 2,624.38 contigs generated per clone. The average scaffold length was 26,767.28 bp, and the average number of scaffolds created was 117.79 (see Appendix I, Table S4.4 for max, min lengths and total number of contigs and scaffolds per clone). ERV sequences were detected in all of the 47 clones sequenced. BLAST searches revealed that all of the assembled clones contained contigs with similarity to the probe sequence. In addition, 40 of the 47 clones had scaffolds that contained the probe sequence.

Pooled assemblies – DNA probe

Unfortunately, the pooling of all 454 reads prior to assembly did not improve the length or contiguity of the ERV sequence assemblies. The pooled assembly produced a total of 5374 contigs and 314 scaffolds. The longest contig was 9690 bp, with an average contig size of 1275 bp. The largest scaffold generated was also 9690 bp. Average scaffold size was 3066 bp. BLAST searches did recover contigs and scaffolds containing the probe sequence, but the majority of these were from contigs that were shorter than the query sequence.

MHC related BACs

A total of 233,759 reads were generated from the nine MHC related BAC clones. The average length of these reads was 309.564 bp. Estimated coverage for each clone ranged from 16× to 102×. The average contig length across the nine clones was 5091.556 bases, with an average of 58 contigs generated per clone (see Appendix I, Table S4.4 for the individual contigs). The longest contig generated was 93,750 bp. Paired-end libraries were not made for these clones.

4.3.4: Classification of ERV genomic sequences

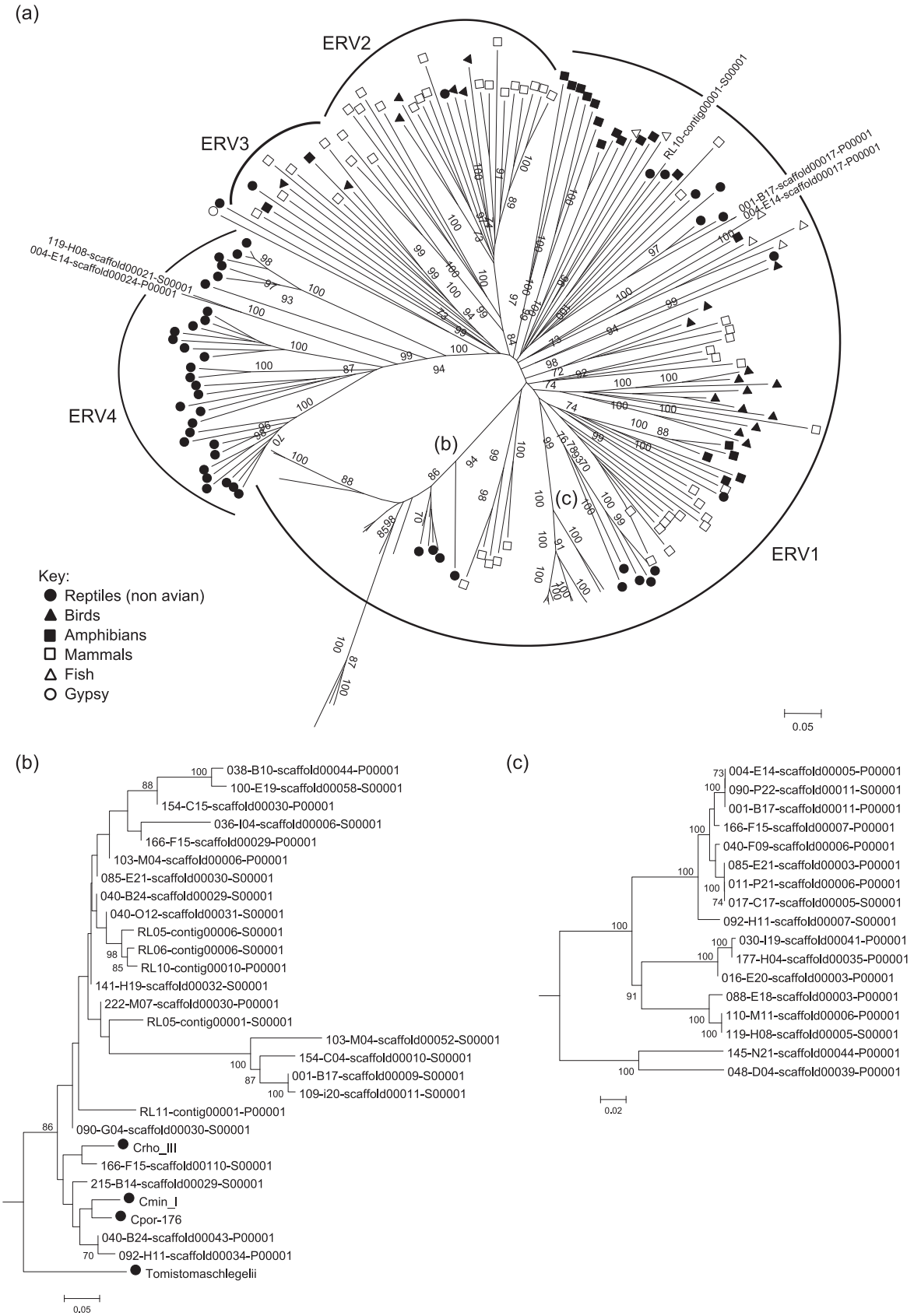
RetroTector identified 281 ERV insertions of varying degrees of completeness. Of these, 194 had identifiable *gag* domains, 147 had the *pro* domain, 197 had detectable *pol* domains, and 118 had detectable *env* domains. A total of 50 of these insertions had detectable LTR regions. The length of the detected insertions (complete and incomplete) ranged between 902 and 14472 nucleotides. The detected *gag* and *pro* domains were highly conserved across the sequences recovered, with 27 and 12 nucleotide sequence haplotypes recovered from each set respectively. These could be assigned to 15 and 6 groups at 80% nucleotide similarity. The *pol* and *env* domains were much more variable, with 80 and 67 haplotypes recovered, forming 21 and 12 groups of sequences respectively at 80% nucleotide sequence similarity.

Overall, the detected *gag* and *env* domains showed very little similarity to known ERVs from other species, although conserved regions were evident. The *pol* region, on the other hand, showed similarity to the previously described crocodilian ERV fragments. This allowed the complete ERV sequences to be assigned to either the described ERV clades, or characterised as novel sequences. Classification and conservation of sequence motifs will be discussed for each major group of sequences below.

As expected a number of ERV sequences were grouped with the *Gammaretrovirus*-like ERV1 sequences described in Chapter 2, and by Jaratlerdsiri et al. (2009). However, a number of sequences were recovered that were assigned to other ERV clades. These included sequences with similarity to the *Epsilonretrovirus*-like fragments, additional divergent ERV1 lineages that are, so far, undescribed, and ERV4 sequences (Figure 4.4; see also Appendix I, Table S4.5). These will be described individually below. For one of these ERV1 lineages,

described below in Section 4.3.5, the probable amino acid sequence could not be predicted from the detected motifs despite showing similarity to ERV1 sequences at the nucleotide level. Consequently, this group of sequences was not included in the phylogenetic analysis.

Figure 4.4 (next page): Five lineages of ERVs were recovered from the *C. porosus* genomic BAC library. The Neighbour Joining phylogeny was generated from the putative amino acid translations of the *pol* domain and ERVs from other species. Sequence names indicate the BAC assembly, contig, and chain ID of the detected insertion. Sequence IDs for published and previously described sequences have been omitted for clarity. Part (a) is the complete tree, (b) and (c) show subtrees containing the majority of the recovered ERVs. The location of these subtrees within the main tree is indicated. The scale bars indicate branch lengths and values on the branches indicate bootstrap support greater than 70%.


(a)
ERV2
ERV3
ERV4
ERV1
(b)
(c)
RL10-contig00001-S00001
001-B17-scaffold00017-P00001
004-E14-scaffold00017-P00001
119-H08-scaffold00021-S00001
004-E14-scaffold00024-P00001
Key:
Reptiles (non avian)
Birds
Amphibians
Mammals
Fish
Gypsy
0.05
(b)
038-B10-scaffold00044-P00001
100-E19-scaffold00058-S00001
154-C15-scaffold00030-P00001
036-I04-scaffold00006-S00001
166-F15-scaffold00029-P00001
103-M04-scaffold00006-P00001
085-E21-scaffold00030-S00001
040-B24-scaffold00029-S00001
040-O12-scaffold00031-S00001
RL05-contig00006-S00001
RL06-contig00006-S00001
RL10-contig00010-P00001
141-H19-scaffold00032-S00001
222-M07-scaffold00030-P00001
RL05-contig00001-S00001
103-M04-scaffold00052-S00001
154-C04-scaffold00010-S00001
001-B17-scaffold00009-S00001
109-i20-scaffold00011-S00001
RL11-contig00001-P00001
090-G04-scaffold00030-S00001
Crho_III
166-F15-scaffold00110-S00001
215-B14-scaffold00029-S00001
Cmin_I
Cpor-176
040-B24-scaffold00043-P00001
092-H11-scaffold00034-P00001
Tomistomaschlegelii
0.05
(c)
004-E14-scaffold00005-P00001
090-P22-scaffold00011-S00001
001-B17-scaffold00011-P00001
166-F15-scaffold00007-P00001
040-F09-scaffold00006-P00001
085-E21-scaffold00003-P00001
011-P21-scaffold00006-P00001
017-C17-scaffold00005-S00001
092-H11-scaffold00007-S00001
030-I19-scaffold00041-P00001
177-H04-scaffold00035-P00001
016-E20-scaffold00003-P00001
088-E18-scaffold00003-P00001
110-M11-scaffold00006-P00001
119-H08-scaffold00005-S00001
145-N21-scaffold00044-P00001
048-D04-scaffold00039-P00001
0.02

4.3.5: Characterisation of ERV genomic sequences

Many structural and conserved domains were detected in the assembled ERV sequences despite the fragmented nature of the assemblies and the apparent divergence of some of the ERV lineages identified. These are outlined below for each of the major ERV clusters.

ERV1 *Gammaretrovirus*-like

As expected, these were the most abundant sequences recovered from the BAC assemblies. RetroTector recovered 129 copies of this ERV from the low coverage assemblies, and nine from the MHC BACs. These could be assigned to 46 haplotypes. The majority of sequences retrieved from the low coverage sequencing were partial sequences consisting of at least one internal coding domain. LTRs were detected in insertions from two of the MHC related BAC clones. All major ERV domains were detected, as well as a putative PPT. No additional coding ORFs were present in any of these sequences.

Within the internal coding regions, RetroTector was able to detect conserved motifs that showed similarity to exogenous *Gammaretroviruses* and Class I ERVs. Motifs from these regions were generally present within the same reading frame, and most of stop codons identified were in locations consistent with the *gag-pro* boundary, and the end of the *pro-pol* region (Figure 4.5). This is typical of *Gammaretroviruses* and *Epsilonretrovirses* with these regions encoded by a single ORF immediately following the *gag* ORF, and in the same reading frame. The *env* domain was also identifiable through homology with other ERV *env* regions, although conserved motifs were not always detected. The positions of the predicted coding domains within these sequences also confirm that this is likely to be related to exogenous *Gammaretroviruses*.

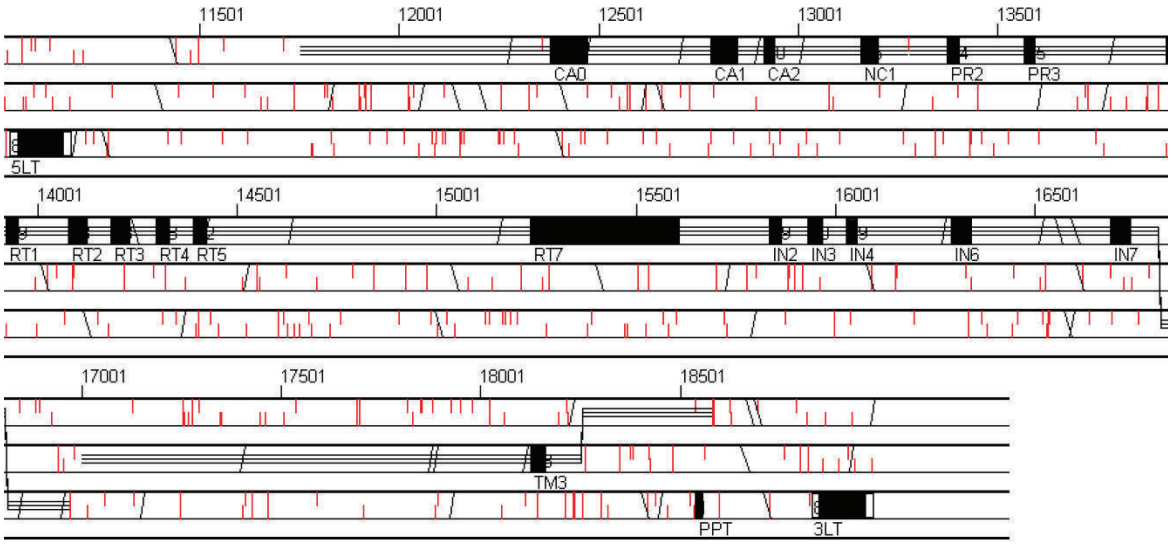


Figure 4.5: Graphical representation of the RetroTector output for the ERV1 G*ammaretrovirus*-like consensus sequence, showing the location and reading frame of putative ORFs, and the location of the detected conserved domains. Numbers above each row indicate nucleotide positions, black boxes indicate detected motifs, vertical lines show the detected stop codons, diagonal lines show potential splice acceptors and donors, and the three horizontal lines indicate predicted puteins.

ERV1 *Epsilonretrovirus*-like

Sequences from this lineage were recovered from each of the low coverage BACs, with the 47 sequences belonging to 4 haplotypes. The majority of these sequences were identical, with 44 sequences forming one haplotype. A comparison of these haplotypes revealed that they were identical except for the number of 'N' bases inserted by the incorporation of the paired reads, suggesting that these may be derived from a single sequence.

This sequence encoded part of the internal portion of an *Epsilonretrovirus*-like ERV, with BLAST searches showing high degree of similarity to previously identified crocodilian *pro-pol* fragments. RetroTector corroborated this, with conserved motifs detected from the *gag*, *pro* and *pol* domains (Figure 4.6). The motifs for each individual domain were mostly encoded in the same reading frame, although these varied between the haplotypes. Stop codons were mostly present at the end of the *gag* domain and within the *pro* domain, although the fragment of the *pol* domain that was recovered appeared to be intact. No LTR regions or additional ORFs were detected for this lineage.

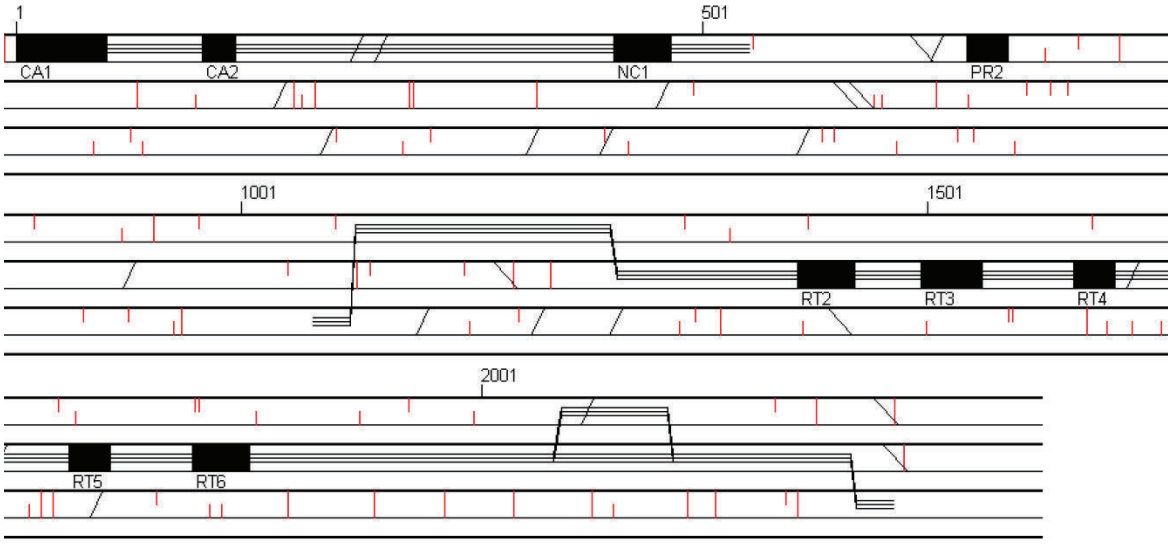


Figure 4.6: Graphical representation of the RetroTector output for the ERV1 *Epsilonretrovirus*-like consensus sequence, showing the location and reading frame of putative ORFs, and the location of the detected conserved domains. Symbols are the same as Figure 4.5.

Novel ERV1 sequences

An additional 54 ERV sequences were recovered that shared very little similarity to the other two ERV1 lineages described above. Comparisons with ERV sequences in GenBank and RepBase suggested these sequences formed three additional ERV1 lineages. The first of these novel lineages was unique to the low coverage BAC clones, and comprised of 47 sequences and 13 sequence haplotypes. All insertions were incomplete, with partial *env* domains detected, and no LTRs (Figure 4.7a). Two novel lineages were identified within the MHC BACs, and were made up of one and six sequences respectively. All of these insertions appeared to be unique and contained predicted LTR regions, although the internal domains were not always complete (Figure 4.7b).

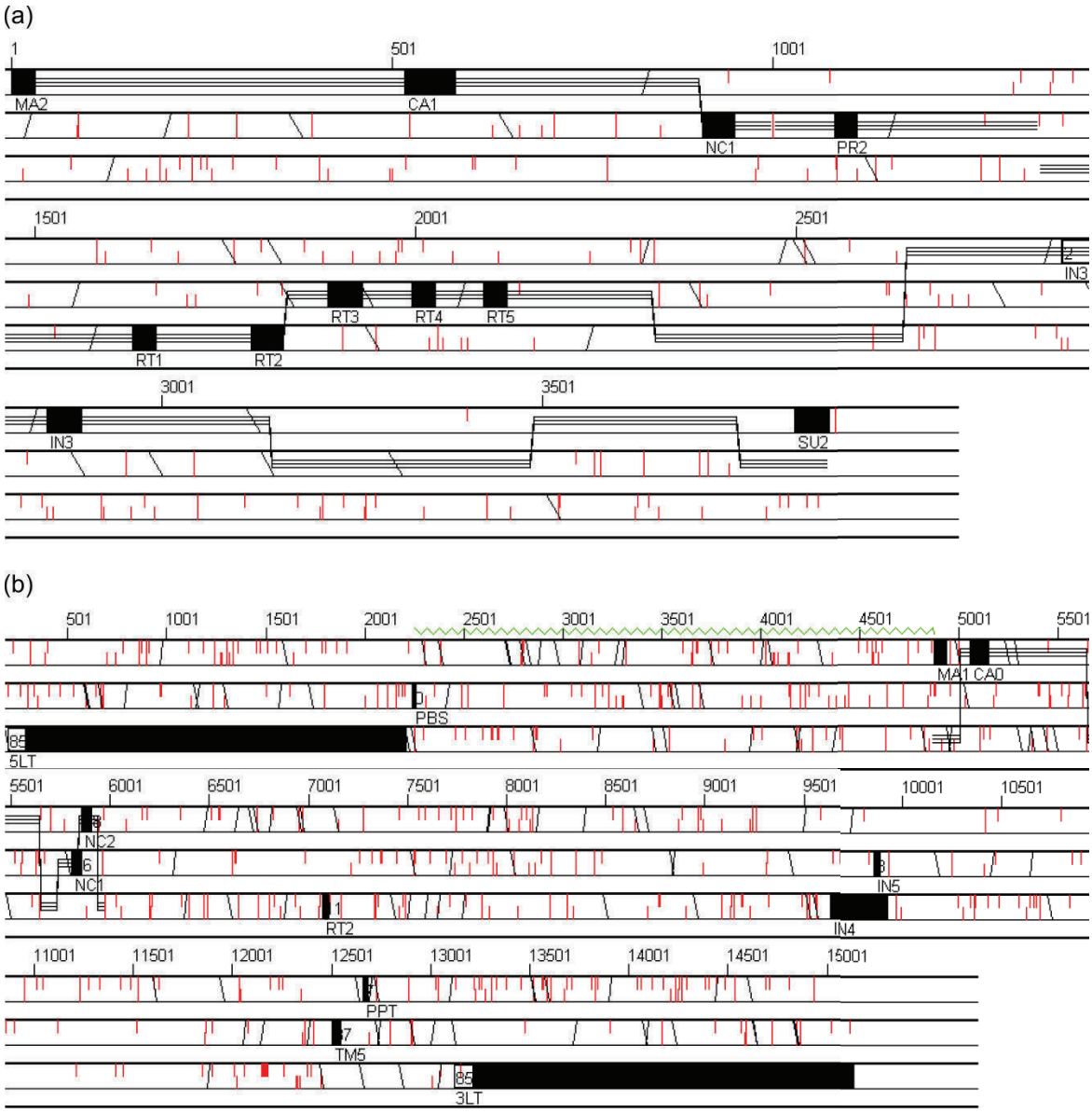


Figure 4.7: Graphical representation of the RetroTector output showing the location and reading frame of putative ORFs, and the location of the detected conserved domains in the novel ERV1 lineages identified in this study. Part (a) shows the predictions from the consensus sequence from the novel low coverage BAC ERV lineage, and part (b) is from the MHC related ERV lineage. Symbols are the same as Figure 4.5.

ERV4 lineage

A large number of insertions that show similarity to crocodilian ERV4 fragments were also identified. With the exception of the LTRs of three sequences, the 41 sequences were identical apart from the number of N bases incorporated during assembly of the scaffolds. Analysis with RetroTector suggested that these sequences also encoded entire ERVs, with

LTRs, PBS, PPT, and motifs from the *gag*, *pro*, *pol*, and *env* domains detected, although these motifs were found across all three reading frames (Figure 4.8).

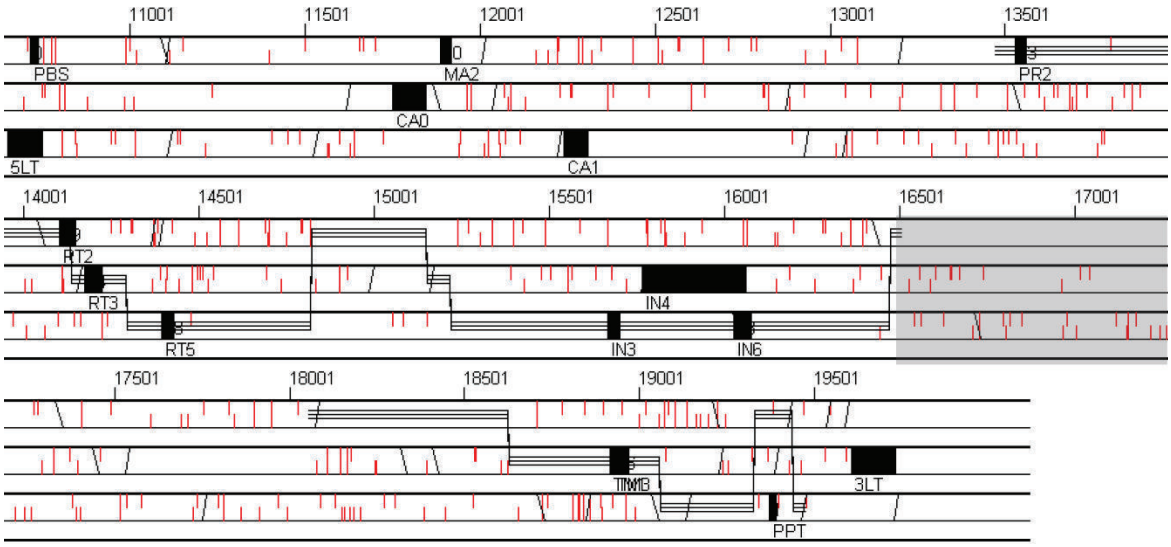


Figure 4.8: Graphical representation of the RetroTector output for the ERV4 consensus sequence, showing the location and reading frame of putative ORFs, and the location of the detected conserved domains. The light grey region indicates the approximate region of the additional ORF that was detected in these sequences. Symbols are the same as Figure 4.5.

An additional ORF was detected in these sequences between positions 5404 and 6231 of the consensus sequence, immediately after the *pol* domain (approximate region shaded in Figure 4.8). This ORF was 825 nucleotides in length (275 amino acids), and showed similarity to mRNA copies of vertebrate KIT-ligand genes. Pairwise genetic distances and phylogenetic analysis of the amino acid sequence of this gene compared with predicted KIT-ligand genes in the crocodilian genomes show that the crocodilian KIT-ligands are more closely related to each other than to the ERV copy. When compared with KIT-ligand transcripts from other species, all the crocodilian sequences, including the ERV sequence, formed a monophyletic sister clade to avian KIT-ligand sequences (Figure 4.9).

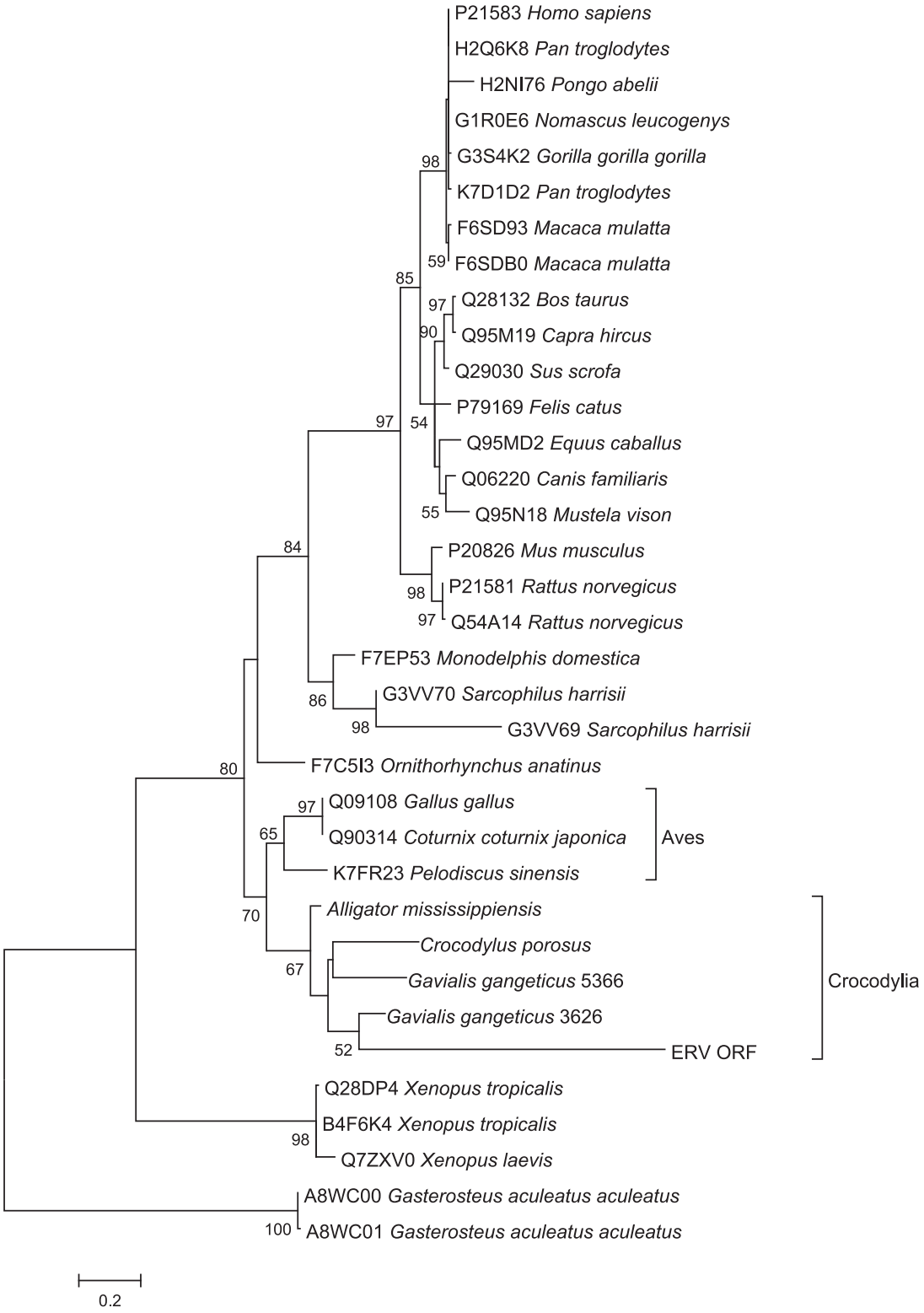


Figure 4.9: The captured KIT-ligand ORF clusters with crocodilian KIT-ligand sequences. The Maximum Likelihood phylogeny was generated from putative amino acid translations of crocodilian KIT-ligand genes and homologs from other species. The scale bar indicates branch length and values on the branches indicate bootstrap support greater than 50%.

4.4: Discussion

4.4.1: Crocodilians may have a lower abundance of ERVs compared to other vertebrates

Overall, *pro-pol* containing ERV lineages make up an estimated 0.1% of the total crocodilian genome. These estimates are low compared with many sequenced species (Table 4.5), but not unexpected since the full extent of ERV integration in the crocodile genome is unknown. More accurate estimations of the ERV content from the sequenced genomes and the reasons for these differences are provided in Chapters 5 and 6. Despite this, these values are similar to predicted ERV abundance in *Xenopus tropicalis* (western clawed frog) and *Canis familiaris* (dog), suggesting that these values, although low, are not unusual among vertebrates.

Table 4.5: Comparison of estimated ERV content in the genomes of *C. porosus* and a selection of model organisms.

Species	Common name	% ERVs in genome	Reference
Danio rerio	Zebrafish	0.8%	Barrio et al. (2011)
C. porosus	Saltwater Crocodile	0.07–0.1%	
Gallus gallus	Chicken	1.3–2.0%	Huda et al. (2008)
Anolis carolinensis	Green Anole	3.0%	Alfoldi et al. (2011)
Xenopus tropicalis	Western Clawed Frog	0.12%	Hellsten et al. (2010)
Monodelphis domestica	Opossum	2.0%	Barrio et al. (2011)
Canis familiaris	Dog	0.15–2.0%	Barrio et al. (2011)
Mus musculus	Mouse	2.0%	Barrio et al. (2011)
Homo sapiens	Human	0.8%	Blikstad et al. (2008)

While the exact biological reasons behind this low ERV complement are not obviously apparent, it may be a consequence of environmental conditions reducing capacity for retroviral proliferation within a population. While not much is known about the stability of crocodilian viruses once shed into the external environment, both temperature and pH are known to affect the stability of virion particles (Beer et al., 2003, Higashikawa and Chang, 2001). Furthermore, retrovirus virions may also have reduced survival in damp and aquatic environments (Moore, 1993), although it has been suggested that some retroviral lineages have developed adaptations for virion persistence in aquatic and semi-aquatic environments. One such example is the extended hydrophobic tail and lack of a cytoplasmic domain in the

env domain of WDSV and WEHV retroviruses (LaPierre et al., 1999). However, due to the lack of a predicted *env* domain in the closest related crocodilian ERVs, similar characteristics cannot be investigated at this stage.

Likewise, crocodile biology may also play a role as cellular characteristics such as divergence or absence of compatible cellular receptors and intracellular restrictions may inhibit retroviral entry into cells (Bishop, 1978), thus preventing retroviral replication. Additionally, the data generated by hybridisation of the *G. gangeticus* BAC library suggest that this species appears to harbour a much different ERV complement to *C. porosus*. The markedly different hybridisation patterns suggest that these species are hosts to very divergent ERV complements, or that the environmental and cellular conditions have resulted in varying levels of prevalence of each ERV lineage. However, there is currently very limited knowledge of crocodilian cellular biology, which prohibits further speculation on the likely role of cellular factors in retroviral, and consequently, ERV proliferation at this stage.

It should also be noted that these estimates only encompass the proportion of the ERVs in the genome that have retained recognisable *pro-pol* regions. Thus, estimates based on this are likely to underestimate the total proportion of the genome that is ERV related, as there may also be more degraded copies of ERVs that have not been detected using this screening methodology. There may also be a number of ERV lineages present in crocodilians that have not been detected in previous PCR screens, and thus will not be represented in the current set of probes.

4.4.2: Crocodilian ERVs appear to show preferential insertion patterns

Based on the values obtained using Holst's Theorem 2, ERVs in the *C. porosus* genome appear to display preferential insertion patterns, with the hybridisation patterns suggesting a highly non-random distribution of ERVs across the genome. This is typical for ERVs, and there are many documented instances of preferences for particular insertion sites or preferential insertion into specific regions of DNA. The physical location of the DNA within the helix can also play a part, with some ERVs, such as MLV, shown to insert primarily in DNA situated on the outer face of nucleosomal DNA (Muller and Varmus, 1994).

Retroviruses generally favour insertion into transcriptionally active regions of DNA, although it is thought that high transcriptional activity actually has an inhibitory effect on the likelihood of ERV integration into that region (Maxfield et al., 2005). ERVs from different genera also display different preferences for insertion within transcriptionally active zones. The *Lentivirus*, Simian Immunodeficiency Virus (SIV), for example, has been shown to integrate in gene dense regions, primarily in the intronic regions of genes (Hematti et al., 2004). Human immunodeficiency virus (HIV) has also been shown to demonstrate similar preferences, with integration predominantly within transcription units (Bushman et al., 2005). On the other hand, the *Gammaretrovirus*, MLV, shows a weak preference for the regions surrounding transcription start sites (Bushman et al., 2005, Hematti et al., 2004). While the assemblies of the BAC clones do not permit further investigation of the surrounding genomic regions, the annotation of the genome sequences may provide further insights.

A number of ERV insertions were also detected within MHC related BAC clones. This was unsurprising, as the accumulation of ERVs and TEs within the MHC region of other vertebrates has previously been noted (Andersson et al., 1998, Edwards et al., 2000, Gasper et al., 2001, Kambhu et al., 1990, Shiina et al., 1999). Studies into the location of these insertions within the MHC suggest that TEs may play a role in promoting the generation of diversity within the MHC region. It has been suggested that ERVs and TEs maintained in this region provide sites for recombination to occur, driving duplication or expansion events (Balakrishnan et al., 2010, Dawkins et al., 1999, Kulski et al., 1999).

4.4.3: ERV1 may be the predominant ERV family in *C. porosus*

The results generated by hybridisation-based screening of the *C. porosus* genome suggest that ERV1 insertions are likely to be the predominant ERV lineage present in this species. Over 80% of the insertions detected using the general crocodilian ERV probe set were also detected using the probe specific to the ERV1 *Gammaretrovirus*-like lineage described in Chapter 2. The higher relative abundance of these ERV1 insertions compared to ERV4 insertions differs from what is suggested by the results of the previous PCR-based surveys. However, the differing hybridisation patterns suggest that the hybridisation conditions were stringent enough to limit detection to those clones containing ERV DNA with high levels of

homology to the probe sequences. This means that it is unlikely that this apparent difference is due to cross hybridisation between the *pro-pol* region of the different ERV lineages.

Likely, this is a consequence of using a different set of techniques, and a different focus between the two surveys. The PCR survey implemented previously was less stringent since it was a general survey with the primary aim of recovering a wide variety of ERV lineages. The hybridisation used here was somewhat more stringent as it used the ERV DNA fragments as probes. This eliminates bias due to reaction conditions or abundance of specific primer sequences. However, this methodology limits detection to already identified lineages, and thus is dependent on the quality of the data used for selection and preparation of the probe sequences.

4.4.4: General genomic characteristics of crocodilian ERVs

The analysis of the *Gammaretrovirus*-like ERV1 lineage suggests that these insertions may represent a more recent infection event, and given the overall completeness of the sequences identified, may be capable of replication by reinfection. This is in concordance with previous observations that some *pro-pol* fragments may encode intact ORFs (Chapter 2). All the major retroviral domains included in the RetroTector screening algorithms (Sperber et al., 2007) appear to be present, although this cannot be ascertained for all of the insertions since many of these insertions were not fully assembled. Notably, all of the detected motifs from within the coding domains were found to be present in the same reading frame within the consensus sequence. Furthermore, a lack of stop codons within the coding domains suggests that any mutations that have occurred in the individual insertions are not shared among many of the other detected insertions. Thus is it highly likely that that ERV lineage has retained the ability to replicate autonomously, and may still be active within the *C. porosus* genome.

Interestingly, the ERV lineage that showed similarity to the exogenous *Epsilonretroviruses* did not appear to contain additional ORFs, although sequences detected were incomplete. Three additional ORFs have been identified in exogenous *Epsilonretroviruses*, and may be considered as a component of the structural characteristic of this genus (Holzschu et al., 1995, Vogt, 1997). These ORFs are located upstream of the *gag* domain, and downstream of

the *env* domain. As these regions are absent from the current consensus sequence, it is possible that these ORFs may be present in crocodilian ERVs as well.

The degree of degradation observed in the other ERV lineages identified from these clones suggests that these may be remnants of ancient infection events that have been inactivated for some time. Based on the sequence data for these insertions, they are unlikely to be capable of replication by reinfection as multiple stop codons were present throughout each of the predicted coding domains. However, recovery of more copies of these lineages from the crocodilian genome sequences should provide further insights into the replicative potential of these lineages.

4.4.5: One ERV4 lineage has captured a host mRNA

The presence of an ERV4 lineage within these BAC clones was unexpected as the probe used to select clones for sequencing was specific to a single ERV1 lineage. However, given that all of the detected ERVs from this lineage were identical, it is more likely to represent a single insertion that was well represented within the paired-end reads rather than multiple insertions across all of the BAC clones. Nevertheless, the identification of additional coding regions within this insertion offers an interesting insight into the evolution of ERVs within the genome of *C. porosus*, and their potential effects on the genomes of these taxa.

The acquisition of an additional ORF through capture of host mRNA is a relatively uncommon occurrence. ERVs are capable of incorporating host genes through recombination and incorporation of the host mRNA into the retroviral genome. This process requires transcription of the cellular gene along with proviral DNA, co-packaging of the chimeric RNA particle, followed by infection of a new cell and recombination of the chimeric RNA with the retroviral RNA genome prior to insertion of the recombinant proviral genome (Figure 4.10) (Muriaux and Rein, 2003). This usually results in the deletion of part of the internal viral coding domains, rendering the resulting provirus incapable of autonomous replication (Katz and Skalka, 1990).

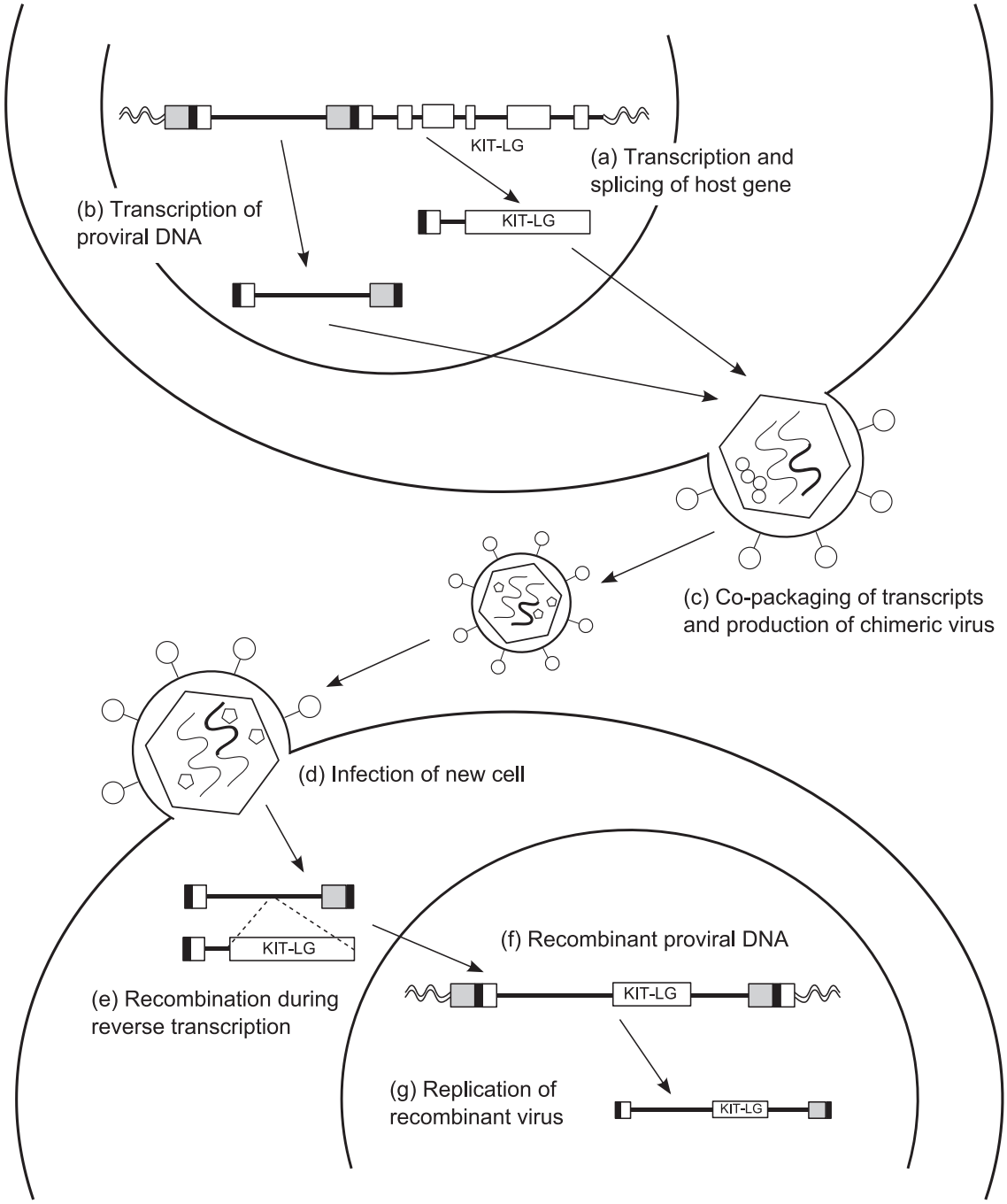


Figure 4.10: Capture of a host mRNA transcript by read-through transcription of the proviral DNA followed by recombination, leading to the incorporation of the host transcript into the retroviral genome. Adapted from Muriaux and Rein (2003).

The ERV4 lineage identified in this study is highly unusual in the respect, as it appears that incorporation of the KIT-ligand mRNA has taken place without significant loss of viral coding regions. To date, only one other example of this occurring has been described, the replication competent Rous sarcoma virus (RSV) (Schwartz et al., 1983, Swanstrom et al., 1983). Identification and comparison of this ERV with other closely related ERV sequences from the crocodilian genomes will be required to confirm that this mRNA has been captured without the loss of any viral domains. However, the transcript itself appears to be complete, and further investigation into potential transcription of this sequence may be warranted.

4.4.6: Limitations of low coverage sequencing

The average coverage for each of the BAC clones was much lower than expected, although the depth of overage of the assembled contigs and scaffolds were sufficient for basic analyses, such as the identification of retroviral domains and prediction of some codign regions. Examination of the number of reads incorporated into each assembly suggested that the low coverage was likely a result of the number of reads that contained no recognisable CAG-SXT tag. Consequently, the resulting assemblies were quite fragmented, with a large number of contigs of varying lengths (Appendix I, Table S4.4).

Although the paired-end libraries increased theoretical coverage of the clone substantially, this did not improve the contiguity of the resulting assembly. This lack of contiguity may be a result of insufficient sequencing coverage. Guidelines for whole genome sequencing recommend between 25× and 70× coverage for *de novo* assembly (Schuster, 2007). While there has been no similar figure published for BAC sequencing, the required coverage is likely to be similar. Despite this, the large number of contigs and singleton reads suggests that a large proportion of each BAC may have been sequenced.

Comparisons of the coordinates of the detected ERV insertions and the length of the contigs that they were located on suggests that the assembly of the ERV proviruses broke over what are likely to be the more repetitive regions of the ERV genome, such as the R region of the LTR. Given that the LTRs of many of the detected insertions were truncated or not present, this may indicate that the assembly program had difficulty in assembling the sequence reads over this region of the ERV genome. While this fragmentation prevented further analysis of

the genomic regions immediately surrounding the ERV insertions, the depth of sequencing over the internal regions of the ERV sequences recovered in this survey suggested that it did not affect the analysis of the assembled internal ERV domains.

4.5: Conclusions

The data presented in this chapter provide an important insight into the overall genomic structure of crocodilian ERVs. It has allowed for a basic characterisation of the major ERV lineages present in *C. porosus*, and provides an overview of the diversity and diversification of crocodilian ERVs. In particular, the discovery of novel lineages and lineage diversification within these ERVs suggests that there is still a lot to learn about crocodilian ERV diversity and evolution.

At the time that this study was initiated, these crocodilian BAC libraries were the only additional genomic resource available for further characterisation of ERV sequences. While the whole genome sequence assemblies of three crocodilian taxa have subsequently been released, and have been utilised in the later analyses within this study, BAC libraries, where available, continue to be a valuable resource for the discovery of ERV insertions and specific investigations into the dynamics of specific ERV lineages or insertions.

Directly interrogating the genomes of key crocodilian species will provide a more comprehensive insight into the evolution of crocodilian ERVs than the methodology implemented here. However, the data and methodology devised in this chapter provide the framework through which this characterisation and investigation can take place.

Chapter 5: A comparison of ERV detection programs

5.1: Introduction

Identification and classification of ERVs, TEs and other repetitive elements is becoming increasingly important with the advent of whole genome sequencing projects. The masking of repeat sequences in genome sequencing and annotation projects is particularly important before performing homology-based searches and predictions, such as used for gene annotation (Price et al., 2005). Repeat families that are capable of autonomous replication can pose a particular problem, as these encode their own set of genes for replication. These can cause problems for large scale gene annotation, particularly if these genes share some similarity with genes within the host genome (Bao and Eddy, 2002, Price et al., 2005).

A wide variety of tools are available for the detection and recovery of ERV sequences. The previous chapters have explored primarily laboratory-based methods, such as PCR and DNA hybridisation. These methods pose a number of challenges and restrictions – notably, the need for prior knowledge of some sequence information to facilitate detection of these elements, and facilities and reagents to carry out the experiments. The advent of next generation sequencing technologies and availability of vast amounts of sequence data mean that *in silico* or bioinformatics-based detection methods may provide a faster, more viable alternative for the recovery of large numbers of these elements. This chapter provides a comparison of a number of commonly used ERV detection programs to establish their relative effectiveness for the detection of divergent ERVs from a subset of the *C. porosus* genome.

5.1.1: *In silico* ERV detection and classification

In silico, or computer-based ERV detection methods take two general forms: Homology-based detection, and *de novo* detection. Both of these require some prior knowledge of ERV structure and sequence, although *de novo* methods are designed to work from minimal prior knowledge, and generally focus more on structural characteristics and common features rather than sequence homology. ERVs and related LTR retrotransposons, such as those of the G*ypsy/Ty-copia* and *bel* classes present a particular challenge for

detection methods due to varying levels of conservation throughout their sequences (Benit et al., 2001, Katz and Skalka, 1990, Katz and Skalka, 1994, McClure et al., 1988).

Homology-based detection of TEs relies primarily on identification of sequences that share similarity to previously identified TEs, using approaches similar to the BLAST searches utilised in previous chapters. Programs such as CENSOR (Jurka et al., 1996) and RepeatMasker (Smit et al., 1996-2010) use this method to identify TEs, producing a set of "masked" sequences where the identified TEs have been replaced with non-coding characters such as 'X' or an asterisk. This form of identification is relatively quick compared to the other methods assessed here, and therefore is a more efficient alternative for most comparative and functional genomic studies where the focus is less on discovery and characterisation of repetitive sequence, and more on the removal of such sequence to minimise the search space for other analyses, such as gene discovery.

However, TEs evolve with their host genomes, presenting a challenge for repeat detection based solely on sequence similarity. Species where there are no closely related characterised genomes, particularly those of non-model taxa such as crocodilians, pose a particular problem for these methods of repeat detection, as TE families in these genomes may be quite divergent from currently classified families and lineages. In particular, current ERV classification is based around similarity to the seven exogenous retroviral genera or a more relaxed clustering into the three major ERV classes (see Section 1.2 in Chapter 1). However, previous studies (Chapters 2, 3 and 4) (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Martin et al., 2002) have demonstrated that ERV lineages do not necessary cluster strongly within these groups, with sequences forming intermediate lineages between exogenous retroviral genera. In particular the complicated nature of ERV classification is demonstrated by the divergent lineages identified in crocodilians by Jaratlerdsiri et al. (2009), Martin et al. (2002), and in Chapters 2, 3 and 4.

The wide variety of genomes currently being sequenced and the identification divergent ERV lineages has highlighted the need for the development of *de novo* detection programs and algorithms, to facilitate detection and preliminary classification of such elements. These *de novo* detection programs have been developed with the intention of being able to detect TE insertions from uncharacterised genomes or those where there is no closely related, annotated reference (summarised in Table 5.1). This can be achieved through two major methods. The

first of these involves creating multiple local alignments of sequences to detect and define the outer boundaries of repetitive sequences, before comparing detected repeats and assigning these to "families", or groups of related sequences. Programs using this strategy include Recon (Bao and Eddy, 2002), RepeatScout (Price et al., 2005), and RepeatModeler (Smit and Hubley, 2008-2010).

The second commonly used method for *de novo* detection utilises conserved structural features of transposons. This is generally used for programs aiming to recover particular groups of transposons, as conserved features usually correspond to important functional regions of the transposon genome. A number of programs, such as LTR_STRUC (McCarthy and McDonald, 2003), LTR_FINDER (Xu and Wang, 2007), and RetroTector (Sperber et al., 2007) have been developed to utilise these features in LTR retrotransposons. Commonly used features of these transposons include the LTR regions, PBS, PPT, and reverse transcriptase domain (*pol* region).

Table 5.1: Summary of commonly used TE detection programs and their detection methods.

Program	Detection type	Detection method	Individual insertions	Classification	Reference
Censor	Homology	Database	Yes	Class	Jurka et al. (1996)
LTR_FINDER	*de novo*	Alignment Structural	Yes	N/A	Xu and Wang (2007)
LTR_STRUC	*de novo*	Alignment Structural	No	Lineage	McCarthy and McDonald (2003)
RECON	*de novo*	Alignment	No	Lineage	Bao and Eddy (2002)
RepeatMasker	Homology	Database	Yes	Class	Smit et al. (1996-2010)
RepeatModeler	*de novo*	Alignment Database	No	Class Lineage	Smit and Hubley (2008-2010)
RepeatScout	*de novo*	Alignment	No	Lineage	Price et al. (2005)
RetroTector	*de novo*	Structural	Yes	Class	Sperber et al. (2007)

Censor, RECON, RepeatMasker, RepeatModeler, and RepeatScout have been designed to recover and classify all repetitive sequences within the dataset, including ERVs, other retrotransposons, DNA transposons, and simple repeats. On the other hand, LTR_FINDER, LTR_STRUC, and RetroTector, have been written to specifically identify LTR retrotransposons. RetroTector, in particular, is specifically designed for the recovery of ERVs and the closely related *Gypsy* transposons.

Given the diversity and divergence that has been detected in crocodilian ERVs to date, simple homology-based searches are unlikely to detect the full complement of ERVs present in the crocodilian genomes. This study compares a number of stand-alone ERV detection programs to evaluate their relative ability to recover crocodilian ERV sequences and return meaningful data that could be used for phylogenetic and evolutionary studies. In particular the focus was on the ability to detect divergent ERV lineages. The data presented in this chapter is primarily aimed at supporting these comparisons, while a detailed analysis of the detected ERVs has been presented in Chapter 4.

5.2: Methodology

5.2.1: Selection of programs for comparison

Initial insights into the ERV complement of crocodilians from previous chapters have demonstrated that crocodilians are hosts to a broad variety of ERV lineages, some of which are quite divergent. Therefore, a selection of commonly used detection programs were evaluated to determine which would provide the most useful information for identification of ERV lineages in the crocodilian genomes and allow for predictions regarding their evolutionary history (Table 5.1).

A comparison of methodologies and program limitations suggested that LTR_STRUC and LTR_FINDER would not be suitable for ERV detection. Both of these programs have been designed to identify complete LTR retrotransposons based on the structural characteristics outlined previously. However, the data from the previous chapters suggested that crocodilian ERVs may be quite divergent, even at these more conserved regions. Additionally, these programs rely on a high degree of similarity within the LTR regions for detection and definition of element boundaries; a feature that may reduce their effectiveness when detecting

older, more degraded ERVs. The reliance on LTRs for definition of proviral insertions also prevents these programs from detecting solo LTRS, which are expected outnumber complete ERVs many times over, limiting their usefulness for estimating the total fraction of the genome that may be made up of ERV insertions.

From the remaining programs, RepeatMasker, RepeatModeler, and RetroTector were compared to determine which program would be most suitable to provide data for downstream ERV analysis. RepeatMasker utilises the BLAST search algorithms to identify repetitive elements within the dataset, providing a useful baseline to assess the specificity and sensitivity of the other programs when detecting elements related to known lineages. However, its usefulness for identification of more divergent elements maybe limited by the absence of known sequences within the database used for classification.

RepeatModeler incorporates RECON and RepeatScout within its detection strategy, making it unnecessary to run these programs separately. It was chosen as it is one of the commonly used tools to screen detect repetitive elements, such as ERVs, and mask these sequences for downstream analysis of non-repetitive genomic sequence. Using a combination of different detection methods, RepeatModeler aligns short lengths of each sequence (*l*-mers) from the input dataset to identify lengths of sequence that are present in multiple copies. These are then expanded until the sequences are no longer similar. This point is taken to be the end of the repetitive sequence. These ends are then adjusted to recover the likely ends of each element based on the multiple alignment. A consensus sequence can then be produced for that lineage of element, and compared to known sequences, such as through an iteration of RepeatMasker, for initial classification of that lineage. Despite this, it is not limited by the collection of known sequences within the database, and provides a collection of 'unknown' lineages that can be manually classified (Smit and Hubley, 2008-2010).

The primary aim of RetroTector is the retrieval and classification of ERV sequences, unlike RepeatMasker and RepeatModeler which can detect all forms of repetitive sequence. This allows the use structural and conserved motifs for identification, rather than sequence alignments. The basic premise of this method, as described in Chapter 4, is the identification of conserved motifs and structural characteristics of ERVs, such as the PBS, PPT, and TSDs. These motifs are then joined into 'chains' based on the order of the motifs, predicted reading frames, and likely distances between motifs derived from previously identified ERVs and

exogenous retroviruses. ERVs are classified into one of the seven exogenous retroviral genera based on the detected motifs and the retroviral or ERV sequences these have been derived from. This also allows for the prediction of 'puteins', or putative proteins, from the retroviral domains (Sperber et al., 2007). While RetroTector does not group ERVs into lineages automatically, these puteins are a valuable source of information for the prediction of lineages during downstream analyses.

5.2.2: Dataset and implementation

As there is limited sequence data available for these ERVs outside of the *pro-pol* region, it is also unclear how effective the various *de novo* detection programs will be at detection and classification of these divergent lineages. The sequence data generated in Chapter 4 provides\ a suitable dataset for the comparison of these methods and the opportunity to evaluate the effectiveness of these programs for the recovery of ERV data for the purpose of identification and characterisation of divergent elements in a non-model species. The data used for comparison of these programs were the assembled BAC clones sequenced in Chapter 4.

Briefly, these data were created by low coverage 454 sequencing of BAC clones from the *C. porosus* genomic library shown in Chapter 4 to contain ERV related fragments, and high coverage sequencing of MHC related BAC clones as part of a separate study that is not associated with the current project, but were included as they are from a region known to be rich in ERVs. The final assemblies comprised 6118 contigs and scaffolds from both the high and low coverage BAC clones totalling 149,498,987 bases (86,602,639 excluding 'N' bases).This dataset was chosen for the comparisons as it is known to contain ERV sequences (see BLAST results in Chapter 4, Section 4.3.3) and is much smaller than a complete genome making it less computationally intensive. This work was carried out following assembly of the BAC clones, but prior to the analysis of ERV structure described in that chapter.

Both RepeatMasker and RepeatModeler were implemented using the RepeatMasker version 20120418 of the RepBase database. RepeatMasker version open-4.0.0 was run using "vertebrata" as the query species and default sensitivity. RetroTector was run using default parameters and RetroTectorEngine, a command-line implementation of the RetroTector

program. All programs were run on a Dell Optiplex-745 desktop computer with 2 Intel Pentium D processors (3.00 GHz) and 2 GB RAM, running Ubuntu Linux 12.04.

5.2.3: Comparisons between programs

Due to the different methods of detection and the format of the final output, direct comparisons between RepeatMasker, RepeatModeler, and RetroTector cannot be made (Sperber et al., 2007). However, it is still possible to assess relative performances of these programs where output formats overlap, and compare these to determine which program, or programs, provides the most meaningful data for characterisation of ERV insertions and studies of ERV dynamics within a genome. For example, RepeatMasker and RetroTector provide the locations and details of specific insertions, while RepeatModeler output consists of consensus sequences from detected lineages. On the other hand, RepeatModeler and RetroTector provide the predicted sequence of the entire detected insertion, while RepeatMasker output is limited to the regions of homology to database sequences (see also Table 5.1).

Comparisons between RepeatMasker and RetroTector were made on the basis of the length and number of insertions that were detected. Of particular interest here were insertions that were detected by RetroTector that were not present in the RepeatMasker collection. These were identified using custom python scripts that compared the contig or scaffold locations of each of the RetroTector insertions with those from the RepeatMasker output. Insertions that were not present in the latter set of sequences, even in fragmented form, were retrieved for manual investigation. This involved comparisons against published sequences from the GenBank and RepBase databases as outlined in in previous chapters.

Sequence clustering and phylogenetic comparisons were used to determine the likely groupings of the RetroTector sequences with the RepeatModeler families as the latter program does not record the locations of the individual insertions used to create the final family consensus sequences reported. The program cd-hit-est in the CD-HIT package (Li and Godzik, 2006) was used to identify clusters of sequences based on nucleotide sequence similarities. Highly similar sequences (99% or greater similarity) were then removed and the remaining sequences aligned in MAFFT (Katoh et al., 2005) using the E-INS-i algorithm as

used previously. After removing non-overlapping sequences, Neighbour Joining trees were created using MEGA5 (Tamura et al., 2011) based on p-distances with 1000 bootstrap replicates to determine statistical support. Novel sequences generated by RepeatModeler were again compared against published sequences from GenBank and RepBase to confirm their classification.

5.3: Results

5.3.1: Comparison of ERV detection statistics

The three programs selected for comparison produced varying estimates of the ERV content present in the assembled BAC clones differing in the proportion of the data that was predicted to be ERV related and the number of potential elements detected. RepeatMasker detected a total of 1069 potential elements from the BAC contigs, out of a total of 6118 contigs. RetroTector detected fewer insertions (280 sequences) but attributed a greater proportion of the input sequences to retroviral related fragments. RepeatModeler identified a total of eight ERV families within the sequenced BAC clones. However, due to the form of output (lineage consensus sequences rather than individual insertions), it was not possible to estimate the proportion of the dataset attributed to ERV sequences.

RepeatMasker attributed 0.28% of the dataset to ERV sequences. Potential retroviral related elements were recovered from the low coverage BAC assemblies were equivalent to approximately 0.25% of these scaffolds. The MHC region appeared to contain a higher proportion of ERV related sequences with approximately 2.33% of the total input attributed to retroviral related fragments. This fraction of the MHC assembly was attributed to 34 potential elements.

RetroTector attributed a much greater proportion of each dataset to ERV related sequence with 0.76% of the assembled BAC contigs and scaffolds incorporated into predicted ERV insertions. For the low coverage BAC assemblies, this made up 0.69% of the assembled scaffolds. Again the MHC region had a higher proportion of retroviral related sequences with 8.2% of the assembled contigs attributed to ERV insertions. However, comparisons of the locations of these insertions with RepeatMasker output suggested that this program may be identifying multiple parts of each insertion and listing these as separate elements.

5.3.2: Classification of ERVs by the tested programs

Most of the insertions detected by the three programs were classed as ERV1 or ERVK (Human endogenous retrovirus-K; HERV-K) sequences. RepeatMasker also identified a number of ERVL (ERV3 Foamy virus-like) fragments that were not present in the RetroTector output. However, as these were short fragments (13–201 bp), it is unlikely that these contained retroviral motifs that would have allowed RetroTector to infer partial ERV chains. These fragments were also not shared between contigs, making it unlikely that RepeatModeler would have been able to detect them.

RetroTector also recovered a number of ERV insertions that shared no similarity with any sequences in the RepeatMasker database. These were subsequently identified as members of the ERV4 lineage previously described from PCR amplified fragments (Chapters 2 and 3), and have been more fully described in Chapter 4. Initial examination of the ERV families reported by RepeatModeler revealed that the ERV4 lineage identified by RetroTector was not found in the RepeatModeler ERV families. This lineage was instead classified as a *Gypsy* retrotransposon. Final alignment sizes ranged from 16 to 30 sequences for each of these families.

Sequence similarities and clustering analyses between the RepeatModeler families and RetroTector insertions suggested that only three of these families clustered with RetroTector insertions at an 80% nucleotide sequence similarity cut-off. Phylogenetic clustering confirmed this with the same three RepeatModeler families clustering within clades of the RetroTector insertions (Figure 5.1). A fourth shared clade was also observed, with *Epsilonretrovirus*-like ERV sequences from both programs clustering together. However, the remaining three RepeatModeler families did not appear to share any similarity with known ERV sequences either from this study or from published material, and did not appear to be detected by RetroTector.

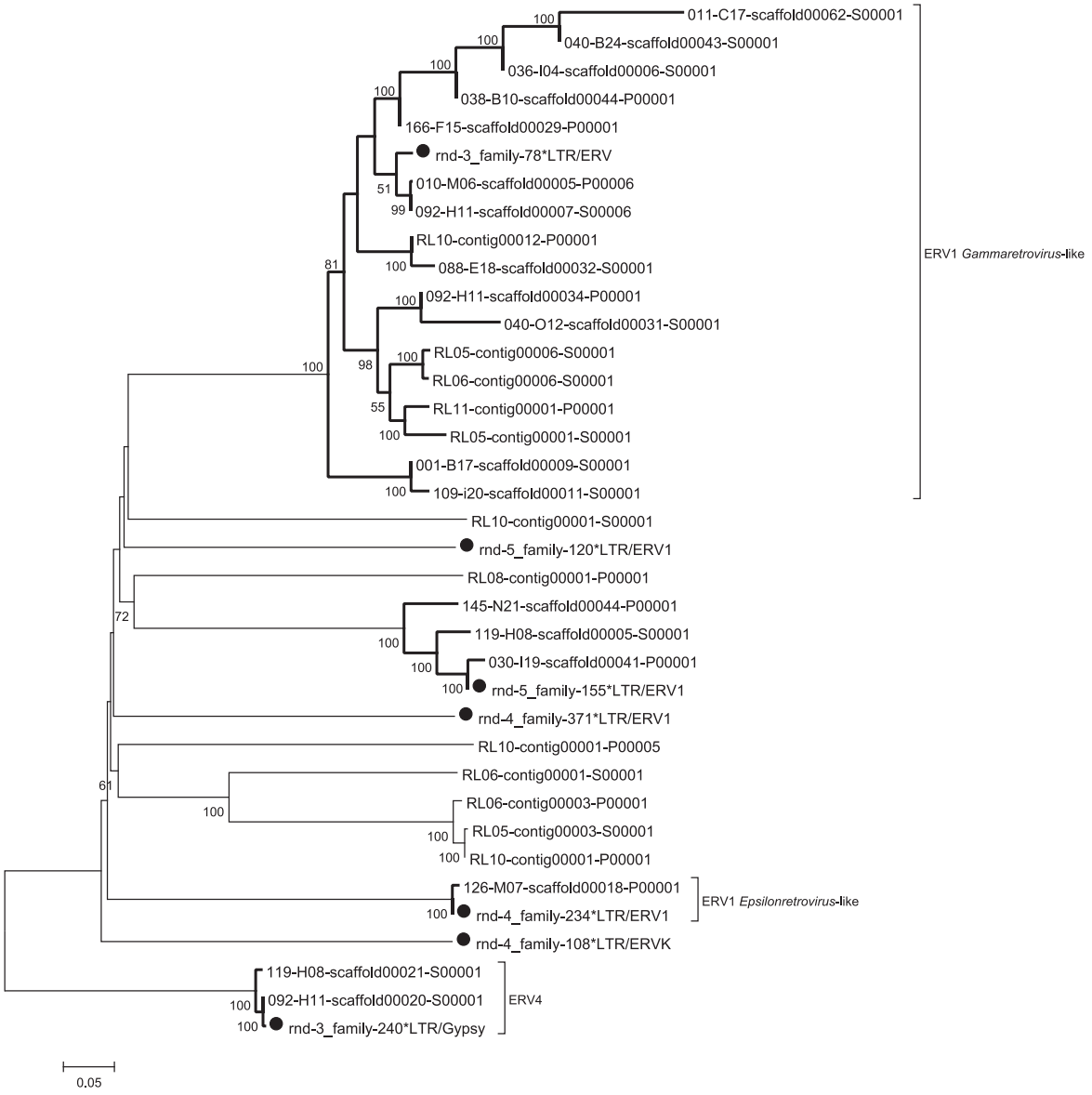


Figure 5.1: Many of the *C. porosus* families identified by RepeatModeler correspond to major ERV lineages within the insertions identified by RetroTector. Phylogenetic tree was generated by Neighbour Joining analysis. The shared clades are indicated by thicker branches. Black circles indicate the RepeatModeler families. Brackets beside the sequence names indicate previously identified ERV clades. The scale bar indicates branch length and the number indicate bootstrap support values greater than 50%.

5.4: Discussion

One of the major constraints of current ERV detection methods is the detection and classification of novel lineages and classes. In particular, the comparisons made in this chapter highlight the constraints of methods oriented around global similarity to sequence databases for ERV discovery. These comparisons show that similarity-based detection methods may have difficulty with this, especially with more divergent classes.

5.4.1: Relative effectiveness of the different detection methods

Overall, the *de novo* detection programs appeared to be more effective than homology based methods at predicting the ERV sequences and lineages present in the dataset. Comparisons of the insertions masked by RepeatMasker, the ERV chains detected by RetroTector, and the families identified by RepeatModeler suggest that RepeatMasker is relatively conservative in its predictions of ERV sequences. This is particularly evident from the much smaller proportion of ERV related sequence reported by RepeatMasker and the relatively short length of the repetitive sequences masked by the program, despite it returning a larger number of detected insertions. Likely, this is a result of different purposes of each program, and consequently, differing detection strategies.

The primary form of detection for sequence homology detection methods such as RepeatMasker relies solely on sequence similarity to previously classified sequences present in the sequence databases (Jurka et al., 1996, Smit et al., 1996-2010). Thus, these programs are reliant on the completeness of these databases to identify potential elements. This form of detection is relatively simple, making these programs faster and less computationally intensive than *de novo* detection. Such programs may be best suited to initial screens to provide an overview of the range of elements that may be present and quickly remove these elements for other genomic analyses such as gene prediction.

The *de novo* programs, on the other hand, are designed to infer the presence of ERVs and other TEs based on minimal structural and sequence characteristics (Smit and Hubley, 2008-2010, Sperber et al., 2007). As a result, this form of program may be more effective for initial characterisation of ERVs, particularly in new genomes. The data generated in this set of comparisons suggest that this form of prediction is more sensitive for detection of TEs as it is

capable of predicting insertions or families at varying degrees of sequence degradation. RetroTector in particular appears to be particularly suited for detection of ancient insetions, with a reported age limit of 200–300 million years for neutral insertions (Sperber et al., 2007).

5.4.2: Ability to detect divergent or novel lineages

The ability of the three tested programs to detect more divergent ERV lineages also varied depending on methodology. Unsurprisingly all three programs appeared to detect the ERV1 sequences in the dataset, including the previously described *Gammaretrovirus*-like and *Epsilonretrovirus*-like lineages, and a novel lineage. These ERV sequences are the most widespread across vertebrate taxa and share a relatively high degree of sequence similarity between taxa (Martin et al., 1997) making these lineages relatively easy to detect and classify regardless of the methods used.

On the other hand, the *de novo* methods were much more effective than RepeatMasker at detecting novel insertions, with both RepeatModeler and RetroTector identifying ERV4 insertions, although initial classification of these insertions differed between the programs. The limited capacity for programs such as RepeatMasker to identify these insertions is unsurprising, as previous attempts to classify these sequences (Chapters 2 and 3) (Jaratlerdsiri et al., 2009) and subsequent analyses (Chapter 4) have shown that this lineage of ERVs shares very little similarity to other recognised ERV classes. Furthermore, this class of ERVs had no representative sequence in the RepBase databases at the time of analysis. This has subsequently been updated, and using subsequent releases of the database may allow for the detection of these elements. Despite this, the results generated by these comparisons highlight the limitations of sequence homology-based detection for the identification of ERVs from non-model species, and the need to consider the desired form of output when selecting programs for ERV detection.

The success of the *de novo* programs at identifying the ERV4 insertions is almost definitely due to their ability to infer sequence characteristics from minimal prior information. In this aspect, RepeatModeler may be less constrained than RetroTector as it utilises local sequence alignments and copy number to identify and define repetitive elements such as ERVs, and

therefore is not reliant on a pre-defined database for identification. However, the reliance on sequence alignments also restricts the ability of RepeatModeler to detect very low and single copy number elements, as there are limited regions of homologous DNA for the construction of the final families and consensus sequences. This detection strategy may also impede the ability of RepeatModeler to detect very ancient insertions due to degradation of these sequences by mutations.

Although it is reliant on a database of conserved sequence motifs and characteristics, the use of these features allows RetroTector to detect and predict the presence of ERV sequence with a degree of confidence. By predicting the insertions individually, the program is able to identify single copy and low copy number lineages. Furthermore, the focus on conserved and functional motifs facilitates the detection of divergent and novel lineages, making this program a particularly useful tool for studies of ERV dynamics in poorly characterised genomes.

5.4.3: Avenues for improvement of ERV detection

The data generated so far from crocodilian ERV sequences highlight the need for new methods or adaptation of current methods to allow the identification and classification of novel and divergent sequences. Although the data generated in this chapter came from a single iteration of each of the programs tested, it has become apparent that multiple iterations of one or more of these detection methods, combined with additional data generated from these iterations, may be desirable for effective discovery of novel ERVs or ERVs from previously uncharacterised species.

Addition of novel data to existing databases has the potential to increase the accuracy and subsequent detection of elements by providing a larger range of 'characterised' sequence that can be used to initiate the search. For example, the addition of sequence data from the ERV4 lineage to the RepeatModeler database would assist in the correct classification of related sequences in subsequent iterations. Likewise, the addition of the novel motifs and classification information to the RetroTector databases would rectify the misclassification of these insertions, and possibly lead to the identification of other related ERV chains.

Unfortunately, the addition of new material and repeated iterations also increases the amount of time required for processing of data, particularly if manual curation of novel insertions is required at each stage. The automation of this process, through the creation of additional, local databases of divergent lineages during processing, has the potential to speed up discovery, potentially reducing the need for manual processing of data until the final stages of the detection process. RepeatModeler implements this system to some extent, since initial definition of the repetitive element families is carried out by clustering of sequences by local alignment (Smit and Hubley, 2008-2010).

Automation of ERV does have some drawbacks, however, in the definition of what may constitute a 'novel' element or insertion. Many attempts have been made to define the criteria for a novel element or family, such as sequence similarities compared to other elements, divergence from 'model' elements, similarity to exogenous retroviruses, and the presence or absence of phylogenetic support for element delineation (Blomberg et al., 2009, Jern et al., 2005, Wicker et al., 2007). However, the difficulties in implementing a consistent criteria across vertebrate taxa means that ERV classification is still a subjective form of analysis, based around arbitrary values and combinations of the listed criteria (Barrio et al., 2011, Garcia-Etxebarria and Jugo, 2010). Thus, complete automation of classification and retrieval may not be possible under current standards for classification and nomenclature.

5.5: Conclusions

The comparisons carried out in this chapter demonstrate the relative effectiveness of three commonly used ERV and TE detection programs to detect ERVs in a relatively unclassified genome, and highlight the need for careful consideration of the detection programs used to facilitate characterisation of ERVs in these species. It is evident from the various outputs that each program and detection method fulfils a different niche in the detection of ERV sequences. Homology and database searches such as implemented in RepeatMasker are more effective for screening of data from characterised taxa, while *de novo* methods are better suited to ERV detection and characterisation of uncharacterised genomes.

Both of the *de novo* programs were able to detect ERV4 sequences in the assembled contigs (see also Chapter 4), which RepeatMasker was unable to identify. While RepeatModeler was

equally as effective as RetroTector at detecting the more divergent ERV sequences, the program provides only the final consensus sequences created by alignment of the detected insertions, necessitating further processing of the raw sequence data to retrieve individual insertions. On the other hand, RetroTector provides details of the individual elements as well as information on their locations within each scaffold, and general structural data. Thus, RetroTector would be a better suited program for the study of ERV dynamics within crocodilians as it provides more useful information for studies of ERV lineage evolution and replication dynamics.

Chapter 6: Automation of ERV detection and classification

6.1: Introduction

Up until now crocodilian ERV characterisation has focussed on fragments from the *pro-pol* domain (Chapters 2 and 3) (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Martin et al., 1999), or ERVs from a single species (Chapters 2, 3, and 4) (Martin et al., 2002). These methodologies are highly reliant on sequence conservation for recovery of ERV data, with the PCR surveys focussing on conserved domains, and therefore likely to have missed more degraded or rarer ERVs. Given that these sequences comprised the probes used for the genomic BAC library screening, it is probable that the estimates generated from this screening underestimates the complete ERV complement of *C. porosus*. As a consequence, estimates of ERV content and diversity from these surveys are likely to be lower than the actual figures. Furthermore, the primary focus of experimental data for the previous chapters has been to identify and detail the ERV complement of *C. porosus*, limiting the ability to draw meaningful comparisons between crocodilians.

Improvements in next generation sequencing technologies have made these techniques accessible and feasible for furthering studies into the biology and evolution of non-model organisms. Crocodilians are one such species, and the initiation of the crocodilian genomes project has the potential to shed light on the evolution and genome biology of reptiles, avians, and modern vertebrate taxa (St John et al., 2012). *De novo* detection of ERVs from the crocodilian genome sequences may therefore provide a more accurate representation of ERV complement of these species. Similarly, a comparative genome-wide approach may be more appropriate for the study of evolutionary dynamics across crocodilians given that there are now whole genome sequences available. This chapter outlines some of the challenges faced when studying ERVs in such an environment and proposes a simple wrapper for the detection and initial processing of ERV data using the ERV detection program RetroTector (Sperber et al., 2007).

6.1.1: The crocodilian genomes

The genomes of three crocodilian species (*A. mississippiensis*, *C. porosus*, and *G. gangeticus*) have been generated, using a combination of Illumina and 454 sequencing technologies (St John et al., 2012). These species represent the three major taxonomic lineages present within the Order Crocodylia, namely the alligators, crocodiles, and gharials, respectively.

The *A. mississippiensis* genome was sequenced primarily using Illumina technology, utilising a combination of high coverage short reads and low coverage BAC-end sequencing. BAC-end sequencing involves the sequencing of the genomic regions immediately adjacent to the integration site of the BAC vector. *C. porosus* and *G. gangeticus* were sequenced with a hybrid methodology incorporating both Illumina and 454 sequencing technologies. These utilised high coverage, short sequence reads from Illumina genomic preparations, as well as single and paired-end reads generated using 454 sequencing. As BAC libraries are also available for these species, low coverage BAC-end sequencing was also used to combine contigs (St John et al., 2012).

The *A. mississippiensis* and *G. gangeticus* genomes have been estimated to be approximately 2.5 Gb in size (Gigabases, 2.5×10^9 bases), whilst *C. porosus* has been estimated to be slightly larger in size (2.78 Gb). Preliminary estimates of the repetitive DNA content of these genomes suggest that upwards of 23.44% for all three species are made up of repetitive DNA (St John et al., 2012). Given the divergence and diversity of ERVs detected in crocodilians to date, bioinformatic analyses of these genomes will facilitate the recovery of novel ERVs and the identification and characterisation of these lineages.

6.1.2: Challenges associated with analysing whole genomes

One of the major challenges of genomic ERV discovery is the handling of the vast amounts of data that may be produced by any of the *de novo* detection programs currently available (see Chapter 5). This challenge is made more complicated by the need to analyse multiple genomes concurrently, as with the three crocodilian genomes.

The *de novo* detection program RetroTector attempts to minimise the need for additional data handling by generating an SQL (structured query language) database containing the pertinent

ERV information for each genome and ERV insertion detected by the program (Sperber et al., 2007). However, setting up and querying the database for information regarding the detected insertions requires knowledge of SQL programming, a skill not always associated with genomic and genetic research. In order to simplify this process, the work outlined in this chapter collates pertinent information regarding the location and general structure of the ERV insertions into a series of simple tables and sequence files that do not require knowledge of a programming language to query or examine.

This chapter provides an overview of the methodology involved with ERV detection from a genomic perspective and initial insights into the ERV content of the crocodilian genomes. It expands on some of the complications associated with automated ERV detection as discussed in the previous chapter and provides an initial overview of the ERV content of the three sequenced crocodilian genomes and complements the comparative analyses in the following chapter. The studies outlined in the previous chapters suggest that crocodilians harbour a number of divergent ERV lineages that show little similarity to ERVs from other vertebrates. Based on the relative efficiencies of the programs discussed in Chapter 5, the *de novo* detection methods such as RetroTector may be more efficient than the traditionally used homology-based methods for ERV detection in these species.

6.2: Methodology

6.2.1: Pipeline design

The methodology discussed in this chapter relates to the automation of ERV detection and an initial exploration of the ERV content of the crocodilian genomes. Where possible, pre-existing programs and packages used in the previous chapters were also incorporated to reduce the amount of optimisation required to adapt previous methodology for automated analyses.

The detection of ERV insertions was carried out using RetroTector, specifically the commandline variant of the program, RetroTectorEngine to facilitate automation of the initial analyses. Due to the large amounts of data that such analysis is likely to produce, custom scripts were written in the python programming language to assist in the retrieval of data, and perform some of the basic data analyses. These scripts included initial splitting of the genome

assembly scaffolds into individual files for analysis, retrieving the detected ERV sequences and ERV gene sequences, identifying and collating sequence haplotypes, and compiling these into single files (see Figure 6.1 and Table 6.1).

The final automated pipeline was written and automated in the python programming language using the python module Ruffus (v2.2) to create the pipeline and control the flow of data between each stage of analysis. The Ruffus module (available from https://code.google.com/p/ruffus/downloads/list) is a python add-on designed to implement computational pipelines and manage the dependencies of each stage or 'task'. It has built-in controls for the processing of multiple tasks in parallel, as well as functions to limit the number of tasks that can be run concurrently. It also allows the pipeline to be started or restarted from an arbitrary point, and will automatically re-run jobs if output files are deemed to be 'out of date', or the modification dates on the input files for a specific task is after the modification date of the output (for example if the input files have been updated to incorporate new data). The files for the pipeline are provided in Appendix V.

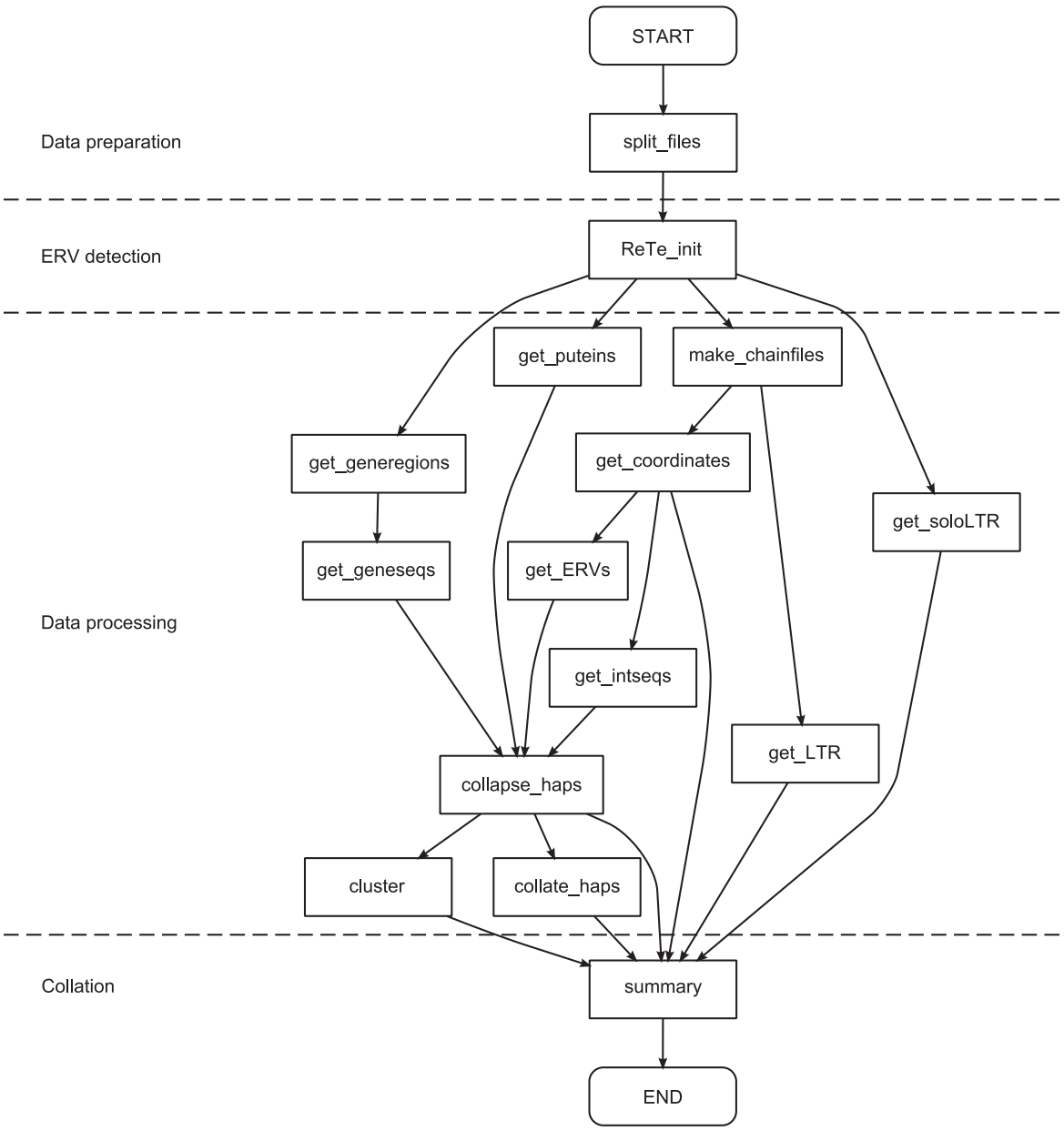


Figure 6.1: Flow diagram of the computational pipeline showing the tasks and their input and outputs. Briefly, genomic data are separated into individual scaffolds before being processed using RetroTector. The pipeline then takes the output from this program, extracts the final sequence data, and collates this into a number of easily accessible files. Arrows indicate the flow of data between tasks.

Table 6.1: Outline of each task specified by the pipeline, and a description of what each does.

Task	Description
split_files	Takes the file containing the genome sequences and writes each sequence to an individual fasta formatted file for RetroTector. These files are created in the directory "-splitfiles" within the main project directory.
ReTe_init	Copies the individual fasta files into the RetroTector "NewDNA" directory. Starts RetroTector – RetroTectorEngine.jar. Copies RetroTector output to the directory "-RetroTectoroutput".
make_chainfiles	Identifies the ERV chains within the RetroTector output and writes the details of each chain into a separate set of files within the directory "-chainfiles".
get_puteins	In addition to identifying ERV sequences, or chains, within the input data, RetroTector attemts to reconstruct amino acid sequences of the retroviral coding regions (puteins). This task identifies these within the Retrotector output and collates them into separate files for each coding region.
get_generegions	Identifies the start and end coordinates of the retroviral coding regions from the RetroTector output. Collates these into a single file for easy viewing.
get_geneseqs	Extracts the nucleotide sequences of the retroviral coding domains and collates these into separate files for each domain.
get_coordinates	Finds and collates the start and end locations of the detected ERV sequences and their LTRs, as well as the predicted PBS, PPT and TSD sequences.
get_LTR	Collates the 5` and 3` LTR sequences from each ERV insertion into a single fasta file.
get_soloLTR	Identifies the start and end coordinates of solo LTRs from the RetroTector output. Collates these into a single file for easy viewing.
get_intseqs	Extracts the internal sequences of each ERV insertion and writes this to a single fasta file.
get_ERVs	Extracts the complete ERV insertion from the input sequence and writes this to a fasta file.
collapse_haps	Identifies identical sequences from all of the fasta files generated in previous tasks. Output is a fasta file containing all of the haplotypes identified, a list of haplotypes and the sequences that belong to each one, as well as a file containing the number of sequences belonging to each haplotype.

Task	Description
collate_haps	Collates the information from each of the haplotype lists and organises this into a single table showing the haplotypes that were identified from each sequence, the complete ERV sequence, the entire internal domain, and for each of the individual coding domains.
cluster	Starts CD-HIT or CD-HIT-EST to perform cluster analyses on the sequence haplotypes identified above. Similarity is set at 80% to provide an overview of the possible number of lineages present within the sequences.
summary	Summarises the data generated by the pipeline into a single overview file. This file contains the sequence haplotypes for each ERV insertion identified, an indication of whether LTRs have been identified, and sequence information for the PBS, PPT and TSDs as generated by "get_coordinates".

6.2.2: Test datasets

Two datasets made up of sequences from Chapter 4 were used for the development of this pipeline. The first comprised of 15 sequences ranging in length from 12,553 to 284,168 bases (Table 6.2). Eight of these contained one of five previously detected crocodilian ERV haplotypes placed within randomly generated sequence, and were used only for initial development to ensure that the data were correctly parsed by the pipeline. The assembled BAC sequences that were obtained in Chapter 4 were used as a larger dataset to test the complete pipeline and gauge the computational resources that may be required at each stage of the pipeline. The results from this dataset are presented in Chapter 4. The pipeline was written and tested in Python 2.7 on a Dell Optiplex-745 desktop computer with 2 Intel Pentium D processors (3.00 GHz) and 2 GB RAM, running Ubuntu Linux 12.04. To prevent problems with insufficient memory, the pipeline was limited to one job at a time.

Table 6.2: Overview of the sequences that made up the preliminary test set for the development of the computational pipeline.

	Scaffold ID	Length	ERV Haplotype
1	scaffold_0001	123368	Haplotype 4
2	scaffold_0002	284186	Haplotype 4 and 5
3	scaffold_00001	54168	Haplotype 4
4	scaffold00002	30735	N/A
5	scaffold00003A	28800	Haplotype 3
6	scaffold00003B	28800	Haplotype 3
7	scaffold00004	28439	N/A
8	scaffold00005	26806	N/A
9	scaffold00006	15962	N/A
10	scaffold00007	13946	N/A
11	scaffold00008	13800	N/A
12	scaffold00009	13643	N/A
13	scaffold00010	12553	N/A
14	001-B17-scaffold00003	131869	Haplotype 2
15	001-B17-scaffold00001	140642	Haplotype 1

6.2.3: Analysis of the crocodilian genomes

The crocodilian genome assemblies were obtained from the ICGWG webpage (ftp://ftp.crocgenomes.org/pub/). The final implementation of the pipeline was on a Dell PowerEdge R815 server with 2 AMD Opteron 6220 8-core processors (3.00 GHz) and 128 GB RAM, running Redhat Enterprise Linux 6. On this machine, it was run with Python 2.6 with the argparse module installed, and no restrictions on the number of jobs that could be run concurrently. Each genome was analysed using the above pipeline, and summary statistics were obtained from the resulting output.

Duplicates may occur where scaffolds are too long to be processed by RetroTector and randomly split by the program. When this occurs, the ERV insertion may be present in more than one of these split sequences giving rise to multiple detections. These can be found by comparing the location of the insertion from the program output as the locations within the scaffold are maintained regardless of where the division is made. Occasionally these ERV sequences had differing internal regions as a result of the RetroTector predictions. When this occurred, predictions were manually checked and the information from each entry was merged. The estimated proportion of each genome that was likely to be ERV related was

calculated from the lengths of the detected ERV chains and the total length of the assembled scaffolds (Table 6.3).

6.3: Results

6.3.1: Overview and completeness of ERV chains

RetroTector recovered a total of 4531 ERV chains from the three crocodilian genomes. Of these, 576 were treated as 'complete' as they had all three coding domains and both 5` and 3` LTRs present. These chains were retrieved for further analysis as outlined in Chapter 7. Between 16% and 25% of detected ERV chains were predicted to contain two LTRs but were missing at least one of the internal coding domains (Table 6.3). The average length of ERV chains were 8,239 bases in *A. mississippiensis*, 7,826 in *C. porosus*, and 7,363 in *G. gangeticus* (see Appendix I, Table S6.1 for individual insertions). The detected ERV chains make up between 0.29% and 0.70% of the crocodilian genomes (species specific figures below in Table 6.3). Furthermore, a number of ERVs were detected that did not appear to encode a recognizable *pol* domain, and thus would not have been detected by hybridisation.

An additional 339,610 solo LTRs were detected from the three genomes. The average length of solo LTRs ranged from 1473 to 1573 bases across the three species. Based on these figures, it is estimated that between 6.53% and 9.30% of the genomes are made up of ERV related sequences.

Table 6.3: Summary of ERV insertions (including *Gypsy*-like transposons) detected from each of the genomes, and an approximation of the ERV content.

	A. mississippiensis	*C. porosus*	*G. gangeticus*
Genome size[a]	2.5	2.78	2.5
Version	aMiss_AKHW01000000	croc_sub2	ggan_v0.2
Date uploaded	28/11/2012	10/08/2012	05/09/2012
Date downloaded[b]	10/12/2012	15/10/2012	15/10/2012
Length of assembly	2,144,153,452	2,123,474,087	2,882,656,219
Number of scaffolds	8,897	23,365	47,351
Estimated proportion of genome in assembly	0.86×	0.76×	1.15×
Number of ERV chains	1,817	1,593	1,121
Complete ERVs	207	249	120
Both LTRs	1,316	1,013	613
One LTR	134	130	104
No LTRs	367	450	404
Solo LTRs	79,445	119,753	140,412
Estimated ERV content (excluding solo LTRs)	6.53% (0.70%)	9.30% (0.59%)	7.46% (0.29%)

[a] Estimates of genome size taken from St John et al. (2012).

[b] Genomes were downloaded before the 'freeze' of assembly and refinement. Therefore these may not represent the latest, or the published version of the genomes.

6.3.2: Detection of coding domains

The *pol* domains were consistently the most commonly detected domain across all three species, followed by the *gag* and *pro* domains (Table 6.4). The *env* domain was again the least likely of the retroviral domains to be detected in the ERV chains. Similarly, proportionally more predicted amino acid sequences could be generated based on predicted *pol* nucleotide sequences than the other domains. The *pro* domain appeared to be the most difficult domain to generate predicted amino acid sequences from, with proportionally less puteins generated from these sequences. Additional results and further analysis of these data will be discussed in Chapter 7.

Table 6.4: Summary of the coding domains detected and the number of haplotypes recovered from each.

		A. mississippiensis	*C. porosus*	*G. gangeticus*
Number of sequences	Gag	1,234	1,143	752
	Pro	1,087	1,065	723
	Pol	1,736	1,515	1,008
	Env	666	653	472
Nucleotide haplotypes	Gag	1,213	1,131	749
	Pro	1,062	1,050	713
	Pol	1,711	1,505	1,002
	Env	656	652	468
Number of Puteins [a]	Gag	597 (0.48)	683 (0.60)	457 (0.61)
	Pro	420 (0.39)	536 (0.50)	373 (0.52)
	Pol	958 (0.55)	987 (0.65)	743 (0.74)
	Env	348 (0.52)	407 (0.62)	288 (0.61)
Putein haplotypes	Gag	580	673	452
	Pro	407	526	370
	Pol	946	982	739
	Env	344	406	285

[a] Brackets indicate the number of predicted puteins as a proportion of the number of predicted nucleotide sequences.

6.4: Discussion

6.4.1: Preservation of ERVs within the genome

An initial overview of the ERV insertions recovered from the genomes of *A. mississippiensis*, *C. porosus*, and *G. gangeticus*, suggests that many of the ERVs recovered are highly degraded. Degradation of ERVs is generally the result of accumulation of mutations through mis-incorporation of nucleotides or production of indels (insertions and deletions) during DNA replication. Consequently, the accumulations of these mutations provide a relative measure of the age of the insertions. As the majority of ancient ERVs within a genome are expected to be selectively neutral, and thus of little consequence to genome function, these ERVs are likely to represent remnants of older insertions. Despite this, the presence of a large number of relatively intact ERVs suggests that there is some degree of sequence preservation, and as such, it will be interesting to compare the relative conservation of the different retroviral families within crocodilians.

Likewise, the high proportion of solo LTRs detected suggests that ERVs have previously proliferated in the genomes of these species before being removed by LTR-LTR recombination (Stoye, 2001). These remnants of ERV proviral sequences may be present in genomes in tens to thousands of copies per ERV family (Stoye, 2001). Although they are of little consequence for classification and description of ERV families in general, solo LTRs can still have significant impacts on genome function, altering gene transcription, proving alternative splice sites, or causing the formation of pseudogenes by disruption of regulatory domains (Jern and Coffin, 2008). While it is not currently possible to establish the impacts of solo LTRs on crocodilian genomes at present, the presence of these remnant ERVs is an important consideration for studies into gene regulation in these species.

6.4.2: Comparison of ERV estimates between methods

The estimated ERV content of the three crocodilian genomes is much higher than the previously predicted 0.07−0.1% for *C. porosus* (Chapter 4). Based on these revised figures, the estimated ERV content of crocodilians appears to be similar to that of most other characterised vertebrate species (Table 6.5), rather than the smaller ERV complement suggested from densitometry analyses. This is particularly apparent when compared with other species where ERV content has also been estimated with RetroTector. It is possible that some of the ERV sequences included within these estimates are false positives or represent mis-assembled regions of the genome. However, these are likely to constitute only a small proportion of the estimated ERV content, and therefore do not explain the significantly different estimates obtained here and in Chapter 4.

It is more likely that these differing estimates of ERV content are due to a large number of novel lineages within the data that were not present in the probe set used for hybridisation screening. This is unsurprising given the limited knowledge of the ERV complement of crocodilians from PCR-based analyses, and therefore the limited number of linages represented in the probe sets used for screening of the BAC libraries. As discussed in Chapters 2, 3, and 4, the PCR-based screening was unlikely to detect all ERV lineages present in the crocodilian species assessed as only a limited number of plasmid clones were selected for sequencing from each individual. The amplicons selected were also restricted to what was considered to be close to full length amplicons to reduce the possibility of

nonspecific amplification and highly degraded ERVs being included in these surveys. Furthermore, a number of ERVs were detected in the genomes that did not appear to encode a recognisable *pol* domain, and thus would not have been detected by hybridisation.

Table 6.5: Estimated ERV content based on retroviral chains, and a comparison with previous estimates and other species.

Species	Common name	% ERVs in genome	Reference
A. mississippiensis	American Alligator	0.70%	
C. porosus	Saltwater Crocodile	0.59%	
G. gangeticus	Gharial	0.29%	
Danio rerio[a]	Zebrafish	0.8%	Barrio et al. (2011)
Monodelphis domestica[a]	Opossum	2.0%	Barrio et al. (2011)
Canis familiaris[a]	Dog	0.15%	Barrio et al. (2011)
Mus musculus[a]	Mouse	2.0%	Barrio et al. (2011)
Homo sapiens[a]	Human	0.8%	Blikstad et al. (2008)
Gallus gallus	Chicken	2.9% (0.2% [a])	Huda et al. (2008)
Anolis carolinensis	Green Anole	3.0%	Alfoldi et al. (2011)
Xenopus tropicalis	Western Clawed Frog	0.12%	Hellsten et al. (2010)

[a] These values were also generated using RetroTector, and comparisons to these may be more representative as figures generated using other methods may reflect different detection biases from the various projects and methodologies (see Chapter 5 for an outline of this).

6.4.3: Differences in ERV complement between crocodilian species

Surprisingly, the estimated proportion of ERVs in the genomes of the three crocodilian species appeared to vary greatly between species, with the predicted content of the *A. mississippiensis* and *C. porosus* genomes at least double that of the *G. gangeticus* genome. While some variation in the proportion of genomic ERV content may be expected between species as a result of species specific acquisitions, the large difference in estimated content was largely unexpected. Differing ERV complements could be due to differing activity of ERVs, or acquisition of novel viral strains post speciation, or may be a reflection of the completeness of each genome assembly.

This latter possibility is most likely to explain the large difference in ERV complement between *G. gangeticus* and the two other genomes, as *A. mississippiensis* and *C. porosus*

were the first two genomes sequenced, and therefore are likely to be the most complete in terms of assembly at the time the genomes were obtained. In particular, *A. mississippiensis* was the most advanced, both in terms of contiguity as well as annotation (see also Table 6.3), and consequently, is more likely to be representative of the actual crocodilian genomes. The more fragmented nature of the *G. gangeticus* genome may reduce the ability of RetroTector to detect ERVs (Barrio et al., 2011, Blikstad et al., 2008) due to potential fragmentation of the ERV chains, leading to lower estimates of ERV number and content.

Interspecies variation in ERV content is also likely to be present, although it is likely to have a much lesser impact on the variation observed compared with genome contiguity and coverage. It has also been noted that genome complexity is reflected in retroelement content of the host genome, with more retroelements being present in more complex genomes (Rowe and Trono, 2011), although it is unknown if ERV content correlates with genome size. However within vertebrates, and between closely related families, this is less likely to be the sole cause of differences in the ERV complement between species. The primary sources of interspecies variation are independent acquisitions of novel ERV infections, and differing levels of ERV proliferation and loss as result of ERV evolution within host genomes (Johnson and Coffin, 1999, Kim et al., 1999). These trends are not immediately obvious from the raw data presented here, and therefore will be investigated further following the classification of these ERV sequences in the following chapter.

Genome biology and the genomic environment may also play a role in determining the final ERV complement of a genome. Acquisition of specific control mechanisms, exaptation of ERV domains, and the insertion location can all dictate the preservation or removal of ERVs from a genome. It has also been suggested that some species, notably avians (represented here by *G. gallus*; chicken) and *C. familiaris* (dog), may have additional mechanisms for purging ERVs from the genome or the restriction of retroviral activity (Barrio et al., 2011). As such, it is possible that similar mechanisms have evolved in *G. gangeticus*, and to a lesser extent, possibly *C. porosus* and *A. mississippiensis*. Unfortunately, as with *C. familiaris*, the paucity of retroviral data outside of this project, and the current limited understanding of crocodilian genome biology limits the extent to which further conclusions can be drawn on this.

6.5: Conclusions

The methodology and data presented in this chapter highlight the complexity of *de novo* ERV discovery. The pipeline presented here extends current ERV detection to include basic clustering analyses to facilitate the definition of ERV lineages and families. Furthermore, the pipeline acts as a wrapper around the detection program and associated data processing, performing the preparation of the data, initiation of the program and collation of the ERV data into easily accessible files for further analysis.

The ERV data suggest that the overall proportion of the crocodilian genomes that can be attributed to these elements does not differ greatly from other characterised species, despite containing a more divergent range of ERV lineages. The next stage of analysis will involve definition and description of the ERVs families that are present in these sequences, and will provide more comprehensive insights into the evolution of ERVs within the crocodilian genomes.

Chapter 7: Comparative studies of crocodilian ERVs

7.1: Introduction

7.1.1: ERV classification

Classification of ERVs and other TE families is an ongoing problem as there are a number of different taxonomic groupings that may be applied to ERV sequences. The first of these classification systems relates ERV sequences to the closest related exogenous retroviral genus (e.g. the *Gammaretroviruses*). However, surveys of ERVs across vertebrates suggest that the delineations between these genera are not distinct, with many forming intermediary lineages between genera. As a result, broader groupings of ERV lineages may also be used to denote classification. Under this system, ERVs are generally assigned to broad classes according to general sequence characteristics. These classes correspond approximately with the definition of a genus (e.g. *Gammaretroviruses*, ERV1 or Class I ERVs) and are further divided into 'families' which can be loosely defined as a group of related elements, likely to have originated from a single insertion or infection event.

There are three main methods that are commonly used to define ERV families: classification by similarity to recognised lineages or families; sequence similarity-based classification; and classification based on phylogenetic grouping and support. These methods may be used in combination or separately. Classification based on similarity to named sequence families can be useful in species where large amounts of research have been done, or where the family structures are well defined, such as the HERV families in humans and primates, or the endogenous avian retrovirus (EAV) and Rous-associated virus (RAV) strains in the domestic chicken. However, this becomes more difficult in newly characterised taxa where insertions may show only passing similarity to currently classified ERVs (Barrio et al., 2011, Garcia-Etxebarria and Jugo, 2010). In these cases *de novo* delineation methods may prove more effective.

One of the major drawbacks to sequence-based classification is the assignation of an arbitrary similarity cut-off value for the definition of ERV sequence families. Reviews that have dealt with this area (Blomberg et al., 2009, Wicker et al., 2007), suggest relatively conservative values, such as 80% amino acid sequence similarity across coding regions, with sub-lineages within the major groups designated arbitrarily, on a case by case basis. Classification by

phylogenetic clustering can also pose similar problems of support values and the definition of sub-lineages (Blomberg et al., 2009) as these analyses are dependent on the dataset and methodology used to create the alignment (Chapter 2), and then the quality of the final alignment.

The designation of newly defined families to the major ERV classes is an additional challenge, particularly as new species, and therefore new ERV lineages, are characterised. Traditionally ERV sequences have been classified by similarity to exogenous retroviruses, and thus assigned to one of the seven retroviral genera. However, the identification of apparent intermediates that share sequence characteristics with two or more genera suggests that this may not be prudent (Jern et al., 2005). Likewise, classification by genomic structure can be problematic, as demonstrated by the *Epsilonretrovirus*-like insertions that were recovered in Chapter 4. While these insertions clustered phylogenetically within the *Epsilonretrovirus* lineage along with the exogenous WDSV, analysis of the complete proviral genome suggested that these elements did not contain the additional ORFs that have been deemed characteristic of this genus (Holzschu et al., 1995, LaPierre et al., 1999, Vogt, 1997).

7.1.2: ERV evolution in the genome

The replication and subsequent divergence of ERV lineages within a host genome is driven by a number of factors and mechanisms, including mode of replication, selective pressures, and consequent effects on genomic function. As outlined in Chapter 2, there are three major routes for replication within a host genome: reinfection, retrotransposition, and complementation. Reinfection requires an intact provirus with functional coding domains and regulatory regions (Bannert and Kurth, 2006, Belshaw et al., 2004, Katzourakis et al., 2005). Retrotransposition requires that only the regulatory domains, *gag*, *pro* and *pol* regions are functional, as these are important to facilitate transcription and re-integration into the genome without the need for viral budding (Bannert and Kurth, 2006, Katzourakis et al., 2005). Complementation does not require that the retroviral domains are functional but is dependent on transcription and co-packaging by a complementary helper virus (Gifford and Tristem, 2003, Katzourakis et al., 2005).

ERV expansions within the genome tend to follow one of three patterns or models. The master gene model involves replication from a single, replication competent integration, and is reflected in phylogenetic reconstruction by 'star-like' trees where all retroelement sequences share a common node (Clough et al., 1996). On the other end of the spectrum is the 'transposon' model that assumes all elements are capable of replication and will be producing new insertions (Deininger et al., 1992). The alternative 'intermediate' or 'random template' model posits that a number of insertions are capable of replication at any given time and have evolved and replicated in distinguishable bursts (Clough et al., 1996, Cordaux et al., 2004). Phylogenetic reconstruction of these latter two models are characterised by a much more random tree structure with multiple clades representing replication events evident across the entire phylogeny (Cordaux et al., 2004).

The data generated in Chapters 2, 3 and 4 have suggested that crocodilian ERVs may be capable of replication within the genome, as intact ORFs have been found among those retroelements. However, these experiments were limited to hypotheses drawn from fragments of the *pro-pol* domains (Chapters 2 and 3) and consensus sequences from a small number of elements (Chapter 4). Subsequent recovery of complete proviral insertions from the crocodilian genome sequences will provide the data required for more informed inferences about the replicative potential of these ERVs and the mechanisms by which this occurs in these species.

This study aims to establish the distribution and processes driving ERV evolution in crocodilians using resources derived from the three sequenced genomes. For this, the ERV lineages must first be grouped into likely families, and assigned to the various ERV classes. The distribution and sequence characteristics of each family will be described based on relatively intact insertions. Following this, ERV families showing unusually high levels of proliferation within crocodilians will be examined to determine the mechanisms by which they have become established within the genome. Similarity of the surrounding genomic regions was also estimated to determine whether regional duplications were also playing a role in ERV replication.

7.2: Methodology

7.2.1: Definition of ERV families

ERV families were defined based on the predicted amino acid sequences of the *pol* domains obtained in Chapter 6 from searching the *A. mississippiensis*, *C. porosus*, and *G. gangeticus* genomes using RetroTector (Sperber et al., 2007). Due to the large number of insertions from all three genomes, only the sequences deemed to be 'complete' ERVs were used. These sequences were those where both LTRs and all three retroviral coding domains could be predicted by RetroTector. Sequences with more than five consecutive ambiguous amino acid residues were also excluded to ensure that fragmented insertions and potential assembly artefacts were not incorporated into the final dataset. While these criteria may bias analyses to insertions that are better preserved or more recently integrated, it also reduces the amount of sequence divergence and evolutionary 'noise' that may be introduced by the inclusion of highly degraded sequences.

BLASTX (Altschul et al., 1997) was used to classify the predicted *pol* puteins into the major ERV classes based on similarity to *pro-pol* and *pol* fragments recovered from previous chapters and published papers (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Martin et al., 1999, Martin et al., 2002). Sequences were then assigned to preliminary groupings based on Bit scores and hit coverage. Sequences that showed no similarity to known crocodilian ERV fragments where then compared to other published sequences in GenBank and RepBase using the NCBI BLAST suite (Johnson et al., 2008) and Censor (Kohany et al., 2006). Sequences within each of the major classes were then aligned in MAFFT (Katoh et al., 2005) using the E-INS-i algorithm as described previously.

To determine the likely lineages within these major classes, phylogenetic trees were then created in CLUSTALW (Thompson et al., 1994) using Neighbour Joining, uncorrected sequence distances (equivalent to the p-distance calculations used in previous chapters), and 1000 bootstrap replicates. Preliminary lineages were the defined based on clades with more than 70% bootstrap support. The sequences from each of these lineages were then re-aligned and refined based on sequence similarity and conservation within the *pol* domain.

To confirm the final classification of each of these lineages, representative sequences were then re-aligned with the *pol* domain of other published ERVs from other species also

recovered and classified using RetroTector (Jern et al., 2005, Sperber et al., 2007), and clustered based on Neighbour Joining phylogenies as described above. The final alignment comprised 66 sequences from the crocodilian ERV families and 60 published sequences (Appendix I, Table S2.1). Representative sequences were those most similar to the consensus sequence for each family. To create consensus sequences, the nucleotide sequences of ERV insertions for each family were first aligned using MUSCLE (Edgar, 2004) before consensus sequences were generated in BioEdit (Hall, 1999). Where two or more sequences were equally similar to the consensus, or there were only two sequences present in the lineage, the sequence with the least number of stop codons and ambiguous residues in the predicted putein was chosen as this is more likely to represent the founding insertion.

The initial classification of ERV families from relatively intact sequences introduces a bias toward better preserved and more recent insertions. To minimise the effect of this bias in defining the distribution of these lineages and provide a more general overview of the ERV distribution among the various families, BLASTN (Altschul et al., 1990, Zhang et al., 2000) was then used to expand the defined lineages to include less intact ERV sequences. To minimise the chance of spurious matches to classified sequences, queries were restricted to insertions with both LTRs present. Classified insertions were used as the search database and output was restricted to the sequence providing the best overall alignment with a minimum length of 400 bp to reduce the presence of short, spurious hits.

7.2.2: Placement of previously described crocodilian ERVs within the ERV phylogeny

The associations of previously described crocodilian ERVs with the insertions recovered from the genome were determined using BLASTN (Altschul et al., 1990, Zhang et al., 2000). Searches with the *pro-pol* fragments described in Chapters 2 and 3 were conducted against the complete proviral sequences of the insertions assigned to each lineage due to the nature of the primer sequences. Similarly, the lineage containing the KIT-ligand-like ORF was identified using the ORF sequence as determined in Chapter 4. To confirm that this acquisition was a lineage specific event rather than an insertion specific occurrence, this search was extended to include the predicted crocodilian KIT-ligand transcripts from the genome annotations using TBLASTX (Altschul et al., 1997).

7.2.3: Investigation of ERV expansions

Most ERV families will undergo low levels of replication within the genome and there is little to be gained from assessing selective pressures and mode of replication for all of these as they are less likely to have a significant impact on genomic function. However, a number of ERV families appeared to have undergone a greater degree of replication within the crocodile genomes. These were investigated further to determine the likely mechanism of replication and the evolutionary pressures acting on these insertions.

To clarify the likely evolutionary relationships between the insertions of each family, and infer the likely model of evolution, phylogenetic trees were created for the coding sequences of each ERV family. Nucleotide sequences of the internal domains of the complete ERV insertions for each family were aligned using the E-INS-i algorithm in MAFFT as described previously. Phylogenies were created using PhyML (Guindon et al., 2010) and the best suited nucleotide evolution model was determined by ModelGenerator (Keane et al., 2006). The likely model of evolution was inferred based on the overall topology of the resulting phylogenies.

7.3: Results

7.3.1: Classification of crocodilian ERVs

A total of 591 'complete' ERVs were recovered from the crocodilian genomes. After the removal of sequences that did not meet the criteria listed above, 320 sequences from the three genomes were used for the definition of ERV families. Of these 317 could be classified as ERV1, ERV3, ERV4 or *Gypsy*-like insertions. These sequences could be grouped into 66 families ranging from 1 to 64 sequences per family (Figure 7.1 and Figure 7.2). Families CrocERV1, CrocERV14 and CrocERV40 had much higher numbers of classified insertions and were investigated further as described in Section 7.3.2. The remaining three sequences shared very little similarity to retroviral sequences and were excluded from further analyses. Twenty one ERV1 families, 6 ERV3, 19 ERV4 and 5 *Gypsy*-like families were defined, as well as 15 families that appeared to be intermediates between the major ERV classes and could not definitively be placed (Table 7.1).

The distribution of each ERV family was predicted based on the species that the ERV sequences were recovered from. The majority of families were lineage specific, with 21, 14, and 11 families found only in *A. mississippiensis*, *C. porosus*, and *G. gangeticus*, respectively. A further 15 were found in both *C. porosus* and *G. gangeticus* (classed as *Longirostres* in Table 7.1 as defined by Harshman et al. (2003)), and two families were found in all three species. An additional two families were identified in *A. mississippiensis* and *C. porosus* only; and one family was found only in *A. mississippiensis* and *G. gangeticus*. Given that *Alligatoridae* is basal to the other two crocodilian families, it is unusual to observe lineages showing this distribution, although it is possible that other less intact insertions are present in the third species.

BLAST searches against less intact sequences suggested that most families had maintained relatively low levels of replication with small to moderate numbers of insertions detected (Table 7.1). However, five families (CrocERV1, 13, 14, 27, and 40), including three with large numbers of complete ERVs, appeared to have undergone much higher levels of replication with 50 or more related insertions detected. Comparisons of the distribution of these less complete ERVs across the three species suggested that many of the families identified may in fact be older, showing wider distributions than initially suggested by analysis of the complete ERV sequences.

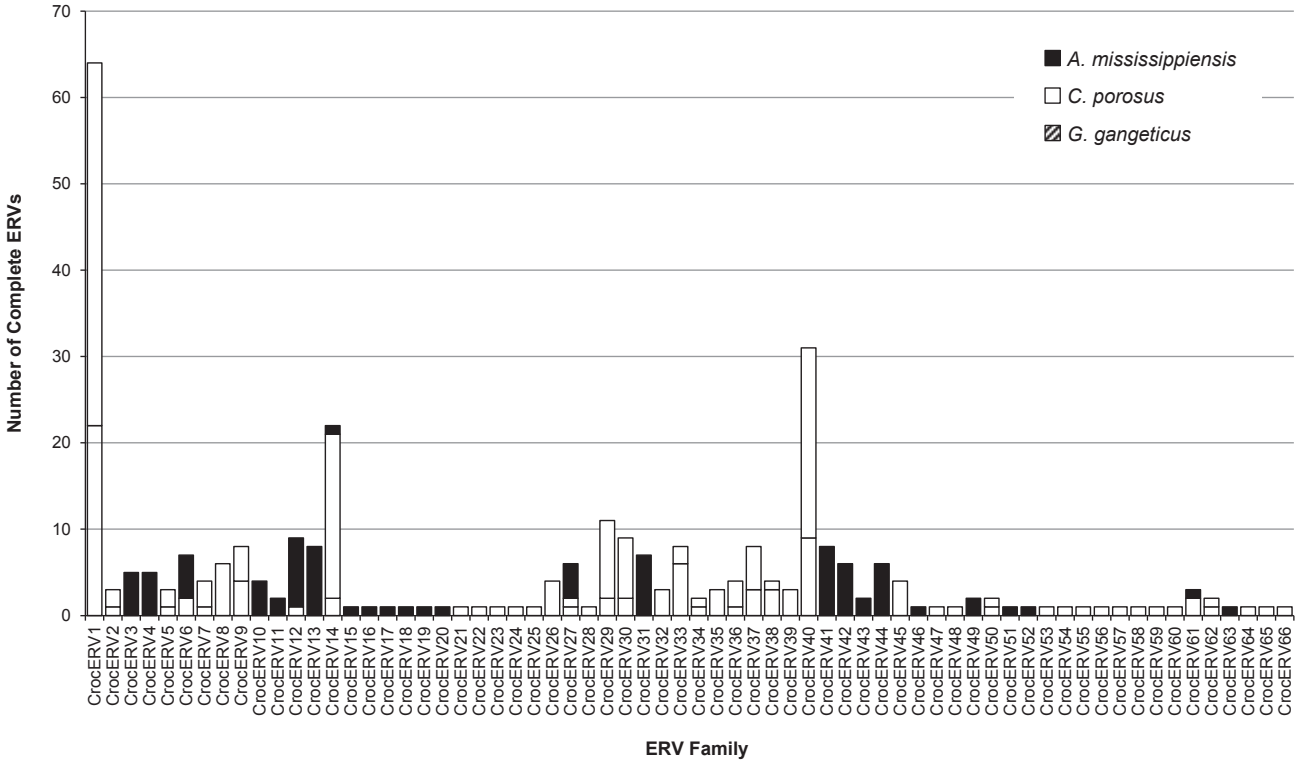


Figure 7.1: Distribution of the 'complete' ERV sequences from each species within each of the defined ERV families.

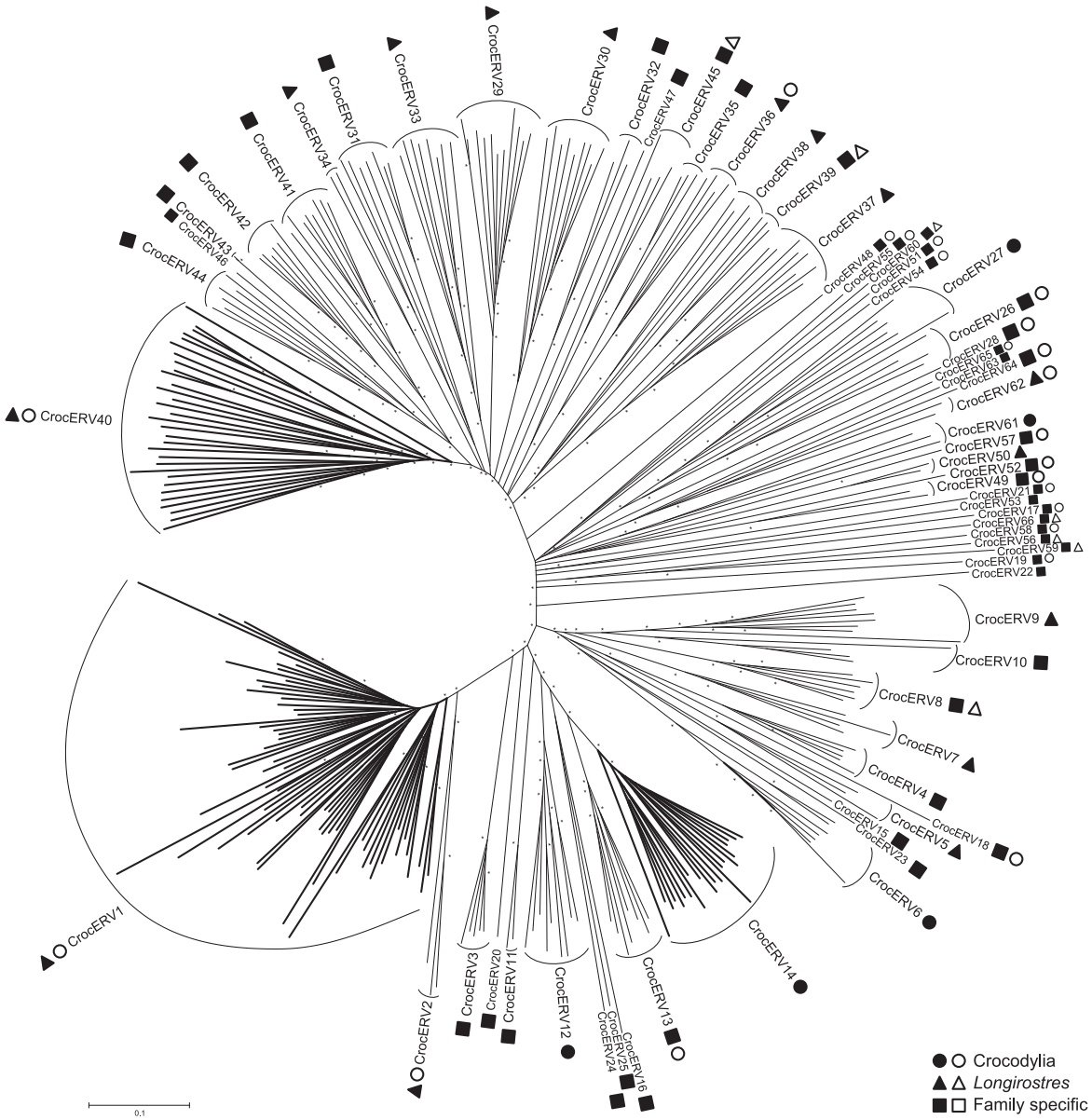


Figure 7.2: Clustering and distribution of sequences within the ERV families as defined using the complete ERV sequences. Specific families of interest are indicated by bold branches. An unrooted Neighbour Joining tree was created from amino acid alignments of the *pol* domain. Sequence IDs have been removed for clarity and only the family names are shown. Symbols indicate distribution of sequences across the three species: 'Crocodylia' refers to those found in all three crocodilian families, '*Longirostres*' to those found in *Crocodylidae* and *Gavialidae*, and 'Family specific' to those found in one of the three crocodilian families. Solid symbols indicate complete ERV sequences and unfilled symbols represent the less intact sequences. Scale bar on the left indicates branch length and the asterisks within the tree indicate bootstrap support values greater than 70%.

Table 7.1: Summary of ERV families and their predicted distribution among crocodilians. The distribution Crocodylia refers to families found in all three species whereas *Longirostres* refers to those only found in both *C. porosus* and *G. gangeticus*. *Alligatoridae*, *Crocodylidae*, and *Gavialidae* refer to the predicted distributions of families found only in a single species.

Family	ERV Class	Complete ERVs		2 LTRs	
		Distribution	No. seqs	Distribution	No. seqs
CrocERV1	ERV1	*Longirostres*	64	Crocodylia	279
CrocERV2	ERV1	*Longirostres*	3	Crocodylia	24
CrocERV3	ERV1	*Alligatoridae*	5	*Alligatoridae*	5
CrocERV4	ERV1	*Alligatoridae*	5	*Alligatoridae*	12
CrocERV5	ERV1	*Longirostres*	3	*Longirostres*	9
CrocERV6	ERV1	Crocodylia [a]	7	Crocodylia [a]	29
CrocERV7	ERV1	*Longirostres*	4	*Longirostres*	10
CrocERV8	ERV1	*Crocodylidae*	6	*Longirostres*	19
CrocERV9	ERV1	*Longirostres*	8	*Longirostres*	31
CrocERV10	ERV1	*Alligatoridae*	4	*Alligatoridae*	11
CrocERV11	ERV1	*Alligatoridae*	2	*Alligatoridae*	4
CrocERV12	ERV1	Crocodylia [a]	9	Crocodylia	32
CrocERV13	ERV1	*Alligatoridae*	8	Crocodylia	56
CrocERV14	ERV1	Crocodylia	22	Crocodylia	135
CrocERV15	ERV1	*Alligatoridae*	1	*Alligatoridae*	8
CrocERV16	ERV1	*Alligatoridae*	1	*Alligatoridae*	6
CrocERV17	Unknown	*Alligatoridae*	1	Crocodylia	46
CrocERV18	ERV1	*Alligatoridae*	1	Crocodylia [a]	11
CrocERV19	Unknown	*Alligatoridae*	1	Crocodylia	3
CrocERV20	ERV1	*Alligatoridae*	1	*Alligatoridae*	8
CrocERV21	Unknown	*Crocodylidae*	1	*Crocodylia*	23
CrocERV22	Unknown	*Crocodylidae*	1	*Crocodylidae*	6
CrocERV23	ERV1	*Crocodylidae*	1	*Crocodylidae*	4
CrocERV24	ERV1	*Crocodylidae*	1	*Crocodylidae*	1
CrocERV25	ERV1	*Gavialidae*	1	*Gavialidae*	1
CrocERV26	ERV3	*Gavialidae*	4	Crocodylia	44
CrocERV27	ERV3	Crocodylia	6	Crocodylia	95
CrocERV28	Unknown	*Crocodylidae*	1	Crocodylia	10
CrocERV29	ERV4	*Longirostres*	11	*Longirostres*	19
CrocERV30	ERV4	*Longirostres*	9	*Longirostres*	24
CrocERV31	ERV4	*Alligatoridae*	7	*Alligatoridae*	18
CrocERV32	ERV4	*Crocodylidae*	3	*Crocodylidae*	7
CrocERV33	ERV4	*Longirostres*	8	*Longirostres*	20
CrocERV34	ERV4	*Longirostres*	2	*Longirostres*	5
CrocERV35	ERV4	*Crocodylidae*	3	*Crocodylidae*	3
CrocERV36	ERV4	*Longirostres*	4	Crocodylia	21

Family	ERV Class	Complete ERVs		2 LTRs	
		Distribution	No. seqs	Distribution	No. seqs
CrocERV37	ERV4	*Longirostres*	8	*Longirostres*	21
CrocERV38	ERV4	*Longirostres*	4	*Longirostres*	6
CrocERV39	ERV4	*Gavialidae*	3	*Longirostres*	4
CrocERV40	ERV4	*Longirostres*	31	Crocodylia	81
CrocERV41	ERV4	*Alligatoridae*	8	*Alligatoridae*	16
CrocERV42	ERV4	*Alligatoridae*	6	*Alligatoridae*	38
CrocERV43	ERV4	*Alligatoridae*	2	*Alligatoridae*	6
CrocERV44	ERV4	*Alligatoridae*	6	*Alligatoridae*	24
CrocERV45	ERV4	*Gavialidae*	4	*Longirostres*	8
CrocERV46	ERV4	*Alligatoridae*	1	*Alligatoridae*	6
CrocERV47	ERV4	*Crocodylidae*	1	*Crocodylidae*	2
CrocERV48	Unknown	*Gavialidae*	1	Crocodylia	6
CrocERV49	ERV3	*Alligatoridae*	2	Crocodylia	12
CrocERV50	ERV3	*Longirostres*	2	*Longirostres*	5
CrocERV51	Unknown	*Alligatoridae*	1	Crocodylia [a]	4
CrocERV52	ERV3	*Alligatoridae*	1	Crocodylia [a]	3
CrocERV53	Unknown	*Crocodylidae*	1	*Crocodylidae*	2
CrocERV54	Unknown	*Crocodylidae*	1	Crocodylia [a]	19
CrocERV55	Unknown	*Crocodylidae*	1	Crocodylia	4
CrocERV56	Unknown	*Crocodylidae*	1	*Longirostres*	6
CrocERV57	ERV3	*Gavialidae*	1	Crocodylia	15
CrocERV58	Unknown	*Gavialidae*	1	Crocodylia	8
CrocERV59	Unknown	*Gavialidae*	1	*Longirostres*	8
CrocERV60	Unknown	*Gavialidae*	1	*Longirostres*	3
CrocERV61	Gypsy-like	Crocodylia [a]	3	Crocodylia	33
CrocERV62	Gypsy-like	*Longirostres*	2	Crocodylia	7
CrocERV63	Gypsy-like	*Alligatoridae*	1	*Alligatoridae*	1
CrocERV64	Gypsy-like	*Crocodylidae*	1	Crocodylia	15
CrocERV65	Gypsy-like	*Gavialidae*	1	Crocodylia	6
CrocERV66	Unknown	*Gavialidae*	1	*Longirostres*	3

[a] Insertions were detected from *Alligatoridae* and either *Crocodylidae* or *Gavialidae*, suggesting either absence of detected insertions meeting the criteria within the third species, or that ERVs insertions were the result of multiple infections

7.3.2: Comparisons with crocodilian ERVs from previous studies

The association of the defined ERV families with previously described ERV sequences produced varying results. The *Gammaretrovirus*-like ERV1 lineage identified by Jaratlerdsiri et al. (2009) and in Chapter 2 corresponded to family CrocERV6, which appears to be present in *Alligatoridae* and *Crocodylidae*, with a possible distribution across all three crocodilian species. The *Epsilonretrovirus*-like lineage represented by haplotype 58 from Chapter 2 corresponds to CrocERV8 and appears to be specific *to C. porosus*. Unexpectedly, the ERV3 lineage identified in *C. johnstoni* in Chapter 3 showed similarity to the two major ERV3 families, CrocERV26 and CrocERV27. Given the similarity between the *pol* domains of these two families, it is possible that they are closely related, although the more variable domains suggest that these are in fact separate lineages. In general, the ERV4 *pro-pol* fragments varied more with fragments distributed among most of the ERV4 families. In particular the complete ERV sequences recovered from *C. niloticus* by Martin et al. (2002) corresponded with family CrocERV31.

The KIT-ligand containing ERV4 insertions described in Chapter 4 were found within family CrocERV40. Regions sharing similarity to the KIT-ligand transcript were recovered from all complete sequences from this family. A small number of these insertions appear to have maintained relatively intact ORFs, including one sequence from which the intact ORF was recovered (Figure 7.3). Subsequent alignment of this sequence with the consensus sequence from Chapter 4 suggests that this insertion is very likely to be the same one that was sequenced from the BAC clones (99.2% similarity across the entire sequence, 100% similarity across internal domains), confirming the hypothesis that the ERV4 lineage from the BAC clones was derived from a single insertion that was well represented in the shared paired-end libraries.

7.3.3: Detailed investigation of large ERV families

Three families (CrocERV1, CrocERV14 and CrocERV40) appear to have undergone a greater degree of replication within the crocodilian genomes with over 20 relatively complete insertions included and over 80 less intact insertions compared to the other families. All putein sequences contained stop codons, with a number of domains from all three families sharing stop codons within at least some of the sequences. No sequences from the coding domains appeared to encode an intact ORF.

Phylogenies for the two ERV1 families were characterised by short internal branches and longer terminal branches, typical of a rapid proliferation followed by divergence of the insertions by mutation (Figure 7.3a and b). On the other hand, the ERV4 families displayed longer branch lengths overall, suggesting a much slower rate of replication and evolution (Figure 7.3c). All families displayed multiple clades of sequences within the phylogenies, indicating bursts of replication followed by periods of differentiation. None of the ERV sequences examined appeared to have originated from regional duplications of the crocodilian genomes.

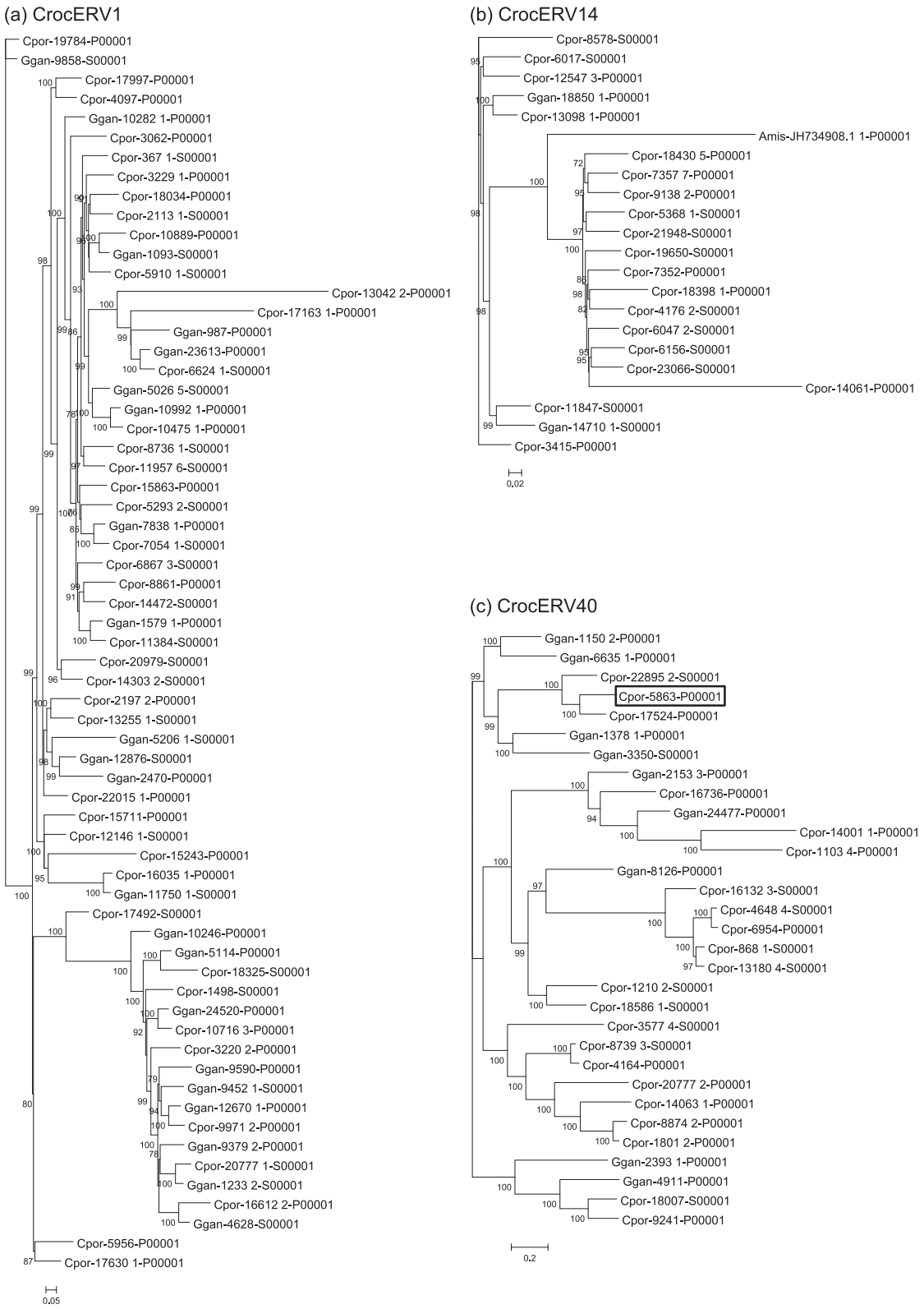


Figure 7.3: Three ERV families showing multiple bursts of radiation resulting in a large number of insertions within the crocodilian genomes. Phylogenetic trees were created from the internal domains of the three ERV families of interest. The ERV insertion encoding a complete KIT-ligand ORF is highlighted by a box. Numbers within the trees indicate statistical support for the branches and the scale bars indicate branch length.

7.3.4: Interspecies comparisons

Phylogenetic comparisons with published ERV *pol* sequences (Jern et al., 2005) largely supported previous observations that the ERV families defined in this chapter are primarily made up of families from the ERV4, ERV3, ERV1 and *Gypsy*-like ERV classes, with a number of intermediate lineages (Figure 7.4). While phylogenetic analyses occasionally grouped these intermediary sequences within one of the major clusters, statistical support for these groupings were very low, and placements were inconsistent. Interestingly, only three ERV families were closely associated with mammalian *Gammaretrovirus*-like ERVs, while the remaining ERV1 families appeared to be more divergent *Epsilonretrovirus*-like lineages.

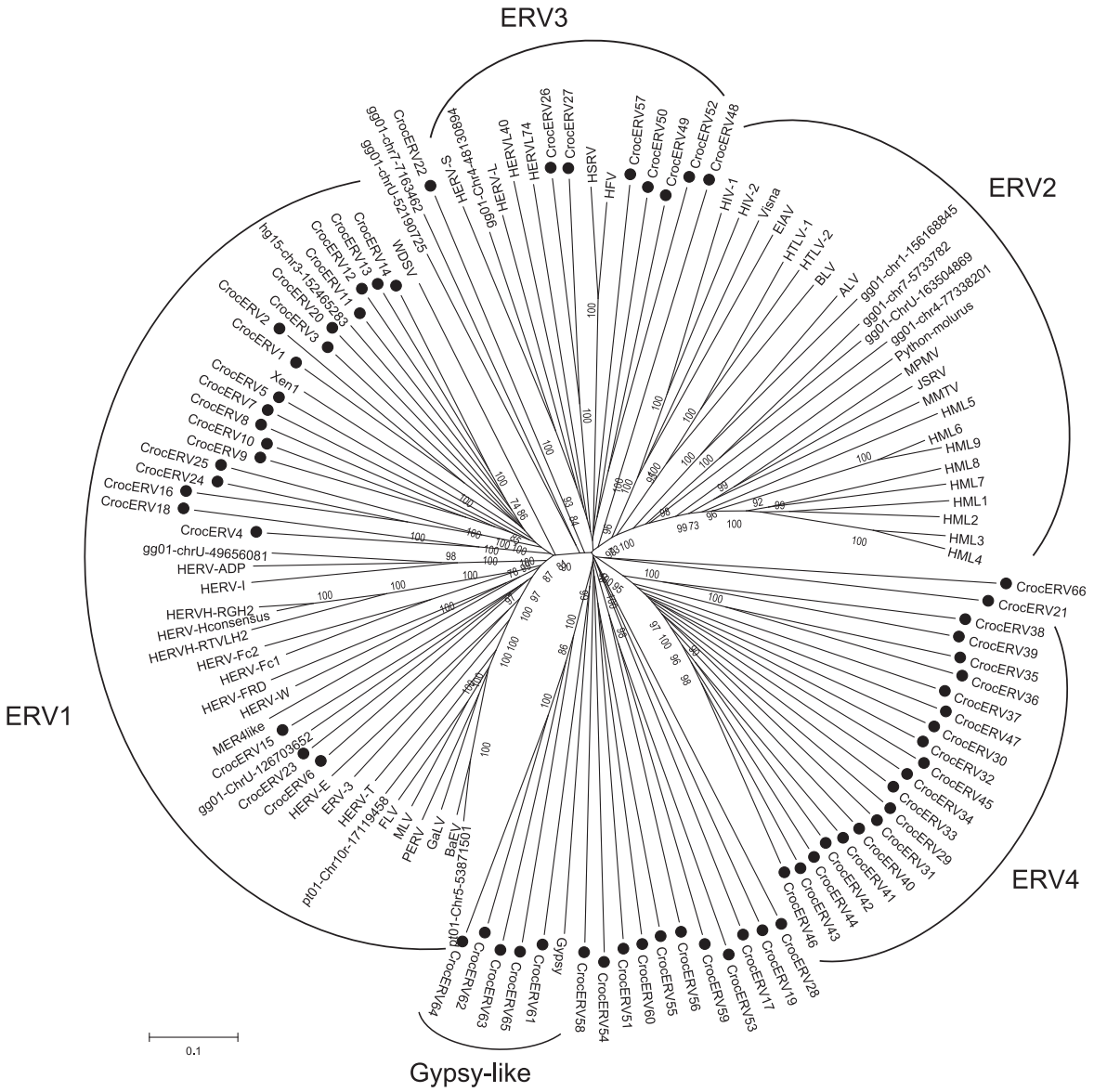


Figure 7.4: Classification and likely relationships between the ERV families and ERVs from other species based on representative sequences and a selection of previously published ERV sequences (Jern et al., 2005). An unrooted Neighbour Joining tree was created from amino acid alignments of the *pol* domain. ERV families defined in this chapter are indicated by a black circle. The scale bar indicates relative branch lengths and the numbers within the tree indicate bootstrap support greater than 70%.

7.4: Discussion

7.4.1: Crocodilian ERVs may represent ancestral retroviral states

Crocodilian ERV families belong to a limited subset of known ERV classes. Excluding intermediates and unassigned families, the crocodilian ERV sequences were distributed among ERV1, ERV3, and ERV4, as well as a number of *Gypsy*-like lineages. There were no sequences or families that showed similarity to ERV2. Unsurprisingly, ERV1 and ERV4 were the predominant lineages in the genomes, with 21 and 19 families respectively. This is in agreement with findings by Jaratlerdsiri et al. (2009) who observed a widespread distribution of insertions related to these lineages across a number of crocodilian species. It is also supported by the surveys within species (Chapters 2 and 3) where only single *Gypsy*-like and ERV3 insertions were isolated in *C. porosus* and *C. johnstoni* compared to numerous ERV1 and ERV4 insertions.

The crocodilian ERV families displayed very weak associations with ERVs from other taxa. Notably, the crocodilian ERVs tended to cluster separately from mammalian ERVs. The current findings are in accordance with previous studies where it was suggested that phylogenetic and evolutionary distance between potential host species may affect the potential distribution of ERV and retroviral linages (Herniou et al., 1998, Martin et al., 1999). As discussed in Chapter 6, genome biology and the genomic environment may also play a role in determining the ERV complement of a genome. These limitations suggest that the distinction between crocodilian and mammalian ERVs is the result of co-evolution between retroviruses and their host lineages.

Many of the crocodilian ERV families also appeared to cluster separately from exogenous retroviral genera suggesting that these may represent ancient ERV insertions that are intermediates or novel lineages between the currently recognised taxa. These findings support the theory that the currently recognised exogenous retroviral genera represent a process of gradual evolution within the three broader branches of the retroviral phylogeny rather than seven separate evolutionary lineages (Herniou et al., 1998, Jern et al., 2005). Notably, further investigation of the *Epsilonretrovirus*-like ERV families did not recover any additional ORFs, nor regions sharing similarity to accessory genes from either the Walleye retroviruses (WDSV and Walleye epidermal hyperplasia virus; WEHV) (Holzschu et al., 1995, LaPierre et al., 1999) or Xen1 (Kambol et al., 2003). Therefore it is possible that these crocodilian

ERV families represent an infection by an early predecessor of the modern *Epsilonretroviruses*.

Retroviruses, ERVs, and related LTR retrotransposons share some common functional domains, the most notable of these being the *pol* domain. It has been proposed that retroviruses and ERVs evolved from LTR transposons via the acquisition of an *env* domain, facilitating extracellular movement and production of infectious particles (Xiong and Eickbush, 1990). It has been proposed that *Gypsy* transposons, such as the *Tf1*/*sushi* group may potentially represent the predecessor of retroviruses and their ERV counterparts (Butler et al., 2001). These transposons do not appear to have an exogenous phase in vertebrates, although many families encode on ORF similar to the retroviral *env* domain (Boeke and Stoye, 1997). Further investigations to identify similar lineages within other vertebrates, particularly the basal taxa such as fish, amphibians, and reptiles, may shed light on the distribution of these intermediate elements, and the evolution of retroviruses.

7.4.2: Crocodilian genomes are host to a larger than expected number of ERV families

The analyses conducted in this chapter have revealed the presence of a large number of ERV families in all three crocodilian genomes. Based on classification of relatively intact ERV sequences, the *A. mississippiensis* genome contains 26 ERV families, *C. porosus* contains 33, and *G. gangeticus* contains 29 families. Of these, two families are common to all three genomes, and 15 are shared between *C. porosus* and *G. gangeticus*. While the number of putative ERV families among non-avian reptilian taxa has not been calculated, these values are significantly higher than those reported in RepBase for *Danio rerio* (zebrafish; 18), *Xenopus tropicalis* (western clawed frog; 9) (Jurka et al., 2005), and *Gallus gallus* (chicken; 17) (Huda et al., 2008). Contrary to this, these values are comparable to some mammalian taxa (24 in *Bos taurus*; European cattle (Garcia-Etxebarria and Jugo, 2010), 20 in *Mus musculus;* mouse (McCarthy and McDonald, 2004), and 31 in humans (Katzourakis et al., 2005)), although much lower than other mammals, such as 42 in *Pan troglodytes* (chimpanzee) (Polavarapu et al., 2006). These findings suggest that crocodilian genomes, and possibly those of many other reptilian species, may contain a more similar number of ERV lineages to mammalian taxa than avian, amphibian or piscine species.

One of the major challenges in defining and characterising crocodilian ERV families is the divergence and degradation of ancestral ERV sequences. To combat this, the ERV families described here have been defined from relatively complete ERV sequences. This conservative approach minimises the likelihood of incorporating false positives into subsequent analyses at the risk of underestimating the true number of ERV families present. When these families were expanded to include more degenerate copies of ERV insertions, approximately one third of the ERV insertions incorporated in this manner were not able to be placed within one of the defined families, suggesting that if these were to be classified in a similar manner, additional families could be defined. However, it is unlikely that these additional families would provide further insights into the mechanisms or impacts of ERVs on the genomes of these species, as they are more likely to be highly degraded and therefore non-functional.

7.4.3: Distribution of ERV families among crocodilians

A large number of ERV families were recovered from all three genomes although insertions showed varying degrees of degradation. Very few of these families were present as relatively complete insertions suggesting that the crocodilian genomes have been subject to a large number of ancestral infection events followed by subsequent degradation of silenced ERV copies. Surprisingly, a large number of ERV families from ERV1 and ERV4 were found to be species specific, even when the search was expanded to include less intact ERVs. This implies that the exogenous retroviruses that gave rise to these endogenous families were active relatively recently in crocodilian evolution. This is particularly significant for the ERV4 lineage as no exogenous counterpart has been described for these proviruses. As such, it would be interesting to obtain an estimated infection date for these elements and investigate their distribution among crocodilians and other vertebrate taxa.

The species specificity identified may be due, in part, to the high similarity of the ERV4 *pro-pol* fragments (Chapters 2 and 3) to the same region across many of the ERV4 families. Based on current knowledge of these families, this similarity suggests that these families may be the result of a single ancestral infection event followed by host family specific replication and evolution. If this is the case, this replication is unusual for ancient ERVs, as it would be expected that insertions with a detrimental effect on the host would be quickly silenced or removed from the population (Barr et al., 2005, Gifford and Tristem, 2003). That these

insertions have maintained some capacity for replication suggests a low level of pathogenicity and virulence (Blikstad et al., 2008). This in turn may also support the suggestion that these elements represent a less pathogenic pre-cursor to modern retroviruses.

Another surprise was the recovery of ERV1 families from *A. mississippiensis*. Given the widespread distribution of ERV1 lineages among vertebrate taxa, it is plausible that some ERV1 families would be isolated from *Alligatoridae*. However, the observation of *A. mississippiensis* lineages from family CrocERV6 was unexpected, as this lineage had previously been thought to be specific to *Crocodylidae* and *Gavialidae* (Chapter 2) (Jaratlerdsiri et al., 2009). The alignment and subsequent phylogeny of sequences within this lineage show that *A. mississippiensis* sequences appear to form a separate sublineage to those from *C. porosus*, suggesting either the presence of species or host family specific sublineages within this ERV family, or concurrent infection by closely related strains of the same exogenous retrovirus. Subsequent comparisons within less intact ERVs revealed that the *C. porosus* sublineage is likely to be specific *to C. porosus*, and may represent the lineages previously identified in *Crocodylidae*, while the *A. mississippiensis* lineage appears to be shared between all three species.

7.4.4: Mechanisms of ERV replication

Most of the ERV families defined herein appear to have undergone very low levels of replication, with only a few families represented by more than 50 insertions across the three genomes, even when less intact sequences were included. A small number of families appear to have undergone a greater degree of replication. The reasons behind this disparity is unclear, although differences in pathogenicity and virulence of the infecting exogenous retroviruses might be a contributing factor (Blikstad et al., 2008). However, many families also appear to be remnants of ancient retroviral infections, predating divergence of the major crocodilian lineages. Thus, it is also possible that there are more degenerate insertions present that were not detected or included due to accumulation of mutations or loss of coding domains.

Three of the families showing greater replication were investigated further and the topology of their phylogenies imply that all expanded through multiple bursts of replication followed

by periods of differentiation, indicating that ERV replication is likely to follow either the transposon or random template models of replication. A comparison of the relative branch lengths between families (Figures 7.2 and 7.4) indicates that the ERV4 family (CrocERV40) has maintained a more steady rate of replication and divergence than the two ERV1 families, as internal branches are much longer relative to the terminal branches (Cordaux et al., 2004, Katzourakis et al., 2005). On the other hand, the two ERV1 families display very short internal branches across many of the clades, characteristic of rapid replication followed by periods of reduced activity (Cordaux et al., 2004, Katzourakis et al., 2005).

It has been suggested that ERV activity and replication within the host genome corresponds to radiation of the host taxa, and therefore, that ERV dynamics may mirror the population dynamics of the host (Gherman et al., 2007, Romano et al., 2007). None of the three families studied in detail showed evidence for host specific divergence or replication, further supporting the concept that these represent ancestral radiation events. Thus it is possible that the observed bursts of replication correspond to significant periods of radiation and speciation in ancient crocodilians. Given an estimated nucleotide substitution rate for crocodilians, it would be possible to date these insertions based on divergence of the LTRs (Garcia-Etxebarria and Jugo, 2010, Romano et al., 2007), and it would be interesting to examine whether these radiation events in the ERV phylogeny correspond with divergence of the respective host taxa.

Despite these bursts of replication, the accumulation of stop codons within the coding domains of these ERVs suggests that they are no longer capable of replicating autonomously. Autonomous replication either by reinfection or retrotransposition requires that some, if not all, of these domains retains some degree of functionality. The large number of shared stop codons within the *env* domain of all three families implies that retrotransposition is the primary mechanism of replication, with individual insertions then silenced by accumulation of other silencing mutations in other domains. Interestingly, shared stop codons found in these other domains indicate that other replication mechanisms have also had an impact on some ERV elements. These mechanisms may include complementation by a functional virus or duplication of genomic regions (Gifford and Tristem, 2003). As no evidence for regional duplications were found, it is more likely that these insertions are the result of replication through complementation.

Loss of a functional *env* domain appears to permit a greater level of replication within the host genome and has been observed in a number of mammalian ERV families. One leading example is the IAP superfamily that has been observed primarily in rodents, as well as Lagomorpha (rabbits, hares, and pikas) and some primates (Magiorkinis et al., 2012). These Class II ERVs have a highly degraded *env* domain (Mietz et al., 1987) necessitating replication by retrotranposition rather than reinfection, and account for up to 75% of ERV loci in rodent genomes (Magiorkinis et al., 2012). Other notable examples include the HERV-L family and related ERV-L elements (Benit et al., 1999, Magiorkinis et al., 2012), and the HERV-K, -H, and W families which have replicated through a combination of retrotransposition and complementation but display a low level of reinfection (Belshaw et al., 2005b).

It is unclear whether the loss of the *env* domain is a consequence or a result of a transition to intracellular replication methods such as retrotransposition and complementation. Furthermore, it is not clear why the loss of a functional *env* promotes greater proliferation within the host genome, although a switch to retrotranposition may facilitate replication within germline cells while replication through reinfection may lead to greater levels of somatic cell integration. Thus reinfection may result in an overall reduction in host fitness through the greater chance of insertional mutagenesis (Magiorkinis et al., 2012). Alternatively loss of *env* and transition to intracellular replication may allow the virus to evade innate host anti-viral defences (Magiorkinis et al., 2012), and thus these elements are more likely to be observed within the genome.

It was also interesting to note that one of these crocodilian ERV families appears to have captured a host KIT-ligand mRNA (see also Chapter 4, Section 4.4.5). While selection on this additional ORF was not assessed, this lineage appears to have captured the KIT-ligand mRNA relatively early in its integration cycle, with remnants of this ORF present in all complete insertions assigned to this family. The recovery of an intact ORF in *C. porosus* suggests that there may be some functional significance associated with it in this species. If this is the case, it would be expected that this insertion or the captured mRNA would be expressed in at least some tissues or at some stage in the development of *C. porosus*.

The KIT-ligands play an important role in a variety of functions ranging from gametogenesis, melanogensis and haematopoesis (Huang et al., 1992). Acquisition of a second, homologous

sequence has been described in *D. rerio* (Hultman et al., 2007) where these genes have co-evolved to share complementary functions and display tissue specific expression patterns. A similar duplication has been observed in *G. gangeticus* (ICGWG, unpublished data) although the functional significance has not yet been established.

At this point in time, there is limited knowledge of the role and effects of KIT-ligand and these potential duplications in crocodilians. Due to the limited annotation of the crocodilian genomes, it is still unclear whether the ERV insertion containing the complete ORF is in a region of the genome likely to promote transcription. As further genomic resources become available for crocodilians, it will be interesting to establish if and where this ORF is being expressed, and what potential functional role this may have in *C. porosus*.

7.5: Conclusions

Overall, the data generated in this chapter indicate that crocodilian ERVs stem from infection events by retroviruses from a wide variety of lineages. These data provide a framework to facilitate further studies into crocodilian ERV diversification as well as other basal vertebrate species. Distributions of the ERV families across the sequenced crocodilian taxa suggest that most of these are ancient integration events predating the divergence of the crocodilian families. The recovery of apparent intermediates between the major ERV classes and recognised retroviral genera lends support for a gradual evolution of the exogenous retroviral genera recognised today, further highlighting the need for detailed studies into the ERVs of the basal vertebrate families.

There is evidence that a small number of these families have undergone significant levels of replication within crocodilian genomes at some stage in their evolution. Using the resources generated here, it will be possible to extend ERV studies in crocodilians to assess the interactions of these ERVs with the crocodilian genomes, and the roles they may play in the biology of these species. Further investigation into the demographics of these ERVs may provide insights into the population demographics of ancient crocodilians and corroborate molecular and fossil evidence of crocodilian radiation. In particular, the capture of a host mRNA by an ERV insertion followed by the subsequent replication of this family merits

further investigation, and highlights the potential impacts and significance of ERV replication and maintenance in crocodilians.

Chapter 8: General Discussion

8.1: Crocodilian ERVs provide a unique perspective on ERV evolution

One of the main recurring findings from these studies has been the diversity of ERVs present in crocodilians. From this, it is evident that crocodilians and other non-mammalian vertebrate genomes may be host to a large repertoire of previously uncharacterised ERV diversity. Early analyses of crocodilian genomes (Chapters 2, 3, and 4) suggested that there were at least five distinct groups of ERVs present in crocodilians, belonging to three major ERV classes: ERV1, ERV3, and the newly recognised ERV4, as well as a number of intermediary families that share characteristics of two or more ERV classes. The availability of the crocodilian genome sequences and subsequent analyses of these genomes (Chapters 6 and 7) indicate that not only are crocodilians host to a wide range of ERVs, but many of these ERVs are well preserved copies of ancient insertions.

Of particular interest is the range of novel ERV families and intermediate lineages recovered from crocodilians, especially the discovery of the novel Class ERV4. Sequences from this class were previously characterised from *C. niloticus* where they were recognised as a divergent and possibly basal member of the *Retroviridae* (Martin et al., 2002). Subsequent analyses of these ERVs from other crocodilians have shown that remnants of these infections are found in all crocodilian species (Chapters 2, 3, and 7) (Jaratlerdsiri et al., 2009). However, the presence of shared functional motifs makes it difficult to determine whether they are the result of multiple infection events by related viral strains or the result of multiple episodes of replication and mutation. Regardless, the diversity and divergence of crocodilian ERVs suggests that non mammalian vertebrates provide a largely untapped resource for ERV studies and their impact on genome evolution and function.

These findings highlight the importance of non-mammalian vertebrates for evolutionary studies of ancient retroviruses and related retroelements. As remnants of an early strain of retrovirus, this class of ERVs is of particular evolutionary significance for studies into the origins of retroviruses and their subsequent evolution. Low mutation rates (Eo and DeWoody, 2010, Hugall et al., 2007, Hughes and Mouchiroud, 2001, Ray et al., 2004) and the slow evolution of crocodilian genomes (D. Ray, pers comm.) make these species particularly useful for further studies of ancient insertion events as they are well preserved, providing

there is no significant decrease in survival fitness for the host. Detailed investigations of these insertions may allow for the reconstruction of these ancient ERVs and shed light on the possible routes of evolution from LTR retroelements to viral particles with the capacity to infect new hosts.

8.1.1: Interaction between ERVs and crocodilian genomes

The impact that ERVs and other TEs have had on shaping the crocodilian genomes is still largely unknown, as is the impact of the genomic environment on the replication dynamics of these elements. The analysis of ERVs within the MHC region of *C. porosus* (Chapter 4) provides some initial insights into ERVs in a region that it known to be rich in repetitive sequences, and demonstrates evidence of ERV replication within that region, However, generating the ERV integration profile of the rest of the crocodilian genome is also important as these sequences have the potential to shape and alter the structure, stability, and transcription profile of surrounding genomic regions.

ERVs and other TEs are a major contributing factor to genome plasticity, through recombination, transposition, and gene shuffling (Jern and Coffin, 2008, Lower et al., 1996). In particular, recombination between ERVs has been associated with large scale chromosomal rearrangements, likely to play a role during speciation (Hughes and Coffin, 2001). Crocodilian karyotypes display a large amount of variation both in the number of chromosomes and the arrangement between chromosomes, although the make-up of these blocks of genomic material appear to be highly conserved (King et al., 1986). As the annotation and mapping of the crocodilian genomes progress, it will be interesting to establish the distribution of ERVs across chromosomes, particularly at these chromosomal breakpoints across species.

The orientation of ERV proviruses relative to genes and coding domains is another significant aspect of ERV-host genome co-evolution that should be explored, particularly with respect to the better conserved insertions. ERV LTRs in particular contain transcription promoters and other regulatory factors that have the potential to affect the expression of nearby genes by providing alternative transcription start sites or regulatory domains thereby altering the transcription profile of that region (Jern and Coffin, 2008, Rosenberg and

Jolicoeur, 1997). Orientation of proviruses relative to the surrounding genomic region can also affect the preservation or degradation of the insertion, with an observable bias in selection against sense insertions and better preservation of antisense insertions (Jern and Coffin, 2008).

8.1.2: Potential for ERV activity in the crocodilian genomes

Studies of the ERV insertions and families suggest that a small proportion of these may be capable of replication within crocodilians, either through reinfection, or transposition within the genome (Chapters 2, 3, 4, and 7). However, these data have not yet been correlated with crocodilian genome coding regions and, as such, the potential for each of these insertions to be transcribed is still unknown. As ERV transcription is reliant on host cellular mechanisms, the underlying genomic structure of surrounding regions, such as chromatin structure and chromosomal location, dictates the likelihood of ERV expression (Rabson and Graves, 1997). Thus, as the annotations of the crocodilian genome sequences progress, it will be important to map the locations of likely ERV sequences to the genomes to identify integration sites and the genetic make-up of the surrounding genome. As tissue specific RNAseq data becomes available to complement the genomic sequence data, it will also be possible to identify ERV transcription within particular tissues.

The recovery of a potentially exapted ERV in *C. porosus* (Chapters 4 and 7) is particularly interesting as it appears that the host mRNA transcript was captured at the time of the initial ERV integration. One insertion appears to have been maintained in *C. porosus* and, while the ERV appears to have been silenced, the captured ORF is still intact. Identification of transcripts from this ORF may shed light on the co-evolution of host and viral genes, and determine whether these genes are co-expressed or if tissue specific expression is evident.

8.1.3: Evolution of ERV silencing mechanisms

Vertebrate genomes have developed a variety of mechanisms for silencing ERVs and other autonomous TEs, including gene products that specifically target various components of the TE genomes or gene products, and epigenetic controls such as methylation. Some examples of these mechanisms in mammals include the tripartite motif containing 5 protein (TRIM5∝) which has been shown to restrict HIV-1 infection in primates, the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3) subfamily (Jern and Coffin, 2008), tandem zinc finger genes (Thomas and Schneider, 2011), and microRNAs (Cullen, 2006). However, while these genes and their activity against ERVs and related TEs are relatively well understood in mammalian genomes, equivalents have not yet been identified in reptiles.

ERV transcription and activity may also be restricted by transcriptional silencing either through methylation of the genomic region where they have integrated, or reliance on particular host transcription factors (Rabson and Graves, 1997). Thus, while there is no direct evidence to suggest that crocodilian ERVs are actively replicating, the possibility of replication competent insertions cannot be excluded. Evidence of recent ERV expansions and possible ERV activity (Chapters 2 and 7) suggest that some ERV families have retained some ability to replicate if released from transcriptional suppression. Thus, situations resulting in the relaxing of these suppression mechanisms, such as stress as a result of environmental conditions or disease, may release these ERVs from suppression, resulting in proliferation within the individual.

Given the divergence and diversity of ERVs observed in the crocodilian genomes, further research into the evolution of transcriptional control of ERVs and TEs may uncover novel mechanisms with significance for understanding both virus-host evolution and the development of novel anti-viral measures. Interrogation of the genome for genes encoding products with similar function to those described in mammals may shed light on the acquisition and evolution of these defence mechanisms in vertebrates. It is also possible that non-mammalian vertebrates have acquired or retained a separate set of anti-viral and ERV suppression mechanisms that has not yet been identified in mammals.

8.2: Technical perspectives

8.2.1: A defined structure for nomenclature of crocodilian ERVs

During the course of these studies, the need for an unambiguous, consistent system for naming ERVs was identified. Early studies (Chapter 2) (Jaratlerdsiri et al., 2009) used the name "CERV" to denote "Crocodilian ERV". However, deeper searches of ERV literature revealed that this acronym had previously been used to denote ERVs isolated from *Pan troglodytes* (chimpanzee) (Hughes et al., 2005, Polavarapu et al., 2006, Shih et al., 1991). In an attempt to resolve this and avoid further confusion, "CrocERV" has subsequently been used as a general term for crocodilian ERVs. ERV insertions from a particular species are identified by a four letter species designation based on the first two letters of the genus and species names followed by "ERV"; for example: "CrpoERV" for *C. porosus*, "CrjoERV" for *C. johnstoni*, "AlmiERV" for *A. mississippiensis*, and "GagaERV" for *G. gangeticus*.

The use of PCR to isolate sequence fragments, and the subsequent need to differentiate between these fragments and ERV insertions resulted in a further tier of identification. Under this scheme fragments should be identified by providing the gene designation, such as *pro-pol* or *pol*, followed by a unique identification number. The PCR data presented in Chapters 2 and 3 refer primarily to sequence haplotypes, and so were numbered numerically. For complete insertions, as identified in Chapters 4, 6 and 7, unique identifiers based on scaffold should be used in lieu of chromosome location as the genome sequences have not yet been mapped.

8.2.2: Relative merits of different techniques for ERV studies

A wide variety of tools are available for the study of ERV sequences, many of which form the basis of the studies presented in the previous chapters. These include PCR, high stringency hybridisation, and bioinformatics-based mining of genomic sequence data (see also Chapter 1, Section 1.4.3). While the relative merits of different ERV detection software have previously been discussed with regards to analysis of whole genome sequencing projects (Chapter 5), it is also important to consider the use of various data generation techniques for ERV characterisation and the study of ERV diversity.

Increasing popularity of whole genome sequencing and improvements in sequencing technologies have made genomic analyses a more accessible and attractive option for the wide-scale identification and classification of ERVs. This is particularly evident from the wide range of bioinformatics tools available for such analyses (Chapters 5 and 6). Such mining of information from genomic resources provides a rich source of sequence data for the classification and subsequent characterisation of ERVs from vertebrate genomes, and an important resource for the discovery of novel ERV diversity (Chapter 7). Additionally, whole genome sequences allow for the mapping of ERV sequences to specific locations of the genome and analysis in relation to the surrounding genomic environment, thereby facilitating studies into ERV-host co-evolution and the interactions between ERVs and their host genome (see also Section 8.1.1 above) (Gifford and Tristem, 2003).

On the other hand, the associated (computational) resource costs of *de novo* assembly and annotation of newly sequenced genomes make it unfeasible for such projects to be carried out for the primary purpose of studying ERV diversity in a species. Instead, genome-wide ERV studies are better carried out in conjunction with, or as a part of, a larger genome sequencing project (Gifford and Tristem, 2003). In the absence of such a project, wide-scale analyses may be facilitated by the use of genomic BAC libraries where available, using strategies similar to those used in Chapter 4, or the creation of Cot-libraries (Choulet et al., 2010, Peterson et al., 2002, Wicker et al., 2006).

The data generated in Chapter 4 suggest that BAC libraries also represent a useful resource for the study of specific ERV families or insertions. While the preparation of these libraries is a time consuming process, there are a large number of species where these have already been constructed. High stringency hybridisation using appropriate probes can be used to identify regions of the genome containing the family of interest, and densitometric analyses of the resulting arrays provides a means to approximate the ERV content of these genomes (Chapter 4) (Magbanua et al., 2011, Shan et al., 2009). Furthermore, targeting of specific insertions in this manner, combined with next generation sequencing technologies, facilitates the analysis of regions surrounding the insertion of interest without the need for techniques such as primer walking or inverse PCR.

In spite of technological advances and a shift towards genome level studies, PCR surveys still offer valuable insights into ERV evolution in situations where the previous technologies

would be unsuitable. In the absence of a sequenced genome, Chapters 2 and 3 demonstrate that PCR still remains a valuable tool for initial surveys of the ERV complement of a species, and that the data generated from such surveys is sufficient for an initial characterisation of the ERV present in these taxa (Gifford and Tristem, 2003). Furthermore, studies involving diversity of ERV domains across multiple individuals, such as within a population (Chapter 2) or across species (Herniou et al., 1998, Jaratlerdsiri et al., 2009, Martin et al., 1999, Martin et al., 1997, Martin et al., 2002), are best suited to PCR amplification-based methods, such as those targeting conserved domains (Chapter 2). Likewise, studies canvassing multiple tissue types may be better suited to PCR-based approaches in situations where re-sequencing projects are not feasible (Chapter 3).

PCR can also be used to study insertional polymorphisms across individuals or closely related taxa in situations where an ERV is not yet fixed in a population, or it is suspected that species specific expansion events have occurred (Belshaw et al., 2005a, Turner et al., 2001). Moreover, ERVs that are polymorphic across species may be a useful diagnostic tool for the resolution of uncertain phylogenies (Johnson and Coffin, 1999), particularly in the case of cryptic species such as the two sublineages of *C. niloticus*. These approaches usually require prior knowledge of ERV sequences and their integration sites, relying on amplification, or lack thereof depending on the presence of the insertion of interest. Although it was not incorporated into these current studies, the data generated from the crocodilian genomes (Chapters 6 and 7) will allow for targeted studies such as these to be carried out, particularly to assess the distribution of unusual or potentially significant ERV families across crocodilian species.

8.3: Final comments and future studies

Crocodilians and other non-mammalian vertebrates are a rich source of novel ERV diversity that has been largely untapped, despite the occasional study that has previously identified unique and divergent lineages of ERVs and other TEs. These studies include two exogenous piscine retroviruses (Holzschu et al., 1995, LaPierre et al., 1999), an amphibian ERV (Kambol et al., 2003) and a description of a small number of sequences from the ERV4 class of crocodilian ERVs (Martin et al., 2002). The studies presented here represent the first

comprehensive survey of ERVs in crocodilians, and include an assessment of ERV diversity at the species and population levels, as well as the characterisation of ERVs at the genomic scale and across tissues.

The recovery of novel ERV families, intermediate lineages, and a new Class of ERVs highlights the importance of studying non-mammalian vertebrates, and the unique evolutionary insights that these taxa can provide. The studies presented suggest the presence of recently or currently active ERV lineages in crocodilians that warrant further investigation. While no further conclusions can be drawn on the current activity of ERVs based on DNA sequence data, the development of transcriptome resources will facilitate detection and confirmation of ERV activity, and therefore will be important for further studies into the potential replicative ability of these ERVs.

The data presented provide a significant resource for targeted studies of ERV diversity in crocodilians and non-mammalian vertebrates, and the means by which functional analyses of ERVs in crocodilian genomes may be carried out. The correlation of ERV locations with gene regions and chromosomal breakpoints will be important to unravel the underlying interactions between these ERVs and the crocodilian genomes, and the impacts that ERV replication has imposed on crocodilian evolution. These correlations and identification of ERV transcripts from RNAseq data is essential for furthering investigations into the possibility of actively replicating ERVs within crocodilians. The possibility of exapted ERVs and ERV mediated gene duplications within crocodilians, and *C. porosus* in particular, should also be investigated further as these may play a key role in crocodilian development genome evolution.

References

Alfoldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., Russell, P., Lowe, C. B., Glor, R. E., Jaffe, J. D., Ray, D. A., Boissinot, S., Shedlock, A. M., Botka, C., Castoe, T. A., Colbourne, J. K., Fujita, M. K., Moreno, R. G., ten Hallers, B. F., Haussler, D., Heger, A., Heiman, D., Janes, D. E., Johnson, J., de Jong, P. J., Koriabine, M. Y., Lara, M., Novick, P. A., Organ, C. L., Peach, S. E., Poe, S., Pollock, D. D., de Queiroz, K., Sanger, T., Searle, S., Smith, J. D., Smith, Z., Swofford, R., Turner-Maier, J., Wade, J., Young, S., Zadissa, A., Edwards, S. V., Glenn, T. C., Schneider, C. J., Losos, J. B., Lander, E. S., Breen, M., Ponting, C. P. & Lindblad-Toh, K. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature,* 477**,** 587-591.

Alibardi, L. & Toni, M. 2007. Characterization of keratins and associated proteins involved in the corneification of crocodilian epidermis. *Tissue and Cell,* 39**,** 311-323.

Allen, G. R. 1974. The marine crocodile, *Crocodylus porosus*, from Ponape, eastern Caroline Islands, with notes on food habits of crocodiles from the Palau Archipelago. *Copeia,* 1974**,** 553-553.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

Andersen, P. R., Barbacid, M., Tronick, S. R., Clark, H. F. & Aaronson, S. A. 1979. Evolutionary relatedness of viper and primate endogenous retroviruses. *Science,* 204**,** 318-321.

Andersson, G., Svensson, A.-C., Setterblad, N. & Rask, L. 1998. Retroelements in the human MHC class II region. *Trends in Genetics,* 14**,** 109-114.

Anisimova, M. & Gascuel, O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology,* 55**,** 539-552.

Applied Biosystems. Available: http://www.appliedbiosystems.com [Accessed 22/06/2013].

Arnaud, F., Caporale, M., Varela, M., Biek, R., Chessa, B., Alberti, A., Golder, M., Mura, M., Zhang, Y. P., Yu, L., Pereira, F., DeMartini, J. C., Leymaster, K., Spencer, T. E. & Palmarini, M. 2007a. A paradigm for virus-host coevolution: Sequential counter-adaptations between endogenous and exogenous retroviruses. *PLoS Pathogens,* 3**,** 1716-1729.

Arnaud, F., Murcia, P. R. & Palmarini, M. 2007b. Mechanisms of late restriction induced by an endogenous retrovirus. *Journal of Virology,* 81**,** 11441-11451.

Arnaud, F., Varela, M., Spencer, T. E. & Palmarini, M. 2008. Coevolution of endogenous Betaretroviruses of sheep and their host. *Cellular and Molecular Life Sciences,* 65**,** 3422-3432.

Balakrishnan, C. N., Ekblom, R., Volker, M., Westerdahl, H., Godinez, R., Kotkiewicz, H., Burt, D. W., Graves, T., Griffin, D. K., Warren, W. C. & Edwards, S. V. 2010. Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology,* 8**,** 29.

Bannert, N. & Kurth, R. 2006. The evolutionary dynamics of human endogenous retroviral families. *Annual Review of Genomics and Human Genetics,* 7**,** 149-173.

Bao, Z. & Eddy, S. R. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research,* 12**,** 1269-1276.

Barbulescu, M., Turner, G., Seaman, M. I., Deinard, A. S., Kidd, K. K. & Lenz, J. 1999. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Current Biology,* 9**,** 861-S1.

Barr, S. D., Leipzig, J., Shinn, P., Ecker, J. R. & Bushman, F. D. 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *Journal of Virology,* 79**,** 12035-12044.

Barrio, A. M., Ekerljung, M., Jern, P., Benachenhou, F., Sperber, G. O., Bongcam-Rudloff, E., Blomberg, J. & Andersson, G. 2011. The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships. *PLoS One,* 6.

Beer, C., Meyer, A., Müller, K. & Wirth, M. 2003. The temperature stability of mouse retroviruses depends on the cholesterol levels of viral lipid shell and cellular plasma membrane. *Virology,* 308**,** 137-146.

Belshaw, R., Dawson, A. L. A., Woolven-Allen, J., Redding, J., Burt, A. & Tristem, M. 2005a. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *Journal of Virology,* 79**,** 12507-12514.

Belshaw, R., Katzourakis, A., Pačes, J., Burt, A. & Tristem, M. 2005b. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Molecular Biology and Evolution,* 22**,** 814-817.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Pačes, J., Burt, A. & Tristem, M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 4894-4899.

Benit, L., Dessen, P. & Heidmann, T. 2001. Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. *Journal of Virology,* 75**,** 11709-11719.

Benit, L., Lallemand, J. B., Casella, J. F., Philippe, H. & Heidmann, T. 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *Journal of Virology,* 73**,** 3301-3308.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R. & Willerslev, E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One,* 2**,** e197.

Bishop, J. M. 1978. Retroviruses. *Annual Review of Biochemistry,* 47**,** 35-88.

Blikstad, V., Benachenhou, F., Sperber, G. O. & Blomberg, J. 2008. Evolution of human endogenous retroviral sequences: a conceptual account. *Cellular and Molecular Life Sciences,* 65**,** 3348-3365.

Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G. & Mayer, J. 2009. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. *Gene,* 448**,** 115-123.

Boeke, J. D. & Stoye, J. P. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. *In:* Coffin, J. M., Hughes, S. H. & Varmus, H. E. (eds.) *Retroviruses.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Boni, M. F., Posada, D. & Feldman, M. W. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics,* 176**,** 1035-1047.

Borysenko, L., Stepanets, V. & Rynditch, A. V. 2008. Molecular characterization of full-length MLV-related endogenous retrovirus ChiRV1 from the chicken, *Gallus gallus*. *Virology,* 376**,** 199-204.

Boyce-Jacino, M. T., O'Donoghue, K. & Faras, A. J. 1992. Multiple complex families of endogenous retroviruses are highly conserved in the genus *Gallus*. *Journal of Virology,* 66**,** 4919-4929.

Brazaitis, P. 1987. The identification of crocodilian skins and products. *In:* Webb, G. J. W., Manolis, S. C. & Whitehead, P. (eds.) *Wildlife Management: Crocodiles and Alligators*. Chipping Norton, NSW: Surrey Beatty & Sons in association with the Conservation Commission of the Northern Territory.

Brazaitis, P. 1989. The forensic identification of crocodilian hides and products. *Crocodiles: Their ecology, management and conservation. IUCN, Gland, Switzerland.*

Brochu, C. A. 2003. Phylogenetic approaches toward crocodylian history. *Annual Review of Earth and Planetary Sciences,* 31**,** 357-397.

Buenviaje, G. N., Hirst, R. G., Ladds, P. W. & Millan, J. M. 1997. Isolation of *Dermatophilus sp* from skin lesions in farmed saltwater crocodiles (*Crocodylus porosus*). *Australian Veterinary Journal,* 75**,** 365-367.

Buenviaje, G. N., Ladds, P. W. & Martin, Y. 1998. Pathology of skin diseases in crocodiles. *Australian Veterinary Journal,* 76**,** 357-363.

Buenviaje, G. N., Ladds, P. W., Melville, L. & Manolis, S. C. 1994. Disease-husbandry associations in farmed crocodiles in Queensland and the Northern Territory. *Australian Veterinary Journal,* 71**,** 165-173.

Burmeister, T. 2001. Oncogenic retroviruses in animals and humans. *Reviews in Medical Virology,* 11**,** 369-380.

Bushman, F., Lewinski, M., Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S. & Hoffmann, C. 2005. Genome-wide analysis of retroviral DNA integration. *Nature Reviews Microbiology,* 3**,** 848-58.

Butler, M., Goodwin, T., Simpson, M., Singh, M. & Poulter, R. 2001. Vertebrate LTR retrotransposons of the *Tf1/Sushi* group. *Journal of Molecular Evolution,* 52**,** 260-274.

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

Caldwell, J. 2011. World trade in crocodilian skins 2007-2009. Cambridge: UNEP-WCMC.

Chandra, A. M. S., Jacobson, E. R. & Munn, R. J. 2001. Retroviral particles in neoplasms of Burmese pythons (*Python molurus bivittatus*). *Veterinary Pathology,* 38**,** 561-564.

Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*.

Cho, K., Lee, Y. K. & Greenhalgh, D. G. 2008. Endogenous retroviruses in systemic response to stress signals. *Shock,* 30**,** 105-16.

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.-C., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., Anderson, J. A., Gill, B. S., Appels, R., Keller, B. & Feuillet, C. 2010. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant Cell Online,* 22**,** 1686-1701.

Clark, H. F., Andersen, P. R. & Lunger, P. D. 1979. Propagation and characterization of a C-Type virus from a rhabdomyosarcoma of a corn snake. *Journal of General Virology,* 43**,** 673-683.

Clough, J. E., Foster, J. A., Barnett, M. & Wichman, H. A. 1996. Computer simulation of transposable element evolution: random template and strict master models. *Journal of Molecular Evolution,* 42**,** 52-58.

Coffin, J. M. 1992. Structure and classification of retroviruses. *In:* Levy, J. A. (ed.) *The Retroviridae.* New York, NY: Plenum Press.

Coffin, J. M., Hughes, S. H. & Varmus, H. E. 1997. The interactions of retroviruses and their hosts. *In:* Coffin, J. M., Hughes, S. H. & Varmus, H. E. (eds.) *Retroviruses.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Cogger, H. E. 1992. *Reptiles & Amphibians of Australia,* Ithaca, NY, Cornell University Press.
Cordaux, R., Hedges, D. J. & Batzer, M. A. 2004. Retrotransposition of *Alu* elements: How many sources? *Trends in Genetics,* 20**,** 464-467.
Cornelis, G., Heidmann, O., Bernard-Stoecklin, S., Reynaud, K., Véron, G., Mulot, B., Dupressoir, A. & Heidmann, T. 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proceedings of the National Academy of Sciences,* 109**,** E432–E441.
Cornelis, G., Heidmann, O., Degrelle, S. A., Vernochet, C., Lavialle, C., Letzelter, C., Bernard-Stoecklin, S., Hassanin, A., Mulot, B., Guillomot, M., Hue, I., Heidmann, T. & Dupressoir, A. 2013. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proceedings of the National Academy of Sciences,* 110**,** E828–E837.
Craig, N. L. 2002. Mobile DNA: An introduction. *In:* Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds.) *Mobile DNA II.* Washington, DC: ASM Press.
Craigie, R. 2002. Retroviral DNA integration. *In:* Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. (eds.) *Mobile DNA II.* Washington, DC: ASM Press.
Craven, R. C., Leure-duPree, A. E., Weldon Jr, R. A. & Wills, J. W. 1995. Genetic analysis of the major homology region of the Rous sarcoma virus Gag protein. *Journal of Virology,* 69**,** 4213-4227.
Crittenden, L. B., Smith, E. J., Weiss, R. A. & Sarma, P. S. 1974. Host gene control of endogenous avian leukosis virus production. *Virology,* 57**,** 128-138.
Crocodile Specialist Group 1996a. *Alligator sinensis*. *IUCN 2011, IUCN Red List of Threatened Species, http://www.iucnredlist.org, Version 2011.2, Downloaded 30 May 2012*.
Crocodile Specialist Group 1996b. *Crocodylus johnsoni*. *IUCN 2012, IUCN Red List of Threatened Species, http://www.iucnredlist.org, Version 2012.2, Downloaded 30 May 2012*.
Crocodile Specialist Group 1996c. *Crocodylus porosus*. *IUCN 2011, IUCN Red List of Threatened Species, http://www.iucnredlist.org, Version 2011.2, Downloaded 30 May 2012*.
Cullen, B. R. 2006. Viruses and microRNAs. *Nature Genetics*.
D'Souza, V. & Summers, M. F. 2005. How retroviruses select their genomes. *Nature Reviews Microbiology,* 3**,** 643-655.
Dawkins, R., Leelayuwat, C., Gaudieri, S., Tay, G., Hui, J., Cattley, S., Martinez, P. & Kulski, J. 1999. Genomics of the major histocompatibility complex: Haplotypes, duplication, retroviruses and disease. *Immunological Reviews,* 167**,** 275-304.
DEC 2009. Management Plan for the Commercial Harvest and Farming of Crocodiles in Western Australia: 1 January 2009 – 31 December 2013. Perth: Department of Environment and Conservation, Government of Western Australia.
Deininger, P. L., Batzer, M. A., Hutchison, C. A. & Edgell, M. H. 1992. Master genes in mammalian repetitive DNA amplification. *Trends in Genetics,* 8**,** 307-311.
Denner, J. 1998. Immunosuppression by Retroviruses: Implications for Xenotransplantation. *Annals of the New York Academy of Sciences,* 862**,** 75-86.
Densmore, L. D. & Owen, R. D. 1989. Molecular systematics of the Order Crocodilia. *American Zoologist,* 29**,** 831-841.

Densmore, L. D. & White, P. S. 1991. The systematics and evolution of the Crocodilia as suggested by restriction endonuclease analysis of mitochondrial and nuclear ribosomal DNA. *Copeia*, 602-615.
Doody, J. S., Green, B., Rhind, D., Castellano, C. M., Sims, R. & Robinson, T. 2009. Population-level declines in Australian predators caused by an invasive species. *Animal Conservation,* 12, 46-53.
Doxiadis, G. G. M., de Groot, N. & Bontrop, R. E. 2008. Impact of endogenous intronic retroviruses on major histocompatibility complex Class II diversity and stability. *Journal of Virology,* 82, 6667-6677.
Eaton, M. J., Martin, A., Thorbjarnarson, J. & Amato, G. 2009. Species-level diversification of African dwarf crocodiles (Genus *Osteolaemus*): A geographic and phylogenetic perspective. *Molecular Phylogenetics and Evolution,* 50, 496-506.
Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32, 1792-1797.
Edwards, S. V., Gasper, J., Garrigan, D., Martindale, D. & Koop, B. F. 2000. A 39-kb sequence around a blackbird MHC class II gene: Ghost of selection past and songbird genome architecture. *Molecular Biology and Evolution,* 17, 1384-1395.
Elsey, R. M., Joanen, T., McNease, L. & Lance, V. 1990. Growth rate and plasma corticosterone levels in juvenile alligators maintained at different stocking densities. *Journal of Experimental Zoology,* 255, 30-36.
Eo, S. H. & DeWoody, J. A. 2010. Evolutionary rates of mitochondrial genomes correspond to diversification rates and to contemporary species richness in birds and reptiles. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 277, 3587-3592.
Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A. (eds.) 2005. *Virus taxonomy : Classification and nomenclature of viruses : Eighth report of the International Committee on the Taxonomy of Viruses* San Diego, CA ; London: Elsevier Academic Press.
Fitzsimmons, N. N., Buchan, J. C., Lam, P. V., Polet, G., Hung, T. T., Thang, N. Q. & Gratten, J. 2002. Identification of purebred *Crocodylus siamensis* for reintroduction in Vietnam. *Journal of Experimental Zoology,* 294, 373-381.
Foggin, C. M. 1992. Diseases of farmed crocodiles. *In:* Smith, G. A. & Marais, J. (eds.) *Conservation and utilization of the Nile crocodile in South Africa. Handbook on crocodile farming.* Pretoria, South Africa: The Crocodile Study Group of Southern Africa.
Foster, M. 2009. Emerging Animal and Plant Industries - their value to Australia (Second Edition). Second Edition ed. Canberra: Rural Industries Research and Development Corporation, Union Offset Printing.
Fukuda, Y., Whitehead, P. & Boggs, G. 2007. Broad-scale environmental influences on the abundance of saltwater crocodiles (*Crocodylus porosus*) in Australia. *Wildlife Research,* 34, 167.
Garcia-Etxebarria, K. & Jugo, B. M. 2010. Genome-wide detection and characterization of endogenous retroviruses in *Bos taurus*. *Journal of Virology,* 84, 10852-62.
Gasper, J. S., Shiina, T., Inoko, H. & Edwards, S. V. 2001. Songbird genomics: Analysis of 45 kb upstream of a polymorphic MHC Class II gene in red-winged blackbirds (*Agelaius phoeniceus*). *Genomics,* 75, 26-34.
Gatesy, J., Amato, G., Norell, M., DeSalle, R. & Hayashi, C. 2003. Combined support for wholesale taxic atavism in gavialine crocodylians. *Systematic Biology,* 52, 403-422.

Gatesy, J., Baker, R. H. & Hayashi, C. 2004. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. *Systematic Biology,* 53**,** 342-355.

Gherman, A., Chen, P. E., Teslovich, T. M., Stankiewicz, P., Withers, M., Kashuk, C. S., Chakravarti, A., Lupski, J. R., Cutler, D. J. & Katsanis, N. 2007. Population bottlenecks as a potential major shaping force of human genome architecture. *PLoS Genetics,* 3**,** e119.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics,* 16**,** 573-82.

Gifford, R., Kabat, P., Martin, J., Lynch, C. & Tristem, M. 2005. Evolution and distribution of Class II-related endogenous retroviruses. *Journal of Virology,* 79**,** 6478-6486.

Gifford, R. & Tristem, M. 2003. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes,* 26**,** 291-316.

Gifford, R. J., Katzourakis, A., Tristem, M., Pybus, O. G., Winters, M. & Shafer, R. W. 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 20362-20367.

Gilbert, C., Maxfield, D. G., Goodman, S. M. & Feschotte, C. 2009. Parallel germline infiltration of a lentivirus in two Malagasy lemurs. *PLoS Genetics,* 5**,** e1000425.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources,* 11**,** 759-769.

Govett, P. D., Harms, C. A., Johnson, A. J., Latimer, K. S., Wellehan, J. F. X., Fatzinger, M. H., Christian, L. S., Kelly, T. R. & Lewbart, G. A. 2005. Lymphoid follicular cloacal inflammation associated with a novel Herpesvirus in juvenile alligators (*Alligator mississippiensis*). *Journal of Veterinary Diagnostic Investigation,* 17**,** 474-479.

Gratten, J. 2003. *The molecular systematics, phylogeography and population genetics of Indo-Pacific Crocodylus.* PhD, University of Queensland.

Grigg, G. & Gans, C. 1993. Morphology and physiology of the Crocodylia. *In:* Glasby, C. G., Ross, G. J. B. & Beesley, P. L. (eds.) *Fauna of Australia, Volume 2a: Amphibia and Reptilia.* Canberra, ACT: AGPS.

Grigg, G., Taplin, L., Harlow, P. & Wright, J. 1980. Survival and growth of hatchling *Crocodylus porosus* in saltwater without access to fresh drinking water. *Oecologia,* 47**,** 264-266.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010. New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology,* 59**,** 307-321.

Hall, T. A. Year. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *In:* Nucleic Acids Symposium Series, 1999. 95-98.

Hanger, J. J., Bromham, L. D., McKee, J. J., O'Brien, T. M. & Robinson, W. F. 2000. The nucleotide sequence of Koala (*Phascolarctos cinereus*) Retrovirus: a novel Type C endogenous virus related to Gibbon Ape Leukemia Virus. *Journal of Virology,* 74**,** 4264-4272.

Harshman, J., Huddleston, C. J., Bollback, J. P., Parsons, T. J. & Braun, M. J. 2003. True and false gharials: A nuclear gene phylogeny of Crocodylia. *Systematic Biology,* 52**,** 386-402.

Hart, D., Frerichs, G. N., Rambaut, A. & Onions, D. E. 1996. Complete nucleotide sequence and transcriptional analysis of the snakehead fish retrovirus. *Journal of Virology,* 70**,** 3606-3616.

Hass, C. A., Hoffman, M. A., Densmore, L. D., III & Maxson, L. R. 1992. Crocodilian evolution: Insights from immunological data. *Molecular Phylogenetics and Evolution,* 1**,** 193-201.
He, K., Ye, Q., Zhu, Y., Chen, H., Wan, Q.-H. & Fang, S.-G. 2012. A bacterial artificial chromosome library for the Chinese alligator (Alligator sinensis). *Gene,* 507**,** 74-78.
Heidmann, O., Vernochet, C., Dupressoir, A. & Heidmann, T. 2009. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. *Retrovirology,* 6**,** 107.
Hekkala, E., Shirley, M. H., Amato, G., Austin, J. D., Charter, S., Thorbjarnarson, J., Vliet, K. A., Houck, M. L., Desalle, R. O. B. & Blum, M. J. 2011. An ancient icon reveals new mysteries: mummy DNA resurrects a cryptic species within the Nile crocodile. *Molecular Ecology,* 20**,** 4199-4215.
Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., Kapitonov, V., Ovcharenko, I., Putnam, N. H., Shu, S., Taher, L., Blitz, I. L., Blumberg, B., Dichmann, D. S., Dubchak, I., Amaya, E., Detter, J. C., Fletcher, R., Gerhard, D. S., Goodstein, D., Graves, T., Grigoriev, I. V., Grimwood, J., Kawashima, T., Lindquist, E., Lucas, S. M., Mead, P. E., Mitros, T., Ogino, H., Ohta, Y., Poliakov, A. V., Pollet, N., Robert, J., Salamov, A., Sater, A. K., Schmutz, J., Terry, A., Vize, P. D., Warren, W. C., Wells, D., Wills, A., Wilson, R. K., Zimmerman, L. B., Zorn, A. M., Grainger, R., Grammer, T., Khokha, M. K., Richardson, P. M. & Rokhsar, D. S. 2010. The genome of the western clawed frog *Xenopus tropicalis*. *Science,* 328**,** 633-636.
Hematti, P., Hong, B.-K., Ferguson, C., Adler, R., Hanawa, H., Sellers, S., Holt, I. E., Eckfeldt, C. E., Sharma, Y. & Schmidt, M. 2004. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biology,* 2**,** e423.
Herniou, E., Martin, J., Miller, K., Cook, J., Wilkinson, M. & Tristem, M. 1998. Retroviral diversity and distribution in vertebrates. *Journal of Virology,* 72**,** 5955-5966.
Higashikawa, F. & Chang, L.-J. 2001. Kinetic Analyses of Stability of Simple and Complex Retroviral Vectors. *Virology,* 280**,** 124-131.
Holzschu, D. L., Martineau, D., Fodor, S. K., Vogt, V. M., Bowser, P. R. & Casey, J. W. 1995. Nucleotide sequence and protein analysis of a complex piscine retrovirus, Walleye Dermal Sarcoma Virus. *Journal of Virology,* 69**,** 5320-5331.
Huang, E. J., Nocka, K. H., Buck, J. & Besmer, P. 1992. Differential expression and processing of two cell associated forms of the kit-ligand: KL-1 and KL-2. *Molecular Biology of the Cell,* 3**,** 349-62.
Huang, W., Umbach, D. M. & Li, L. 2006. Accurate anchoring alignment of divergent sequences. *Bioinformatics,* 22**,** 29-34.
Huchzermeyer, F. W. 2003. *Crocodiles: biology, husbandry and diseases*, CABI Pub.
Huda, A., Polavarapu, N., Jordan, I. K. & McDonald, J. F. 2008. Endogenous retroviruses of the chicken genome. *Biology Direct,* 3.
Huder, J. B., Boni, J., Hatt, J. M., Soldati, G., Lutz, H. & Schupbach, J. 2002. Identification and characterization of two closely related unclassifiable endogenous retroviruses in pythons (*Python molurus* and *Python curtus*). *Journal of Virology,* 76**,** 7607-7615.
Hugall, A. F., Foster, R. & Lee, M. S. Y. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Systematic Biology,* 56**,** 543-563.
Hughes, J. F. & Coffin, J. M. 2001. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nature Genetics,* 29**,** 487-489.

Hughes, J. F., Skaletsky, H., Pyntikova, T., Minx, P. J., Graves, T., Rozen, S., Wilson, R. K. & Page, D. C. 2005. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature,* 437**,** 100-103.

Hughes, S. & Mouchiroud, D. 2001. High evolutionary rates in nuclear genes of squamates. *Journal of Molecular Evolution,* 53**,** 70-76.

Hultman, K. A., Bahary, N., Zon, L. I. & Johnson, S. L. 2007. Gene Duplication of the Zebrafish *kit ligand* and Partitioning of Melanocyte Development Functions to *kit ligand a*. *PLoS Genetics,* 3**,** e17.

Illumina. Available: http://www.illumina.com [Accessed 22/06/2013].

International Chicken Genome Sequencing Consortium 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature,* 432**,** 695-716.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Isberg, S., Shilton, C. & Thomson, P. 2009. Improving Australia's crocodile industry productivity: Understanding runtism and survival. Canberra: Rural Industries Research and Development Corporation Union Offset Printing

Isberg, S. R., Chen, Y., Barker, S. G. & Moran, C. 2004a. Analysis of microsatellites and parentage testing in saltwater crocodiles. *Journal of Heredity,* 95**,** 445-449.

Isberg, S. R., Thomson, P. C., Nicholas, F. W., Barker, S. G. & Moran, C. 2004b. A genetic improvement program for farmed saltwater crocodiles. Canberra: Rural Industries Research and Development Corporation Union Offset Printing

Isberg, S. R., Thomson, P. C., Nicholas, F. W., Barker, S. G. & Moran, C. 2005a. Quantitative analysis of production traits in saltwater crocodiles (*Crocodylus porosus*): I. reproduction traits. *Journal of Animal Breeding and Genetics,* 122**,** 361-369.

Isberg, S. R., Thomson, P. C., Nicholas, F. W., Barker, S. G. & Moran, C. 2005b. Quantitative analysis of production traits in saltwater crocodiles (*Crocodylus porosus*): II. age at slaughter. *Journal of Animal Breeding and Genetics,* 122**,** 370-377.

Isberg, S. R., Thomson, P. C., Nicholas, F. W., Barker, S. G. & Moran, C. 2006a. Quantitative analysis of production traits in saltwater crocodiles (*Crocodylus porosus*): III. juvenile survival. *Journal of Animal Breeding and Genetics,* 123**,** 44-47.

Isberg, S. R., Thomson, P. C., Nicholas, F. W., Webb, G. J. W., Manolis, S. C., Barker, S. G. & Moran, C. 2006b. Quantitative analysis of production traits in saltwater crocodiles (*Crocodylus porosus*): IV. number of scale rows. *Journal of Animal Breeding and Genetics,* 123**,** 48-55.

Iyer, S. R., Yu, D., Biancotto, A., Margolis, L. B. & Wu, Y. 2009. Measurement of human immunodeficiency virus type 1 preintegration transcription by using Rev-dependent Rev-CEM cells reveals a sizable transcribing DNA population comparable to that from proviral templates. *Journal of Virology,* 83**,** 8662-73.

Jacobson, E. R., Oros, J., Tucker, S. J., Pollock, D. P., Kelley, K. L., Munn, R. J., Lock, B. A., Mergia, A. & Yamamoto, J. K. 2001. Partial characterization of retroviruses from boid snakes with inclusion body disease. *American Journal of Veterinary Research,* 62**,** 217-224.

Janke, A. & Arnason, U. 1997. The complete mitochondrial genome of *Alligator mississippiensis* and the separation between recent archosauria (birds and crocodiles). *Molecular Biology and Evolution,* 14**,** 1266-1272.

Janke, A., Gullberg, A., Hughes, S., Aggarwal, R. & Arnason, U. 2005. Mitogenomic analyses place the gharial (*Gavialis gangeticus*) on the crocodile tree and provide pre-

K/T divergence times for most crocodilians. *Journal of Molecular Evolution,* 61**,** 620-626.
Jaratlerdsiri, W., Isberg, S., Higgins, D. & Gongora, J. 2012. MHC class I of saltwater crocodiles (*Crocodylus porosus*): polymorphism and balancing selection. *Immunogenetics,* 64**,** 825-838.
Jaratlerdsiri, W., Rodríguez-Zárate, C. J., Isberg, S. R., Damayanti, C. S., Miles, L. G., Chansue, N., Moran, C., Melville, L. & Gongora, J. 2009. Distribution of Endogenous Retroviruses in Crocodilians. *Journal of Virology,* 83**,** 10305-10308.
Jern, P. & Coffin, J. M. 2008. Effects of retroviruses on host genome function. *Annual Review of Genetics,* 42**,** 709-732.
Jern, P., Sperber, G. O. & Blomberg, J. 2004. Definition and variation of human endogenous retrovirus H. *Virology,* 327**,** 93-110.
Jern, P., Sperber, G. O. & Blomberg, J. 2005. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology,* 2**,** 50.
Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. & Madden, T. L. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Research,* 36**,** W5-W9.
Johnson, W. E. & Coffin, J. M. 1999. Constructing primate phylogenies from ancient retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 10254-10260.
Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research,* 110**,** 462-7.
Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. 1996. Censor—a program for identification and elimination of repetitive elements from DNA sequences. *Computers and Chemistry,* 20**,** 119-121.
Kambhu, S., Falldorf, P. & Lee, J. S. 1990. Endogenous retroviral long terminal repeats within the HLA-DQ locus. *Proceedings of the National Academy of Sciences of the United States of America,* 87**,** 4927-4931.
Kambol, R., Kabat, P. & Tristem, M. 2003. Complete nucleotide sequence of an endogenous retrovirus from the amphibian, *Xenopus laevis*. *Virology,* 311**,** 1-6.
Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research,* 33**,** 511-518.
Katz, R. A. & Skalka, A. M. 1990. Generation of diversity in retroviruses. *Annual Review of Genetics,* 24**,** 409-445.
Katz, R. A. & Skalka, A. M. 1994. The retroviral enzymes. *Annual Review of Biochemistry,* 63**,** 133-173.
Katzourakis, A. & Gifford, R. J. 2010. Endogenous viral elements in animal genomes. *PLoS Genetics,* 6.
Katzourakis, A., Rambaut, A. & Pybus, O. G. 2005. The evolutionary dynamics of endogenous retroviruses. *Trends in Microbiology,* 13**,** 463-468.
Katzourakis, A., Tristem, M., Pybus, O. G. & Gifford, R. J. 2007. Discovery and analysis of the first endogenous lentivirus. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 6261-6265.
Kazazian, H. H. 2004. Mobile elements: Drivers of genome evolution. *Science,* 303**,** 1626-1632.
Keane, T., Creevey, C., Pentony, M., Naughton, T. & Mclnerney, J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that *ad hoc* assumptions for choice of matrix are not justified. *BMC Evolutionary Biology,* 6**,** 29.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome Research,* 12**,** 656-64.
Kim, H.-S., Takenaka, O. & Crow, T. J. 1999. Isolation and phylogeny of endogenous retrovirus sequences belonging to the HERV-W family in primates. *Journal of General Virology,* 80**,** 2613-2619.
King, M., Honeycutt, R. & Contreras, N. 1986. Chromosomal repatterning in crocodiles: C, G and N-banding and the in situ hybridization of 18S and 26S rRNA cistrons. *Genetica,* 70**,** 191-201.
Klymiuk, N., Muller, M., Brem, G. & Aigner, B. 2003. Characterization of endogenous retroviruses in sheep. *Journal of Virology,* 77**,** 11268-11273.
Kohany, O., Gentles, A., Hankus, L. & Jurka, J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics,* 7**,** 474.
Kulski, J. K., Gaudieri, S., Inoko, H. & Dawkins, R. L. 1999. Comparison between two human endogenous retrovirus (HERV)-rich regions within the major histocompatibility complex. *Journal of Molecular Evolution,* 48**,** 675-683.
Ladds, P. W., Bradley, J. & Hirst, R. G. 1996. *Providencia rettgeri* meningitis in hatchling saltwater crocodiles (*Crocodylus porosus*). *Australian Veterinary Journal,* 24**,** 397-398.
Ladds, P. W. & Sims, L. D. 1990. Diseases of young captive crocodiles in Papua New Guinea. *Australian Veterinary Journal,* 67**,** 323-330.
LaPierre, L. A., Holzschu, D. L., Bowser, P. R. & Casey, J. W. 1999. Sequence and transcriptional analyses of the fish retroviruses walleye epidermal hyperplasia virus types 1 and 2: evidence for a gene duplication. *Journal of Virology,* 73**,** 9393-9403.
Leach, G. J., Delaney, R. & Fukuda, Y. 2009. Management program for the saltwater crocodile in the Northern Territory of Australia, 2009 - 2014. Darwin: Northern Territory Department of Natural Resources, Environment, The Arts and Sport.
Letnic, M., Webb, J. K. & Shine, R. 2008. Invasive cane toads (*Bufo marinus*) cause mass mortality of freshwater crocodiles (*Crocodylus johnstoni*) in tropical Australia. *Biological Conservation,* 141**,** 1773-1782.
Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G. K.-S. & Wang, J. 2005. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol,* 1**,** e43.
Li, W. & Godzik, A. 2006. CD-HIT: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-1659.
Li, Y., Wu, X., Ji, X., Yan, P. & Amato, G. 2007. The complete mitochondrial genome of salt-water crocodile (*Crocodylus porosus*) and phylogeny of crocodilians. *Journal of Genetics and Genomics,* 34**,** 119-128.
Lloyd, M. & Morris, P. J. 1999. Phlebotomy techniques in crocodilians. *Bulletin of the Association of Reptilian and Amphibian Veterinarians,* 9**,** 12-14.
Lower, R., Lower, J. & Kurth, R. 1996. The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 5177-5184.
Luck, N. L., Thomas, K. C., Morin-Adeline, V. E., Barwick, S., Chong, A. Y., Carpenter, E. L., Wan, L., Willet, C. E., Langford-Salisbury, S. M., Abdelsayd, M., Ang, R. A., Atkinson, S. J., Barcelo, F. G., Booth, M. E., Bradbury, E. J., Branighan, T. L., Brown, J., Castillo, L. E., Chandler, N. D., Chong, J. Y., Collits, K. J., Cook, E., Cruz, R. E., Farrugia, C. A., Fletcher, J. L., Fletcher, S., Gamaliel, N. S., Gurr, J. F., Hallett, N. J., Hargreaves, G., Harris, T., Hollings, S., Hopcroft, R. L., Johinke, D., Kern, P. L., Kiddell, J. L., Kilby, K. E., Kragic, B., Kwan, J. H., Lee, J. I., Liang, J. M., Lillie, M. C., Lui, B. C., Luk, S. W., Lun, K. H., Marshall, K. L., Marzec, J. A., Masters, K.

T., Mazurkijevic, L. J., Medlock, J., Meoli, C., Morris, K. M., Noh, Y. H., Okazaki, H., Orourke, T. J., Payne, E. M., Powell, D. J., Quinlivan, A. R., Reeves, T. J., Robson, K., Robson, K. L., Royle, L. J., Stevenson, R., Sellens, T., Sun, Z., Sutton, A. L., Swan, A., Tang, J. M., Tinker, J. E., Tomlinson, S. C., Wilkin, T., Wright, A. L., Xiao, S. T., Yang, J., Yee, C., Jaratlerdsiri, W., Isberg, S. R., Miles, L., Higgins, D., Lane, A. & Gongora, J. 2012. Mitochondrial DNA analyses of the saltwater crocodile (*Crocodylus porosus*) from the Northern Territory of Australia. *Australian Journal of Zoology,* 60**,** 18-25.

Lynch, M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. *Molecular Biology and Evolution,* 14**,** 914-925.

Macas, J., Neumann, P. & Navratilova, A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics,* 8**,** 427.

Magbanua, Z. V., Ozkan, S., Bartlett, B. D., Chouvarine, P., Saski, C. A., Liston, A., Cronn, R. C., Nelson, C. D. & Peterson, D. G. 2011. Adventures in the enormous: A 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. *PLoS One,* 6**,** e16214.

Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J. & Belshaw, R. 2012. *Env*-less endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 7385-7390.

Maksakova, I. A., Mager, D. L. & Reiss, D. 2008. Keeping active endogenous retroviral-like elements in check: The epigenetic perspective. *Cellular and Molecular Life Sciences,* 65**,** 3329-3347.

Man, Z., Yishu, W., Peng, Y. & Xiaobing, W. 2011. Crocodilian phylogeny inferred from twelve mitochondrial protein-coding genes, with new complete mitochondrial genomic sequences for *Crocodylus acutus* and *Crocodylus novaeguineae*. *Molecular Phylogenetics and Evolution,* 60**,** 62-67.

Martin, D. & Rybicki, E. 2000. RDP: Detection of recombination amongst aligned sequences. *Bioinformatics,* 16**,** 562-563.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefeuvre, P. 2010. RDP3: A flexible and fast computer program for analyzing recombination. *Bioinformatics,* 26**,** 2462-3.

Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research and Human Retroviruses,* 21**,** 98-102.

Martin, J., Herniou, E., Cook, J., O'Neill, R. W. & Tristem, M. 1999. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. *Journal of Virology,* 73**,** 2442-2449.

Martin, J., Herniou, E., Cook, J., Oneill, R. W. & Tristem, M. 1997. Human endogenous retrovirus type I-related viruses have an apparently widespread distribution within vertebrates. *Journal of Virology,* 71**,** 437-443.

Martin, J., Kabat, P., Herniou, E. & Tristem, M. 2002. Characterization and complete nucleotide sequence of an unusual reptilian retrovirus recovered from the Order Crocodylia. *Journal of Virology,* 76**,** 4651-4654.

Maxfield, L. F., Fraize, C. D. & Coffin, J. M. 2005. Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 1436-41.

McAliley, L. R., Willis, R. E., Ray, D. A., White, P. S., Brochu, C. A. & Densmore, L. D. 2006. Are crocodiles really monophyletic? Evidence for subdivisions from sequence and morphological data. *Molecular Phylogenetics and Evolution,* 39**,** 16-32.

McCarthy, E. & McDonald, J. 2004. Long terminal repeat retrotransposons of *Mus musculus*. *Genome Biology,* 5**,** R14.
McCarthy, E. M. & McDonald, J. F. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics,* 19**,** 362-367.
McClure, M. A., Johnson, M. S., Feng, D. F. & Doolittle, R. F. 1988. Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. *Proceedings of the National Academy of Sciences of the United States of America,* 85**,** 2469-2473.
Melville, L., Davis, S., Shilton, C., Isberg, S., Chong, A. & Gongora, J. 2012. Hunting viruses in crocodiles: viral and endogenous retroviral detection and characterisation in farmed crocodiles. Canberra: Rural Industries Research and Development Corporation Union Offset Printing
Meredith, R. W., Hekkala, E. R., Amato, G. & Gatesy, J. 2011. A phylogenetic hypothesis for *Crocodylus* (Crocodylia) based on mitochondrial DNA: evidence for a trans-Atlantic voyage from Africa to the New World. *Molecular Phylogenetics and Evolution,* 60**,** 183-191.
Metzker, M. L. 2009. Sequencing technologies—the next generation. *Nature Reviews Genetics,* 11**,** 31-46.
Mietz, J. A., Grossman, Z., Lueders, K. K. & Kuff, E. L. 1987. Nucleotide sequence of a complete mouse intracisternal A-particle genome: relationship to known aspects of particle assembly and function. *Journal of Virology,* 61**,** 3020-3029.
Miles, L., Isberg, S., Glenn, T., Lance, S., Dalzell, P., Thomson, P. & Moran, C. 2009. A genetic linkage map for the saltwater crocodile (*Crocodylus porosus*). *BMC Genomics,* 10**,** 339.
Miles, L. G., Isberg, S. R., Thomson, P. C., Glenn, T. C., Lance, S. L., Dalzell, P. & Moran, C. 2010. QTL mapping for two commercial traits in farmed saltwater crocodiles (*Crocodylus porosus*). *Animal Genetics,* 41**,** 142-149.
Milián-García, Y., Venegas-Anaya, M., Frias-Soler, R., Crawford, A. J., Ramos-Targarona, R., Rodríguez-Soberón, R., Alonso-Tabet, M., Thorbjarnarson, J., Sanjur, O. I., Espinosa-López, G. & Bermingham, E. 2011. Evolutionary history of Cuban crocodiles *Crocodylus rhombifer* and *Crocodylus acutus* inferred from multilocus markers. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology,* 315A**,** 358-375.
Miller, D. L., Mauel, M. J., Baldwin, C., Burtle, G., Ingram, D. & Hines, M. E. 2003. West Nile virus in farmed alligators. *Emerging Infectious Diseases,* 9**,** 794.
Molnar, R. E. 1993. Biogeography and phylogeny of the Crocodylia. *In:* Glasby, C. G., Ross, G. J. B. & Beesley, P. L. (eds.) *Fauna of Australia, Volume 2a: Amphibia and Reptilia.* Canberra, ACT: AGPS.
Moore, B. E. 1993. Survival of human immunodeficiency virus (HIV), HIV-infected lymphocytes, and poliovirus in water. *Applied and Environmental Microbiology,* 59**,** 1437-1443.
Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. 1999. Exon Shuffling by L1 retrotransposition. *Science,* 283**,** 1530-1534.
Morici, L. A., Elsey, R. M. & Lance, V. A. 1997. Effects of long-term corticosterone implants on growth and immune function in juvenile alligators, *Alligator mississippiensis*. *Journal of Experimental Zoology,* 279**,** 156-162.
Morozova, O. & Marra, M. A. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics,* 92**,** 255-264.
Morpurgo, B., Gvaryahu, G. & Robinzon, B. 1992. Effects of population density, size, and gender on plasma testosterone, thyroxine, hematocrit, and calcium in juvenile Nile crocodiles (*Crocodylus niloticus*) in captivity. *Copeia,* 1992**,** 1023-1027.

Muller, H.-P. & Varmus, H. E. 1994. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *The EMBO Journal,* 13**,** 4707-4714.
Muriaux, D. & Rein, A. 2003. Encapsidation and transduction of cellular genes by retroviruses. *Frontiers in Bioscience,* 8**,** D135-142.
Nakaya, Y., Koshi, K., Nakagawa, S., Hashizume, K. & Miyazawa, T. 2013. Fematrin-1 Is Involved in Fetomaternal Cell-to-Cell Fusion in Bovinae Placenta and Has Contributed to Diversity of Ruminant Placentation. *Journal of virology,* 87**,** 10563-10572.
Nascimento, F., Gongora, J., Charleston, M., Tristem, M., Lowden, S. & Moran, C. 2011. Evolution of endogenous retroviruses in the *Suidae*: evidence for different viral subpopulations in African and Eurasian host species. *BMC Evolutionary Biology,* 11**,** 139.
Nei, M. & Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution,* 3**,** 418-26.
Norell, M. A. 1989. The higher level relationships of the extant Crocodylia. *Journal of Herpetology,* 23**,** 325-335.
Oaks, J. R. 2011. A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. *Evolution,* 65**,** 3285-3297.
Ohno, H., Aguilar, R. C., Fournier, M.-C., Hennecke, S., Cosson, P. & Bonifacino, J. S. 1997. Interaction of endocytic signals from the HIV-1 envelope glycoprotein complex with members of the adaptor medium chain family. *Virology,* 238**,** 305-315.
Origgi, F. 2007. Reptile immunology. *In:* Jacobson, E. R. (ed.) *Infectious diseases and pathology of reptiles. Color atlas and text.* Boca Raton: CRC Press.
Pacific Biosciences. Available: http://www.pacificbiosciences.com/ [Accessed 22/06/2013].
Padidam, M., Sawyer, S. & Fauquet, C. M. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology,* 265**,** 218-25.
Perl, A. 2003. Role of endogenous retroviruses in autoimmune diseases. *Rheumatic Disease Clinics of North America,* 29**,** 123-143.
Peterson, D. G., Schulze, S. R., Sciara, E. B., Lee, S. A., Bowers, J. E., Nagel, A., Jiang, N., Tibbitts, D. C., Wessler, S. R. & Paterson, A. H. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Research,* 12**,** 795-807.
Peucker, S. & Mayer, R. 1995. Runt observation studies. *Crocodile Research Bulletin,* 1**,** 57-62.
Poe, S. 1996. Data set incongruence and the phylogeny of crocodilians. *Systematic Biology,* 45**,** 393-414.
Polavarapu, N., Bowen, N. & McDonald, J. 2006. Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses. *Genome Biology,* 7**,** R51.
Posada, D. & Crandall, K. A. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 13757-13762.
Price, A. L., Jones, N. C. & Pevzner, P. A. 2005. De novo identification of repeat families in large genomes. *Bioinformatics,* 21**,** i351-i358.
QEPA 2007. Nature Conservation (estuarine crocodile) Conservation Plan 2007 and Management Program 2007–2017. Brisbane: Queensland Government - Environmental Protection Agency.

Rabson, A. B. & Graves, B. J. 1997. Synthesis and processing of viral RNA. *In:* Coffin, J. M., Hughes, S. H. & Varmus, H. E. (eds.) *Retroviruses.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E. J. P. 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS One,* 6**,** e22594.
Ray, D. A., Dever, J. A., Platt, S. G., Rainwater, T. R., Finger, A. G., McMurry, S. T., Batzer, M. A., Barr, B., Stafford, P. J., McKnight, J. & Densmore, L. D. 2004. Low levels of nucleotide diversity in *Crocodylus moreletii* and evidence of hybridization with *C. acutus*. *Conservation Genetics,* 5**,** 449-462.
Revol, B. 1995. Crocodile farming and conservation, the example of Zimbabwe. *Biodiversity and Conservation,* 4**,** 299-305.
Roche. Available: http://www.454.com [Accessed 22/06/2013].
Rogel-Gaillard, C., Bourgeaux, N., Billault, A., Vaiman, M. & Chardon, P. 1999. Construction of a swine BAC library: application to the characterization and mapping of porcine type C endoviral elements. *Cytogenetics and Cell Genetics,* 85**,** 205-211.
Romano, C. M., de Melo, F. L., Corsini, M. A. B., Holmes, E. C. & Paolo, M. d. A. 2007. Demographic histories of ERV-K in humans, chimpanzees and rhesus monkeys. *PLoS One,* 2**,** e1026.
Roos, J., Aggarwal, R. K. & Janke, A. 2007. Extended mitogenomic phylogenetic analyses yield new insight into crocodylian evolution and their survival of the Cretaceous-Tertiary boundary. *Molecular Phylogenetics and Evolution,* 45**,** 663-673.
Rosenberg, N. & Jolicoeur, P. 1997. Retroviral pathogenesis. *In:* Coffin, J. M., Hughes, S. H. & Varmus, H. E. (eds.) *Retroviruses.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Rowe, H. M. & Trono, D. 2011. Dynamic control of endogenous retroviruses during development. *Virology,* 411**,** 273-287.
Russello, M. A., Brazaitis, P., Gratten, J., Watkins-Colwell, G. J. & Caccone, A. 2007. Molecular assessment of the genetic integrity, distinctiveness and phylogeographic context of the saltwater crocodile (*Crocodylus porosus*) on Palau. *Conservation Genetics,* 8**,** 777-787.
Sambrook, J. & Russell, D. W. 2006. Purification of nucleic acids by extraction with phenol: chloroform. *Cold Spring Harbor Protocols,* 2006**,** pdb. prot4455.
Schmitz, A., Mausfeld, P., Hekkala, E., Shine, T., Nickel, H., Amato, G. & Böhme, W. 2003. Molecular evidence for species level divergence in African Nile crocodiles *Crocodylus niloticus* (Laurenti, 1786). *Comptes Rendus Palevol,* 2**,** 703-712.
Schuster, S. C. 2007. Next-generation sequencing transforms today's biology. *Nature,* 200.
Schwartz, D. E., Tizard, R. & Gilbert, W. 1983. Nucleotide sequence of rous sarcoma virus. *Cell,* 32**,** 853-869.
Shan, X., Ray, D. A., Bunge, J. A. & Peterson, D. G. 2009. A bacterial artificial chromosome library for the Australian saltwater crocodile (*Crocodylus porosus*) and its utilization in gene isolation and genome characterization. *BMC Genomics,* 10 Suppl 2**,** S9.
Shih, A., Coutavas, E. E. & Rush, M. G. 1991. Evolutionary implications of primate endogenous retroviruses. *Virology,* 182**,** 495-502.
Shiina, T., Tamiya, G., Oka, A., Takishima, N., Yamagata, T., Kikkawa, E., Iwata, K., Tomizawa, M., Okuaki, N., Kuwano, Y., Watanabe, K., Fukuzumi, Y., Itakura, S., Sugawara, C., Ono, A., Yamazaki, M., Tashiro, H., Ando, A., Ikemura, T., Soeda, E., Kimura, M., Bahram, S. & Inoko, H. 1999. Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region. *Proceedings*

of the National Academy of Sciences of the United States of America, 96**,** 13282-13287.
Shim-Prydon, G. & Camancho, H. 2007. New animal products: new uses for by-products and co-products of crocodile, emu, goat, kangaroo and rabbit. Canberra: Rural Industries Research and Development Corporation Union Offset Printing
Smit, A. F. A. & Hubley, R. 2008-2010. RepeatModeler Open-1.0, http://www.repeatmasker.org.
Smit, A. F. A., Hubley, R. & Green, P. 1996-2010. RepeatMasker Open-3.0, http://www.repeatmasker.org.
Smith, J. M. 1992. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution,* 34**,** 126-9.
Sperber, G. O., Airola, T., Jern, P. & Blomberg, J. 2007. Automated recognition of retroviral sequences in genomic data—RetroTector©. *Nucleic Acids Research,* 35**,** 4964-4976.
St John, J., Braun, E., Isberg, S., Miles, L., Chong, A., Gongora, J., Dalzell, P., Moran, C., Bed'Hom, B., Abzhanov, A., Burgess, S., Cooksey, A., Castoe, T., Crawford, N., Densmore, L., Drew, J., Edwards, S., Faircloth, B., Fujita, M., Greenwold, M., Hoffmann, F., Howard, J., Iguchi, T., Janes, D., Khan, S., Kohno, S., de Koning, A. J., Lance, S., McCarthy, F. & McCormack, J. 2012. Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology,* 13**,** 415.
Stoye, J. P. 2001. Endogenous retroviruses: still active after all these years? *Current Biology,* 11**,** R914-R916.
Swanstrom, R., Parker, R. C., Varmus, H. E. & Bishop, J. M. 1983. Transduction of a cellular oncogene: the genesis of Rous sarcoma virus. *Proceedings of the National Academy of Sciences of the United States of America,* 80**,** 2519-2523.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution,* 28**,** 2731-2739.
Taplin, L. E. & Grigg, G. C. 1989. Historical zoogeography of the Eusuchian crocodilians: a physiological perspective. *American Zoologist,* 29**,** 885-901.
Tarlinton, R., Meers, J., Hanger, J. & Young, P. 2005. Real-time reverse transcriptase PCR for the endogenous koala retrovirus reveals an association between plasma viral load and neoplastic disease in koalas. *Journal of General Virology,* 86**,** 783-787.
Temin, H. M. 1976. The DNA provirus hypothesis. *Science,* 192**,** 1075-1080.
Temin, H. M. 1985. Reverse transcription in the eukaryotic genome: retroviruses pararetroviruses retrotransposons and retrotranscripts. *Molecular Biology and Evolution,* 2**,** 455-468.
Temin, H. M. 1992. Origin and general nature of retroviruses. *In:* Levy, J. A. (ed.) *The Retroviridae.* New York, NY: Plenum Press.
ten Bosch, J. R. & Grody, W. W. 2008. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *The Journal of Molecular Diagnostics,* 10**,** 484-492.
Thomas, J. H. & Schneider, S. 2011. Coevolution of retroelements and tandem zinc finger genes. *Genome Research,* 21**,** 1800-1812.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* 22**,** 4673-4680.

Tristem, M. 1996. Amplification of divergent retro-elements by PCR. *BioTechniques,* 20**,** 608-612.
Tristem, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the Human Genome Mapping Project database. *Journal of Virology,* 74**,** 3715-3730.
Tristem, M., Herniou, E., Summers, K. & Cook, J. 1996. Three retroviral sequences in amphibians are distinct from those in mammals and birds. *Journal of Virology,* 70**,** 4864-4870.
Tristem, M., Myles, T. & Hill, F. 1995. A Highly Divergent Retroviral Sequence in the Tuatara (*Sphenodon*). *Virology,* 210**,** 206-211.
Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M. I., Kidd, K. K. & Lenz, J. 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current Biology,* 11**,** 1531-1535.
Turton, J. A., Ladds, P. W., Manolis, S. C. & Webb, G. J. W. 1997. Relationship of blood corticosterone, immunoglobulin and haematological values in young crocodiles (*Crocodylus porosus*) to water temperature, clutch of origin and body weight. *Australian Veterinary Journal,* 75**,** 114-119.
Varmus, H. E. & Brown, P. 1989. Retroviruses. *In:* Berg, D. E. & Howe, M. M. (eds.) *Mobile DNA.* Washington, DC: American Society for Microbiology.
Villesen, P. 2007. FaBox: an online toolbox for fasta sequences. *Molecular Ecology Notes,* 7**,** 965-968.
Vogt, V. M. 1997. Retroviral virions and genomes. *In:* Coffin, J. M., Hughes, S. H. & Varmus, H. E. (eds.) *Retroviruses.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Warr, G. W., Chapman, R. W. & Smith, L. C. 2003. Evolutionary immunobiology: new approaches, new paradigms. *Developmental and Comparative Immunology,* 27**,** 257-262.
Webb, G., Manolis, S., Whitehead, P. & Letts, G. 1984. *A proposal for the transfer of the Australian population of Crocodylus porosus Schneider (1801), from Appendix I to Appendix II of CITES [animal protection; ranching]*, Conservation Commission of the Northern Territory.
Webb, G. J. W., Britton, A. R. C., Manolis, S. C., Ottley, B. & Stirrat, S. Year. The recovery of *Crocodylus porosus* in the Northern Territory of Australia: 1971-1998. *In:* 15th Working Meeting of the Crocodile Specialist Group of the Species Survival Commission of the IUCN, 2000 Varadero, Cuba. Gland: Switzerland: IUCN, 195-234.
Webb, G. J. W. & Manolis, S. C. 2010. Australian freshwater crocodile *Crocodylus johnstoni*. *In:* Manolis, S. C. & Stevenson, C. (eds.) *Crocodiles. Status Survey and Conservation Action Plan. Third Edition.* Darwin: Crocodile Specialist Group.
Webb, G. J. W., Whitehead, P. & Manolis, S. C. 1987. Crocodile management in the Northern Territory of Australia. *In:* Webb, G. J. W., Manolis, S. C. & Whitehead, P. (eds.) *Wildlife management: Crocodiles and alligators.* Chipping Norton, NSW: Surrey Beatty & Sons in association with the Conservation Commission of the Northern Territory.
Wellehan, J. F. X. & Johnson, A. J. 2005. Reptile virology. *Veterinary Clinics of North America Exotic Animal Practice,* 8**,** 27-52.
Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. & Schulman, A. H. 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics,* 8**,** 973-982.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T., Keller, B. & Stein, N. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics,* 7**,** 275.
Wu, Y. 2004. HIV-1 gene expression: lessons from provirus and non-integrated DNA. *Retrovirology,* 1**,** 13.
Wu, Y. 2008. The second chance story of HIV-1 DNA: unintegrated? Not a problem! *Retrovirology,* 5**,** 61.
Xiong, Y. & Eickbush, T. H. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. *Molecular Biology and Evolution,* 5**,** 675-690.
Xiong, Y. & Eickbush, T. H. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal,* 9**,** 3353-3362.
Xu, Z. & Wang, H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research,* 35**,** W265-W268.
Yang, Z. 2007. PAML 4: phylogenetic analysis by Maximum Likelihood. *Molecular Biology and Evolution,* 24**,** 1586-1591.
Ye, C., Wu, X., Yan, P. & Amato, G. 2010. β-Keratins in crocodiles reveal amino acid homology with avian keratins. *Molecular Biology Reports,* 37**,** 1169-1174.
Yohn, C. T., Jiang, Z., McGrath, S. D., Hayden, K. E., Khaitovich, P., Johnson, M. E., Eichler, M. Y., McPherson, J. D., Zhao, S. & Pääbo, S. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biology,* 3**,** e110.
Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology,* 7**,** 203-214.

Appendix I: Supplementary Tables

List of Supplementary Tables

Table S2.1: GenBank accession numbers and chromosome locations for published sequences used in this thesis.

Table S2.2: Summary of the predicted recombinant and parental sequences as determined by RDP.

Table S2.3: PAML analysis showing parameter estimates, average d_N/d_S, and significance of the pairwise test comparisons.

Table S4.1: Additional primers used to confirm the presence of the probe sequences in selected BAC clones.

Table S4.2: CAG_SXT and RL MID tag sequences used for hierarchical tagging of the purified BAC DNA in preparation for sequencing.

Table S4.3: Summary of the 48 sequenced BAC clones and their corresponding CAG_SXT and RL MID tags.

Table S4.4: Assembly statistics for the 48 BAC clones sequences at low coverage and 9 MHC related BAC clones.

Table S4.5: Summary of ERV insertions detected within the sequences BAC clones, including scaffold ID and sequence characteristics.

NOTE: Due to the size and formatting of this table, some columns have been deleted and only selected rows have been presented here. Please see the attached CD for the complete table.

Table S6.1: Summary of the ERV insertions detected within the three crocodilian genomes.

NOTE: Due to the size and formatting of this table, some ERV IDs have been truncated, some columns have been deleted and only selected rows have been presented here. Please see the attached CD for the complete table.

Table S2.1: GenBank accession numbers and chromosome locations for published sequences used in this thesis.

GenBank ID [a]	Sequence ID	Reference
AJ225210	RV-Brown trout	Herniou et al. (1998)
AJ225211	RV-Common possum	Herniou et al. (1998)
AJ225212	RV-Edible frog	Herniou et al. (1998)
AJ225213	RV-European frog	Herniou et al. (1998)
AJ225214	RV-Freshwater houting	Herniou et al. (1998)
AJ225215	RV-Gharial	Herniou et al. (1998)
AJ225216	RV-Horse	Herniou et al. (1998)
AJ225217	RV-Iberian frog	Herniou et al. (1998)
AJ225218	RV-Leopard frog	Herniou et al. (1998)
AJ225219	RV-Painted frog	Herniou et al. (1998)
AJ225220	RV-Palmate newtI	Herniou et al. (1998)
AJ225221	RV-Palmate newtII	Herniou et al. (1998)
AJ225222	RV-Pit viper	Herniou et al. (1998)
AJ225223	RV-Pufferfish	Herniou et al. (1998)
AJ225224	RV-Rhinatremid caecilianI	Herniou et al. (1998)
AJ225225	RV-Rhinatremid caecilianII	Herniou et al. (1998)
AJ225226	RV-Rocket frog	Herniou et al. (1998)
AJ225227	RV-Slider turtleI	Herniou et al. (1998)
AJ225228	RV-Slider turtleII	Herniou et al. (1998)
AJ225229	RV-Stickleback	Herniou et al. (1998)
AJ225230	RV-Stripe faced dunnartI	Herniou et al. (1998)
AJ225231	RV-Stripe faced dunnartII	Herniou et al. (1998)
AJ225232	RV-Tiger salamanderI	Herniou et al. (1998)
AJ225233	RV-Tiger salamanderII	Herniou et al. (1998)
AJ225234	RV-Tiger salamanderIII	Herniou et al. (1998)
AJ225235	RV-Tinamou	Herniou et al. (1998)
AJ225236	RV-Tuatara	Herniou et al. (1998)
AJ225207	RV-African clawed toad	Jaratlerdsiri et al. (2009)
AJ225208	RV-Bower bird	Jaratlerdsiri et al. (2009)
AJ225209	RV-Brook trout	Jaratlerdsiri et al. (2009)
FJ155500	RV-Cpa-II	Jaratlerdsiri et al. (2009)
FJ155501	RV-Cla-I	Jaratlerdsiri et al. (2009)
FJ155502	RV-Cla-II	Jaratlerdsiri et al. (2009)
FJ155503	RV-Cla-III	Jaratlerdsiri et al. (2009)
FJ155504	RV-Asi-III	Jaratlerdsiri et al. (2009)
FJ155505	RV-Asi-IV	Jaratlerdsiri et al. (2009)
FJ155506	RV-Crh-I	Jaratlerdsiri et al. (2009)
FJ155507	RV-Crh-II	Jaratlerdsiri et al. (2009)
FJ155508	RV-Crh-III	Jaratlerdsiri et al. (2009)

GenBank ID [a]	Sequence ID	Reference
FJ155509	RV-Cno-I	Jaratlerdsiri et al. (2009)
FJ155510	RV-Cno-II	Jaratlerdsiri et al. (2009)
FJ155511	RV-Cno-III	Jaratlerdsiri et al. (2009)
FJ155512	RV-Cjo-I	Jaratlerdsiri et al. (2009)
FJ155513	RV-Cjo-II	Jaratlerdsiri et al. (2009)
FJ155514	RV-Cjo-III	Jaratlerdsiri et al. (2009)
FJ155515	RV-Cni-I	Jaratlerdsiri et al. (2009)
FJ155516	RV-Cni-II	Jaratlerdsiri et al. (2009)
FJ155517	RV-Ppa-I	Jaratlerdsiri et al. (2009)
FJ155518	RV-Ppa-II	Jaratlerdsiri et al. (2009)
FJ155519	RV-Ppa-III	Jaratlerdsiri et al. (2009)
FJ155520	RV-Csi-I	Jaratlerdsiri et al. (2009)
FJ155521	RV-Csi-II	Jaratlerdsiri et al. (2009)
FJ155522	RV-Csi-III	Jaratlerdsiri et al. (2009)
FJ155523	RV-Cin-I	Jaratlerdsiri et al. (2009)
FJ155524	RV-Cin-II	Jaratlerdsiri et al. (2009)
FJ155525	RV-Cya-I	Jaratlerdsiri et al. (2009)
FJ155526	RV-Cya-II	Jaratlerdsiri et al. (2009)
FJ155527	RV-Cya-III	Jaratlerdsiri et al. (2009)
FJ155528	RV-Ami-I	Jaratlerdsiri et al. (2009)
FJ155529	RV-Ami-II	Jaratlerdsiri et al. (2009)
FJ155530	RV-Ami-III	Jaratlerdsiri et al. (2009)
FJ155531	RV-Ote-I	Jaratlerdsiri et al. (2009)
FJ155532	RV-Ote-II	Jaratlerdsiri et al. (2009)
FJ155533	RV-Ote-III	Jaratlerdsiri et al. (2009)
FJ155534	RV-Cmo-I	Jaratlerdsiri et al. (2009)
FJ155535	RV-Cmo-II	Jaratlerdsiri et al. (2009)
FJ155536	RV-Cmo-III	Jaratlerdsiri et al. (2009)
FJ155537	RV-Cca	Jaratlerdsiri et al. (2009)
FJ155538	RV-Ccr-I	Jaratlerdsiri et al. (2009)
FJ155539	RV-Ccr-II	Jaratlerdsiri et al. (2009)
FJ155540	RV-Ccr-III	Jaratlerdsiri et al. (2009)
FJ155541	RV-Mni-I	Jaratlerdsiri et al. (2009)
FJ155542	RV-Mni-II	Jaratlerdsiri et al. (2009)
FJ155543	RV-Cmi-I	Jaratlerdsiri et al. (2009)
FJ155544	RV-Cmi-II	Jaratlerdsiri et al. (2009)
FJ155545	RV-Cpo-I	Jaratlerdsiri et al. (2009)
FJ155546	RV-Cpo-II	Jaratlerdsiri et al. (2009)
FJ155547	RV-Cpo-III	Jaratlerdsiri et al. (2009)
FJ155548	RV-Cpo-IV	Jaratlerdsiri et al. (2009)
FJ155549	RV-Cpo-V	Jaratlerdsiri et al. (2009)
FJ155550	RV-Cpo-VI	Jaratlerdsiri et al. (2009)

GenBank ID [a]	Sequence ID	Reference
FJ155551	RV-Cpo-VII	Jaratlerdsiri et al. (2009)
FJ155552	RV-Cpo-VIII	Jaratlerdsiri et al. (2009)
FJ155553	RV-Cpo-IX	Jaratlerdsiri et al. (2009)
FJ155554	RV-Cpo-X	Jaratlerdsiri et al. (2009)
FJ155555	RV-Cpo-XI	Jaratlerdsiri et al. (2009)
FJ155556	RV-Cpo-XII	Jaratlerdsiri et al. (2009)
FJ155557	RV-Cpo-XIII	Jaratlerdsiri et al. (2009)
FJ155558	RV-Cpo-XIV	Jaratlerdsiri et al. (2009)
FJ155559	RV-Csi-IV	Jaratlerdsiri et al. (2009)
FJ155560	RV-Csi-V	Jaratlerdsiri et al. (2009)
FJ155561	RV-Csi-VI	Jaratlerdsiri et al. (2009)
NC001408	ALV	Jern et al. (2005)
D10032	BaEV	Jern et al. (2005)
NC001414	BLV	Jern et al. (2005)
	EIAV	Jern et al. (2005)
Chr7-63865366 [b]	ERV-3	Jern et al. (2005)
NC001940	FLV	Jern et al. (2005)
M26927	GaLV	Jern et al. (2005)
	gg01-chr1-156168845 [c]	Jern et al. (2005)
	gg01-Chr4-48130894 [c]	Jern et al. (2005)
	gg01-chr4-77338201 [c]	Jern et al. (2005)
	gg01-chr7-5733782 [c]	Jern et al. (2005)
	gg01-chr7-7163462 [c]	Jern et al. (2005)
	gg01-ChrU-126703652 [c]	Jern et al. (2005)
	gg01-ChrU-163504869 [c]	Jern et al. (2005)
	gg01-chrU-49656081 [c]	Jern et al. (2005)
	gg01-chrU-52190725 [c]	Jern et al. (2005)
AJ000387	Gypsy	Jern et al. (2005)
AC005741	HERV-ADP	Jern et al. (2005)
M10976	HERV-E	Jern et al. (2005)
AL354685	HERV-Fc1	Jern et al. (2005)
AC019088	HERV-Fc2	Jern et al. (2005)
AC004022	HERV-FRD	Jern et al. (2005)
	HERV-Hconsensus	Jern et al. (2005)
D11078	HERVH-RGH2	Jern et al. (2005)
M18048	HERVH-RTVLH2	Jern et al. (2005)
Chr16-72821350 [b]	HERV-I	Jern et al. (2005)
RepBase	HERV-L	Jern et al. (2005)
RepBase	HERVL40	Jern et al. (2005)
RepBase	HERVL66	Jern et al. (2005)
RepBase	HERVL74	Jern et al. (2005)
AC004385	HERV-S	Jern et al. (2005)

GenBank ID [a]	Sequence ID	Reference
Chr14-104635791 [b]	HERV-T	Jern et al. (2005)
Chr7-9105739 [b]	HERV-W	Jern et al. (2005)
NC001736	HFV	Jern et al. (2005)
	hg15-chr3-152465283 [b]	Jern et al. (2005)
	HIV-1	Jern et al. (2005)
	HIV-2	Jern et al. (2005)
Chr19-21849393 [b]	HML1	Jern et al. (2005)
Chr11-101600013 [b]	HML2	Jern et al. (2005)
Chr1-48344461 [b]	HML3	Jern et al. (2005)
Chr8-75679221 [b]	HML4	Jern et al. (2005)
AC004536	HML5	Jern et al. (2005)
	HML6	Jern et al. (2005)
Chr6-121300220 [b]	HML7	Jern et al. (2005)
Chr3-131452286 [b]	HML8	Jern et al. (2005)
Chr9-62700428 [b]	HML9	Jern et al. (2005)
AF033816	HSRV	Jern et al. (2005)
NC001436	HTLV-1	Jern et al. (2005)
NC001488	HTLV-2	Jern et al. (2005)
M80216	JSRV	Jern et al. (2005)
Chr13-54208300 [b]	MER4like	Jern et al. (2005)
NC001501	MLV	Jern et al. (2005)
NC001503	MMTV	Jern et al. (2005)
NC001550	MPMV	Jern et al. (2005)
AJ293656	PERV	Jern et al. (2005)
	pt01-Chr10r-17119458 [d]	Jern et al. (2005)
	pt01-Chr5-53871501 [d]	Jern et al. (2005)
AAN77283	Python-molurus	Jern et al. (2005)
NC001724	SnRV	Jern et al. (2005)
	Visna	Jern et al. (2005)
NC001867	WDSV	Jern et al. (2005)
AJ506107	Xen1	Jern et al. (2005)
AJ236109	RV-Green anole	Martin et al. (1999)
AJ236110	RV-Puff adder	Martin et al. (1999)
AJ236111	RV-Boa constrictor	Martin et al. (1999)
AJ236112	RV-Pit viper	Martin et al. (1999)
AJ236113	RV-Bower Bird III	Martin et al. (1999)
AJ236114	RV-Bower Bird II	Martin et al. (1999)
AJ236115	RV-Natterjack Toad	Martin et al. (1999)
AJ236116	RV-Rhinatrematid caecilianIII	Martin et al. (1999)
AJ236117	RV-Rhinatrematid caecilianIV	Martin et al. (1999)
AJ236118	RV-Yellow	Martin et al. (1999)
AJ236119	RV-Echidna	Martin et al. (1999)

GenBank ID [a]	Sequence ID	Reference
AJ236120	RV-Garter snake	Martin et al. (1999)
AJ236121	RV-Komodo dragon	Martin et al. (1999)
AJ236122	RV-Koala	Martin et al. (1999)
AJ236123	RV-Opossum	Martin et al. (1999)
AJ236124	RV-African	Martin et al. (1999)
AJ236125	RV-Partridge I	Martin et al. (1999)
AJ236126	RV-Partridge II	Martin et al. (1999)
AJ236127	RV-Pheasant	Martin et al. (1999)
AJ236128	RV-Edible Frog	Martin et al. (1999)
AJ236129	RV-Redwing	Martin et al. (1999)
AJ236130	RV-Rook	Martin et al. (1999)
AJ236131	RV-False gharial	Martin et al. (1999)
AJ236132	RV-European Adder	Martin et al. (1999)
AJ236133	RV-Wood Pigeon	Martin et al. (1999)
AJ236134	RV-Wren	Martin et al. (1999)
FJ155497	RV-Cac-I	Martin et al. (1999)
FJ155498	RV-Cac-II	Martin et al. (1999)
FJ155499	RV-Cpa-I	Martin et al. (1999)
AJ438130	CnEVI	Martin et al. (2002)
AJ438131	CnEVII	Martin et al. (2002)
AJ438132	CnEVIII	Martin et al. (2002)
AJ438133	RV-Chinese alligator I	Martin et al. (2002)
AJ438134	RV-Chinese alligator II	Martin et al. (2002)
AJ438135	RV-Smooth-fronted caiman	Martin et al. (2002)
AJ438136	RV-Broad nosed caiman	Martin et al. (2002)
AJ438137	RV-Orinoco crocodile	Martin et al. (2002)
AJ438138	RV-Gharial II	Martin et al. (2002)

[a] GenBank accession numbers as provided by the respective publications.

[b] Insertions were retrieved and generated from the *Homo sapiens* reference assemblies hg15 and 16 (International Human Genome Sequencing Consortium, 2001).

[c] Insertions were generated from the *Gallus gallus* genome assembly gg01 (International Chicken Genome Sequencing Consortium, 2004).

[d] Insertions were generated from the *Pan troglodytes* reference assembly panTro1 (Chimpanzee Sequencing and Analysis Consortium, 2005).

Table S2.2: Summary of the predicted recombinant and parental sequences as determined by RDP.

Recombinant Sequence	Breakpoint Positions		Major Parental Sequence[a]	Minor Parental Sequence[b]	Probability that the detected sequence is a recombinant[c]						
	Begin	End			RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiSscan	3Seq
CERV1-All											
1	289	906	148	Cpo_VII	0.0230	NS	0.0234	0.0001	0.0107	0.0000	0.0000
Csi_IV	753	872	164	142	0.0301	NS	0.0306	NS	NS	NS	0.0344
Cni_I			164	142	Trace evidence only						
CERV1-porosus											
1	289	869	175	Cpo_VII	0.0005	NS	0.0005	0.0001	0.0021	NS	0.0000
CERV2-All											
81	23	296	53	52	0.0001	0.0020	0.0001	0.0001	0.0001	0.0000	0.0000
Ami_II	698	1	Asi_IV	Ppa_II	0.0398	NS	NS	NS	NS	NS	0.0215
Ami_I			Asi_IV	Ppa_II	Trace evidence only						
CERV2-porosus											
81	23	296	53	52	0.0001	0.0013	0.0001	0.0001	0.0001	0.0000	0.0000
60	16	191	162	34	0.0329	NS	NS	NS	NS	NS	0.0003

[a] Major Parent: Parent contributing the larger fraction of sequence.

[b] Minor Parent: Parent contributing the smaller fraction of sequence.

[c] NS: No significant *P*-value was recorded for this recombination event using this method.

Table S2.3: PAML analysis showing parameter estimates, average d_N/d_S, and significance of the pairwise test comparisons.

H_n	Model	Ave d_N/d_S	PAML parameter estimates	lnL	LRT test statistic	*df*	*P*-value
CERV1							
H0	M0: One ratio	0.4904	ω = 0.49042	-6936.0525	273.3946	4	< 0.001
H1	M3: Discrete	0.5187	p0 = 0.63002 p1 = 0.36034 p2 = 0.00964 ω0 = 0.14883 ω1 = 0.97228 ω2 = 7.73167	-6799.3552			
H2	M1a: Nearly neutral	0.4344	p0 = 0.66860 p1 = 0.33140P2 ω0 = 0.15403 ω1 = 1.00000	-6845.3939	92.0213	2	< 0.001
H3	M2a: Positive selection	0.5234	p0 = 0.64094 p1 = 0.34945 p2 = 0.00962 ω0 = 0.15461 ω1 = 1.00000 ω2 = 7.78576	-6799.3832			
H4	M7: Beta	0.4217	p = 0.37192 q = 0.50996	-6848.2309	96.8512	2	< 0.001
H5	M8: Beta & ω	0.4973	p0 = 0.99019 p = 0.40522 q = 0.54097 (p1 = 0.00981) ω = 7.46662	-6799.8053			
CERV2							
H0	M0: One ratio	0.5539	ω = 0.55393	-4029.5600	413.1006	4	< 0.001
H1	M3: Discrete	0.6898	p0 = 0.79438 p1 = 0.18634 p2 = 0.01928 ω0 = 0.14019 ω1 = 1.66189 ω2 = 13.93868	-3823.0097			
H2	M1a: Nearly neutral	0.3057	p0 = 0.76252 p1 = 0.23748 ω0 = 0.08951 ω1 = 1.00000	-3921.6374	178.2283	2	< 0.001
H3	M2a: Positive selection	0.5864	p0 = 0.72223 p1 = 0.25559 p2 = 0.02218 ω0 = 0.10433 ω1 = 1.00000 ω2 = 11.51932	-3832.5232			
H4	M7: Beta	0.3349	p = 0.17755 q = 0.35268	-3923.0792	164.9022	2	< 0.001
H5	M8: Beta & ω	0.6217	p0 = 0.97733 p = 0.21801 q = 0.39323 (p1 = 0.02267) ω = 12.04341	-3840.6281			

Table S4.1: Additional primers used to confirm the presence of the probe sequences in selected BAC clones.

Primer name	Primer sequence
ERV1-1-956-F	GTK TTG KTG GAY ACG GGG KC
ERV1-1-956-R	ATG AGG AKR TCR TCG ACR TA
ERV1-69-375-F	ACA GCA TGT AYT TGT RGA AG
ERV1-69-375-R	CGY GCA GGG GTT CCA TCA GCC
ERV1-88-845-F	GGG ATT GAG GGA ATG CAA AC
ERV1-88-845-R	ATG GGG GGC TGC TAT AAA TC

Table S4.2: CAG_SXT and RL MID tag sequences used for hierarchical tagging of the purified BAC DNA in preparation for sequencing.

Linker	Sequence	
CAG-SXT	**Upper (_U)**	**Lower (_Lp)**
CAG_SimpleXT03	CAGTCGGGCGTCATCACGTGCTGCGGAACTT	/5`Phos/AGTTCCGCAGCACGTGATGACGCCCGAC
CAG_SimpleXT04	CAGTCGGGCGTCATCAGCAGCAGCGGAATCT	/5`Phos/GATTCCGCTGCTGCTGATGACGCCCGAC
CAG_SimpleXT06	CAGTCGGGCGTCATCAGCACGAGCGGAACTT	/5`Phos/AGTTCCGCTCGTGCTGATGACGCCCGAC
CAG_SimpleXT07	CAGTCGGGCGTCATCAGGTCGAGCGGAATGT	/5`Phos/CATTCCGCTCGACCTGATGACGCCCGAC
CAG_SimpleXT08	CAGTCGGGCGTCATCAGGTGCAGGCGAATGT	/5`Phos/CATTCGCCTGCACCTGATGACGCCCGAC
CAG_SimpleXT11	CAGTCGGGCGTCATCACGAGCGAGCGAAGTT	/5`Phos/ACTTCGCTCGCTCGTGATGACGCCCGAC
CAG_SimpleXT12	CAGTCGGGCGTCATCAGCTGGCGTCGAAGTT	/5`Phos/ACTTCGACGCCAGCTGATGACGCCCGAC
CAG_SimpleXT13	CAGTCGGGCGTCATCACCAGCACCGGAACAT	/5`Phos/TGTTCCGGTGCTGGTGATGACGCCCGAC
CAG_SimpleXT14	CAGTCGGGCGTCATCACCTGGGCACGAAGAT	/5`Phos/TCTTCGTGCCCAGGTGATGACGCCCGAC
CAG_SimpleXT15	CAGTCGGGCGTCATCACGTCGTGCGGAAACT	/5`Phos/GTTTCCGCACGACGTGATGACGCCCGAC
CAG_SimpleXT16	CAGTCGGGCGTCATCAGCAGCGTCGGAAAGT	/5`Phos/CTTTCCGACGCTGCTGATGACGCCCGAC
CAG_SimpleXT17	CAGTCGGGCGTCATCAGCTCCTGGCGAATCT	/5`Phos/GATTCGCCAGGAGCTGATGACGCCCGAC
RL-MID		
RL-1	ACACGACGACT	
RL-2	ACACGTAGTAT	
RL-3	ACACTACTCGT	
RL-4	ACGACACGTAT	
RL-5	ACGAGTAGACT	
RL-6	ACGCGTCTAGT	
RL-7	ACGTACACACT	
RL-8	ACGTACTGTGT	
RL-9	ACGTAGATCGT	
RL-10	ACTACGTCTCT	
RL-11	ACTATACGAGT	
RL-12	ACTCGCGTCGT	

Table S4.3: Summary of the 48 sequenced BAC clones and their corresponding CAG_SXT and RL MID tags.

Plate No.	Well	BAC ID	CAG_SXT-## Pools 1-4		CAG_SXT-## Pool 5
001	B17	001-B17	3		3
004	E14	004-E14	4		
010	F2	010-F2	6		
010	M6	010-M6	7		
011	C17	011-C17	8		
011	P21	011-P21	11	Pool 1	
016	E20	016-E20	12		
017	C17	017-C17	13		
024	L19	024-L19	14		
028	L15	028-L15	15		
030	I19	030-I19	16		4
036	I4	036-I4	17		6
037	K13	037-K13	3		
038	B10	038-B10	4		
039	I20	039-I20	6		
040	B24	040-B24	7		
040	F9	040-F9	8		
040	O12	040-O12	11	Pool 2	
046	O12	046-O12	12		
048	D4	048-D4	13		7
085	E21	085-E21	14		
088	E18	088-E18	15		
090	G4	090-G4	16		
090	P22	090-P22	17		8
092	H11	092-H11	3		11
100	E19	100-E19	4		
102	K13	102-K13	6		
103	M4	103-M4	7		
103	N11	103-N11	8		
109	I20	109-I20	11	Pool 3	
110	M11	110-M11	12		
119	H8	119-H8	13		
126	M7	126-M7	14		12
136	B18	136-B18	15		
141	H19	141-H19	16		
145	N21	145-N21	17		13
153	O20	153-O20	3		
154	C15	154-C15	4	Pool 4	
154	C4	154-C4	6		

Plate No.	Well	BAC ID	CAG_SXT-## Pools 1-4	CAG_SXT-## Pool 5
166	F15	166-F15	7	14
168	L14	168-L14	8	
177	H4	177-H4	11	
179	J15	179-J15	12	
186	N1	186-N1	13	
205	D13	205-D13	14	
212	L13	212-L13	15	
215	B14	215-B14	16	
222	M7	222-M7	17	

Table S4.4: Assembly statistics for the 48 BAC clones sequences at low coverage and 9 MHC related BAC clones.

	Contigs				Scaffolds			
	Ave	**Max**	**Number**	**N50**	**Ave**	**Max**	**Number**	**N50**
Low coverage BACs								
001-B17	1727	25287	2571	2324	29255	213113	108	93335
004-E14	1715	19524	2615	2302	26133	187807	121	74309
010-F02	1725	25286	2598	2327	26552	213368	118	74699
010-M06	1705	19524	2643	2286	26138	181904	122	70448
011-C17	1694	19524	2655	2260	26625	211588	116	74006
011-P21	1702	19525	2620	2264	27409	212125	115	74075
016-E20	1705	15748	2611	2291	24419	204511	127	65396
017-C17	1697	19524	2650	2265	26877	211443	118	75653
024-L19	1688	19524	2636	2236	26513	216771	117	70216
028-L15	1712	19524	2615	2296	27822	211997	114	75202
030-I19	1715	25285	2606	2302	26042	219325	120	71939
036-I04	1709	19524	2605	2291	27558	188651	116	74984
037-K13	1709	19524	2647	2288	27105	188307	116	75022
038-B10	1712	19524	2629	2296	27309	180992	115	75336
039-I20	1706	19523	2647	2285	27637	211422	113	86768
040-B24	1700	15749	2642	2266	26028	209155	121	67117
040-F09	1707	19524	2634	2309	26848	210689	118	73600
040-O12	1722	19524	2589	2314	27189	212143	117	70562
046-O12	1708	25286	2626	2286	25899	189240	121	71031
048-D04	1720	25287	2624	2326	27159	204176	117	82678
085-E21	1690	15748	2713	2263	26103	209248	118	68016
088-E18	1696	15749	2617	2261	23914	204114	131	65965
090-G04	1717	15749	2619	2315	27555	181213	115	75957
090-P22	1707	19525	2671	2302	27711	202416	115	90328
092-H11	1720	25285	2629	2326	25149	212332	125	73315
100-E19	1714	19524	2601	2301	27913	217735	114	85367
102-K13	1709	19524	2597	2291	27415	188256	115	74941
103-M04	1706	19524	2625	2296	28148	211903	112	75385
103-N11	1712	19524	2635	2315	26560	211850	119	73325
109-I20	1710	19524	2628	2294	26269	181314	119	75614
110-M11	1706	19524	2606	2302	27039	213240	116	75855
119-H08	1709	19525	2596	2294	26386	212834	120	74532
126-M07	1719	25286	2594	2315	27734	180457	114	76072
136-B18	1704	25285	2623	2278	28897	188503	109	76067
141-H19	1721	19524	2591	2327	26786	185633	118	70701
145-N21	1701	25286	2688	2301	26894	204109	118	78301
153-O20	1707	19525	2618	2288	26403	181819	119	74283

	Contigs				Scaffolds			
	Ave	**Max**	**Number**	**N50**	**Ave**	**Max**	**Number**	**N50**
154-C04	1714	19525	2637	2302	26096	181721	121	75384
154-C15	1716	25286	2610	2308	27400	213392	114	70877
166-F15	1738	19524	2622	2352	26292	212752	121	76759
168-L14	1708	25287	2629	2278	26323	214043	119	74428
177-H04	1708	25288	2626	2294	27385	211559	115	74326
179-J15	1709	19524	2642	2294	24693	210940	127	71182
186-N01	1719	19524	2595	2309	27194	181707	116	75252
205-D13	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
212-L13	1710	19524	2606	2296	26486	212020	119	85224
215-B14	1713	19523	2616	2294	25567	217046	121	74224
222-M07	1718	19524	2649	2294	27233	211650	116	74058
MHC BACs								
009-O17	7448	27545	19	14824				
012-F13	3592	31246	61	11851				
067-G16	3054	43874	68	26076				
077-H5	2020	69203	237	27948				
082-I19	10801	92953	16	92953				
092-F14	2223	43499	51	43499				
186-I16	5544	93750	32	93750				
192-O18	2237	28338	30	28338				
193-A19	8905	41015	8	41015				

Table S4.5: Summary of ERV insertions detected within the sequences BAC clones, including scaffold ID and sequence characteristics. NOTE: Due to the size and formatting of this table, some columns have been deleted and only selected rows have been presented here. Please see the attached CD for the complete table (Associated file: Appendix_I-Table_S4.5.xlsx).

ERV ID	Class	Haplotype	PBS	PPT	5` TSD	Length
RL05-contig00001-S00001	ERV1	MHC_Hap_1	N/A	aaaatggggagc	tgccc/cg	7915
RL10-contig00002-P00001	ERV1	MHC_Hap_1	N/A	aaaatggggagc	tgccc/cg	7915
RL06-contig00003-P00001	ERV1	MHC_Hap_5	tggagaagctggagagag	aagaaagggaggagagcggat	aaaa/tg	13217
RL08-contig00001-P00001	ERV1	MHC_Hap_7	tggtggcagtggtgggat	aagaagaagaggaagagaagcc	ccccc/tg	14054
RL10-contig00001-P00005	ERV1	MHC_Hap_9	tggctgccaaccccagct	aaggatgaaggggcctg	cccaa/gg	11932
004-E14-scaffold00024-P00001	ERV4	Hap_7	tggaactttccccttgat	agagcaaggaaaaggaat	cccca/tg	9633
119-H08-scaffold00021-S00001	ERV4	Hap_69	tggtgctgcgcaggctca	agagcaaggaaaaggaat	gttat/tg	8607
166-F15-scaffold00023-S00001	ERV4	Hap_69	tggtgctgcgcaggctca	agagcaaggaaaaggaat	gttat/tg	8607
RL05-contig00003-S00001	ERV1	MHC_Hap_2	tggtagtgctcgggggga	aggcagagggagaaaagattg	gcac/tg	14472
RL10-contig00001-P00001	ERV1	MHC_Hap_8	tggtcacagggcaccatg	aggcagagggagaaaagattg	cagggc/tg	13887
RL06-contig00001-S00001	ERV1	MHC_Hap_4	N/A	aggggaaggaaggctg	agatcc/tg	5308
010-F02-scaffold00023-S00001	ERV4	Hap_9	tggaactttccccttgat	N/A	accc/tg	9826
030-I19-scaffold00023-P00001	ERV4	Hap_20	tggaactttccccttgat	N/A	cccca/tg	9726
092-H11-scaffold00020-S00001	ERV4	Hap_50	tgggtatctgtccctctg	N/A	ggctg/tg	10298
102-K13-scaffold00024-P00001	ERV4	Hap_56	tggaactttccccttgat	N/A	cccca/tg	9662
RL10-contig00001-S00001	ERV1	MHC_Hap_10	tggtgctgtgacttggat	N/A	ccatg/tg	8328

Table S6.1: Summary of the ERV insertions detected within the three crocodilian genomes. NOTE: Due to the size and formatting of this table, some ERV IDs have been truncated, some columns have been deleted and only selected rows have been presented here. Please see the attached CD for the complete table (Associated file: Appendix_I-Table_S6.1.xlsx).

ERV ID	Start	End	PBS	PPT	5`TSD	3`TSD	Length
gi|397455753_1-S00001	67660	59431	tggcgtcatgaacaggat	agggtgaaaggggatg	tgaagg/tg	ca/tgacag	8230
gi|397455787-P00001	10278	18386	tggcgacgaggatggga	agcaaggggcatg	atcc/tg	ca/atcc	8109
gi|397455981_4-S00001	361187	351581	tggtgtcaccacccggtg	aagaagagagtagc	cttgg/tg	ca/cttgg	9607
gi|397456088_2-P00001	111174	119872	tggtgccgtgactcagat	aagaagagaaaaatggagttg	gtct/tg	ca/gtct	8699
gi|397456325_1-S00001	63244	49519	tggcgtaattccagattt	agcaggagggaaggggt	ggccag/ga	ta/ggcaag	13726
gi|397456616-S00001	34882	27450	tggcaccccagatgggac	aaaaggagggangatg	catt/tg	ca/cagt	7433
gi|397456628_4-P00001	376448	385898	tggtgtggtggaagggga	aagacaaagaggtg	tgtagg/gg	ga/tcaagg	9451
gi|397456663_17-S00001	1679462	1670810	tgggcaccatgtttatag	aggagaggagcaggt	tttgca/tt	ca/tgtgcc	8653
scaffold-10156_1-S00001	89351	76122	tggcaggcgtatataccc	agagagaagtgggggtg	attgg/tg	ca/agtat	13230
scaffold-10204-S00001	45156	35571	tggcgacccaggtggtgt	aaaaggggggattg	ctgt/tg	ca/ctgt	9586
scaffold-10522_9-P00001	862467	872340	tggcactcctgtgtctct	aaaggggggagttg	cggct/tg	ca/cggct	9874
scaffold-10715-P00001	17551	25592	tggtgtcatgactcggat	aaaaaaatgagggaatg	caagg/tg	ca/caagg	8042
scaffold-10982-P00001	31320	45569	tggttcttccccatcagg	agggaatgaggaggcctg	gggat/tg	ca/gggat	14250
scaffold-1103_4-P00001	398493	405707	tgggctacctcacaagcc	aaggggacgaagagcttg	caag/tg	ca/caag	7215
scaffold-1204_1-P00001	105833	114397	tggtgtcagcagtgctta	aaagcaggagtg	cccac/tg	aa/cccac	8565
scaffold-12115-S00001	46995	39187	tggcaacccagatgggtc	aaagaggggattg	catc/tg	ca/catc	7809
scaffold10111-S00001	65174	53750	tggcgtagttggcaggat	agagaaacaaaagaggttg	agcaag/tg	ca/agtagg	11425
scaffold10246-P00001	43181	51461	tgggtgcttcgtctgcca	aagggggaggctg	gtat/tg	ca/gtat	8281
scaffold10282_1-P00001	70042	77937	tgggggcttggctggaat	aggatggaggggggaagatctg	cagct/tg	ca/ccgct	7896
scaffold10324_1-S00001	32490	21879	tggtgtaaggacataggc	agcaaggggggcatg	ctaatg/tg	ca/atgatt	10612
scaffold1040_5-P00001	450052	463017	tggtgctgtgactcagat	aaaaggggggaatgtg	atag/tg	ca/atag	12966
scaffold1075_2-S00002	130709	122578	tggtaaagaaacaggaatc	agaaaaaaagaaaacaaggt	tcata/tg	ca/tgaaa	8132

Appendix II: Supplementary Figures

List of Supplementary Figures

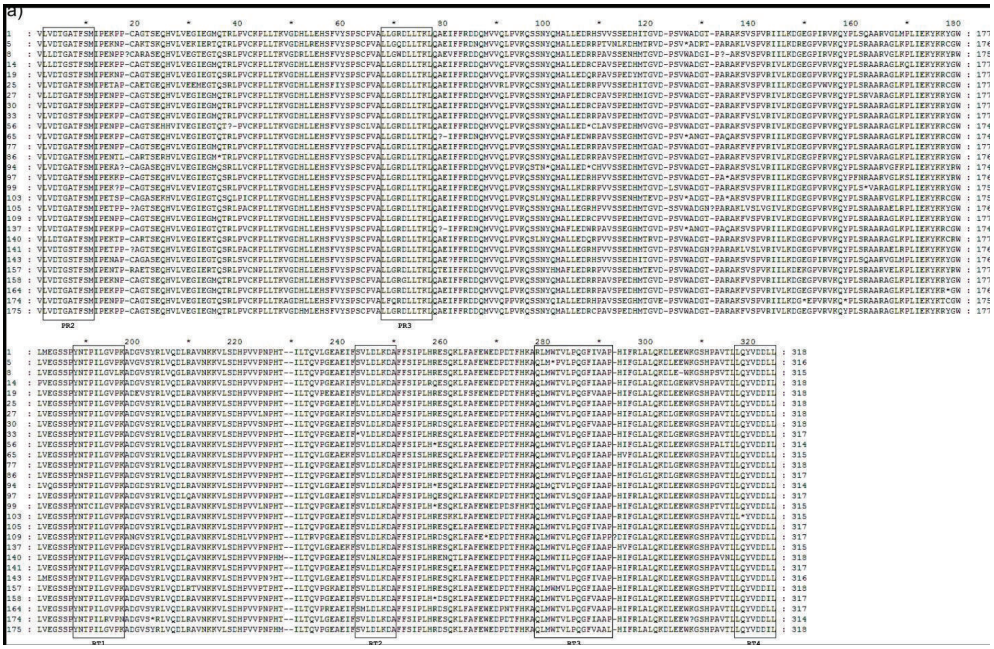
Figure S2.1: Alignments of putative amino acid translations of selected sequences from clades CERV1 (a) and CERV2 (b).

Figure S2.2: Neighbour Joining trees based on the nucleotide alignments of *C. porosus* sequences from the CERV1 (a) and CERV2 (b) clades.

Figure S3.1: *C. johnstoni pro-pol* haplotypes clustered into four major clades belonging to ERV3 and ERV4.

Figure S3.2: *C. johnstoni* ERV lineages cluster with crocodilian ERV4 sequences and within ERV3.

a)



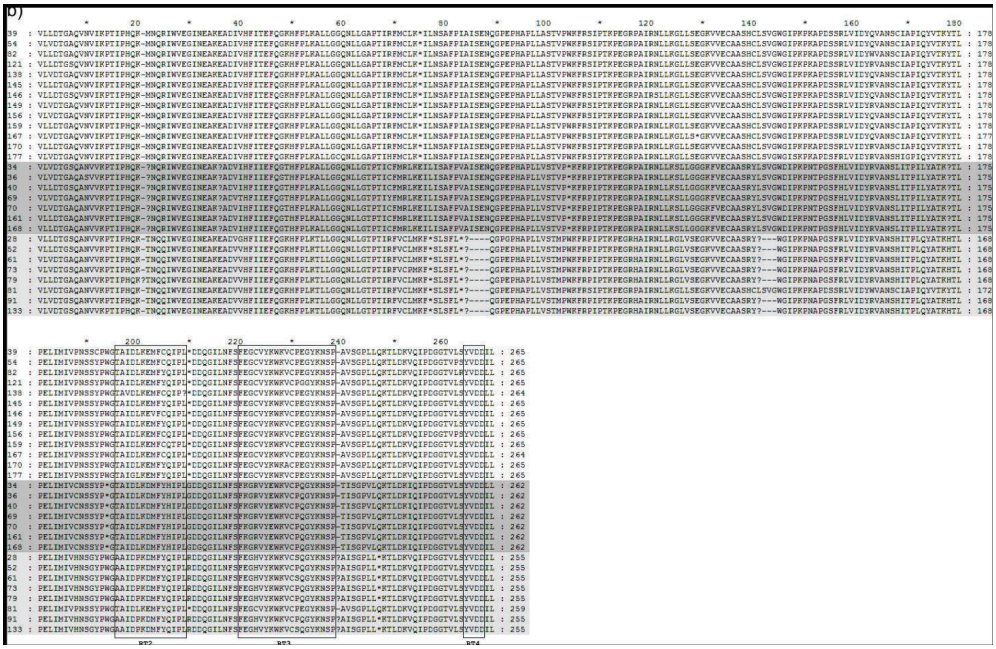


Figure S2.1: Alignments of putative amino acid translations of selected sequences from clades CERV1 (a) and CERV2 (b). Conserved retroviral motifs are bound by black boxes. The dark and light shaded boxes in (b) highlight the proposed sub-lineages within clade CERV2: the unshaded sequences are group CERV2a, dark grey are CERV2b, and lighter grey are CERV2c.

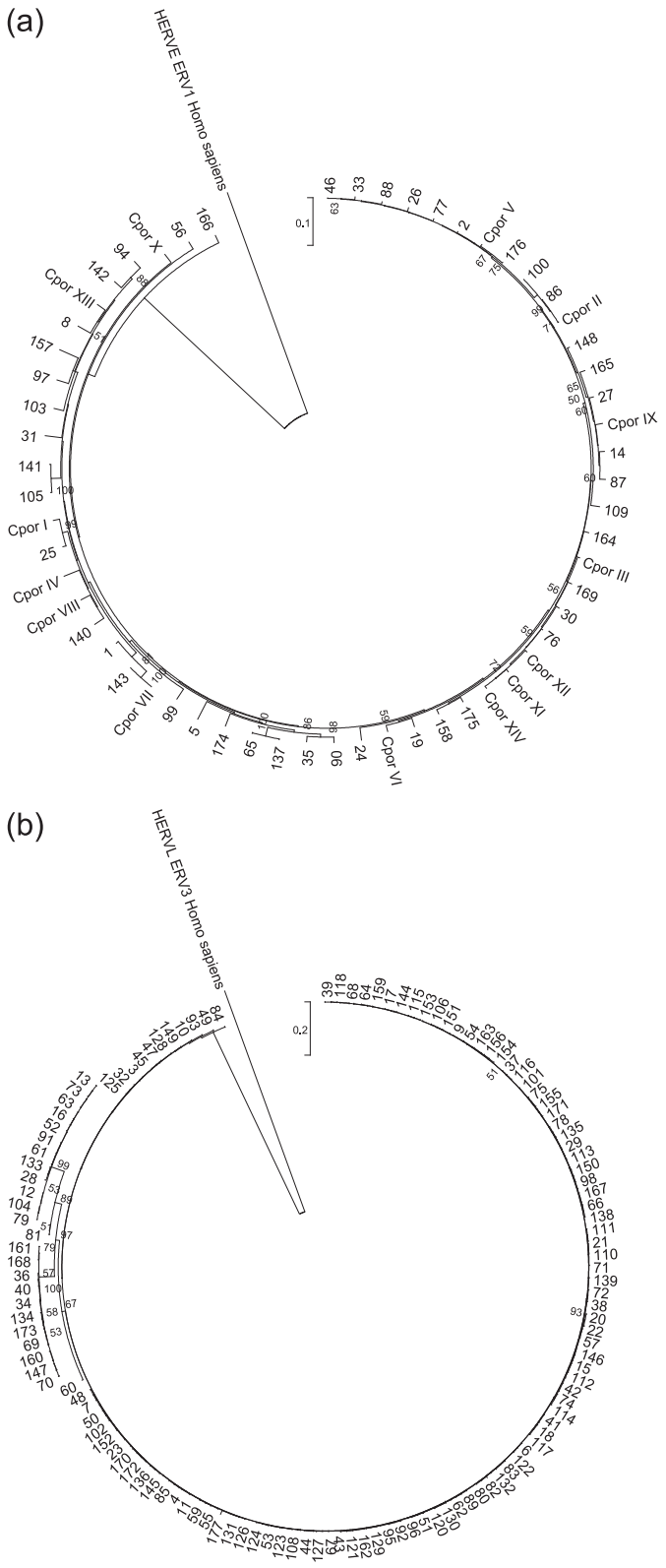


Figure S2.2: Neighbour Joining trees based on the nucleotide alignments of *C. porosus* sequences from the CERV1 (a) and CERV2 (b) clades. Outgroups are human ERV sequences extracted from the RepBase database and from the same general ERV classes as these sequences. Numbers outside the circles indicate the haplotypes included in each tree. Smaller numbers within the circle represent bootstrap support values greater than 50%.

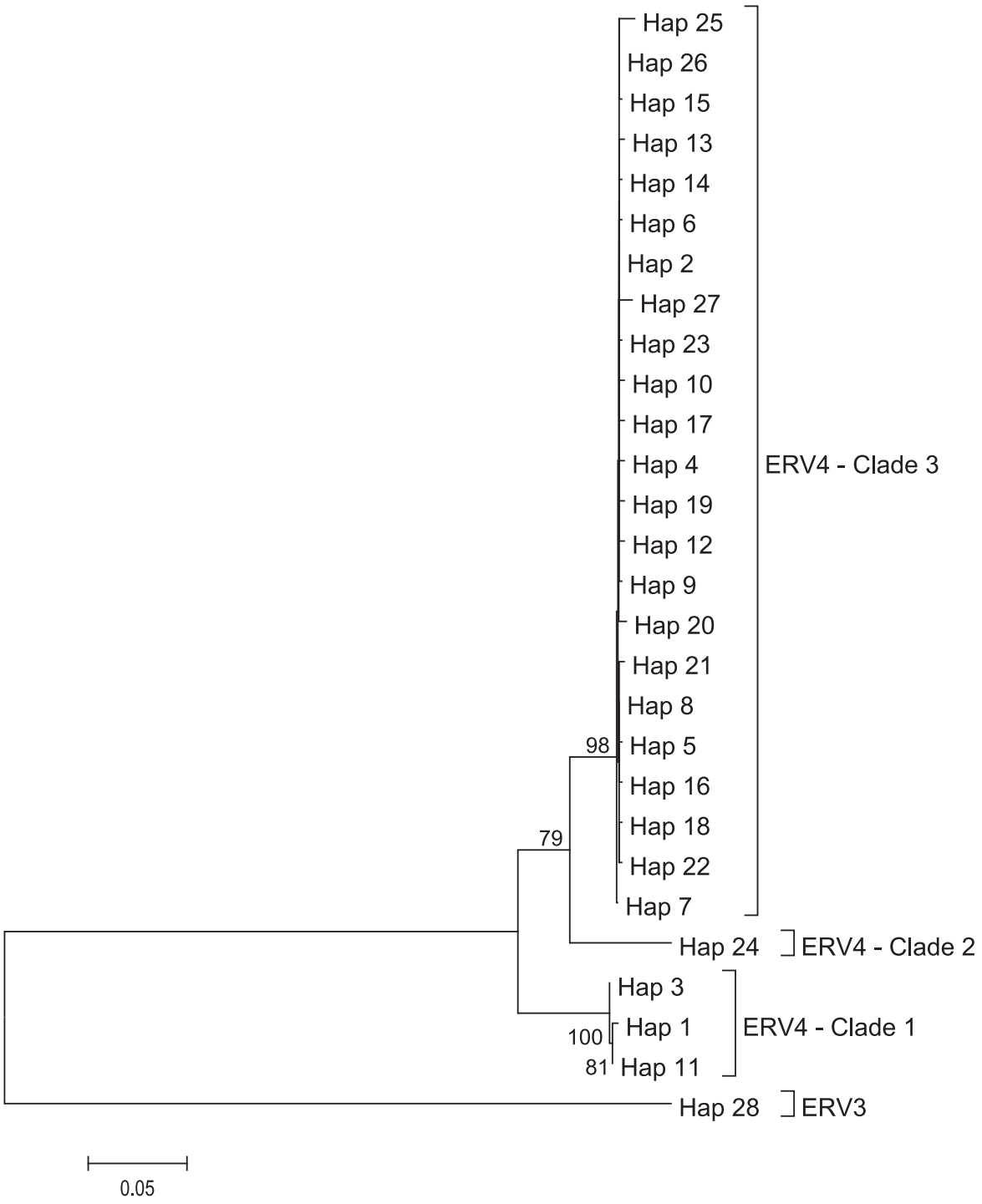


Figure S3.1: *C. johnstoni pro-pol* haplotypes clustered into four major clades belonging to ERV3 and ERV4. The Neighbour Joining phylogeny was derived from 28 nucleotide sequences, and generated using the Jukes-Cantor correction for multiple substitutions and 1000 bootstrap replicates. Numbers beside the branches indicate bootstrap support values greater than 50%. Scale bar on bottom left indicates branch length.

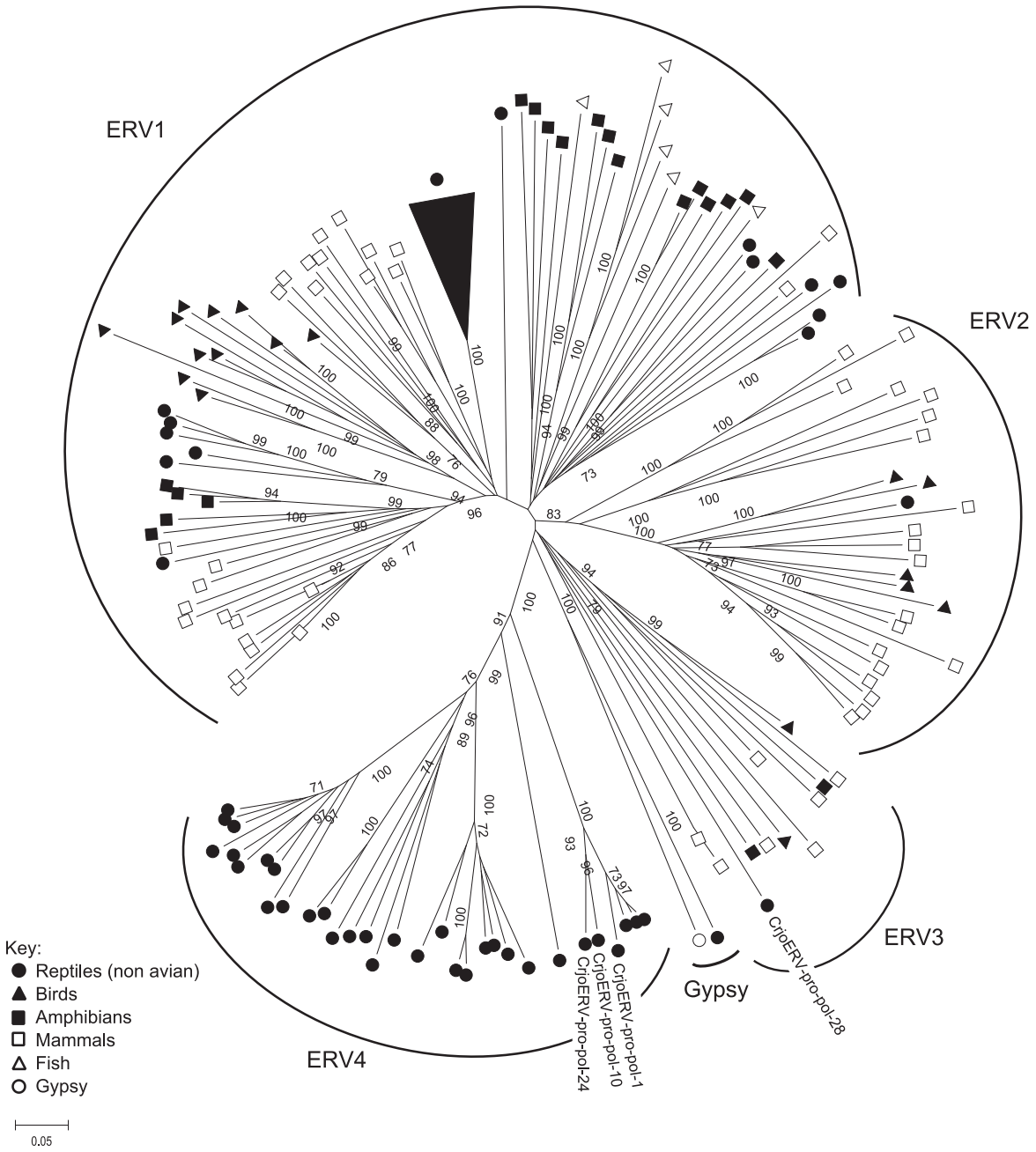


Figure S3.2: *C. johnstoni* ERV lineages cluster with crocodilian ERV4 sequences and within ERV3. The Neighbour Joining phylogeny was generated from the putative amino acid translations of the *pol* domain and ERVs from other species. Sequence names indicate the representative sequences from the four *C. johnstoni* lineages. Sequence IDs for published and previously described sequences have been omitted for clarity. The scale bars indicate branch lengths and values on the branches indicate bootstrap support greater than 70%.

Appendix III: Supplementary Methods

Chapter 2

PCR conditions

Forward primer sequence: GTK TTI KTI GAY ACI GGI KC

Reverse primer sequence: ATI AGI AKR TCR TCI ACR TA

PCR was carried out in duplicate, in 25 μL reaction volumes, containing 100 pmol of each primer, 2 mM $MgCl_2$, 0.16 mM dNTPs, PCR buffer and 1 U of high fidelity *Taq* polymerase. PCR cycles were as follows: initial denaturation at 94°C for 2 minutes, 35 cycles of 45°C (30 seconds), 72°C (60 seconds) and 94°C (30 seconds), followed by final annealing period of 3 minutes at 45°C and a final extension period of 10 minutes at 72°C.

RDP settings

Default program settings were used, implementing the RDP (Martin and Rybicki, 2000), GENECONV (Padidam et al., 1999), Bootscan (Martin et al., 2005), MaxChi (Smith, 1992), Chimaera (Posada and Crandall, 2001), SiScan (Gibbs et al., 2000) and 3seq methods (Boni et al., 2007) for detection of recombinants with a significance cut off of $P = 0.05$ and the Bonferroni correction. Tests were carried out on sequences within *C. porosus* and across species. Sequences were considered to be potential recombinants if two or more of the above methods returned a significant value.

Selection criteria for representative sequences

Due to the large number of sequences recovered, representative sequences from *C. porosus* selected based on similarity to a consensus sequence from each clade. Since we are also interested in the functionality of sequences, sequences with fewer perceived indels or stop codons were selected over those with more of these mutations where there were multiple equally similar sequences. Sequences from other crocodilians were treated similarly except in cases where there were clearly two divergent lineages represented.

PAML model comparisons

Tests for selection on specific sites were conducted under the assumption of a single rate of substitution across branches. The models M0 and M3 were compared to determine if selection differed between sites, and the model pairs M2 and M3, and M7 and M8 were used to test for selection at each site. Likelihood ratio test (LRT) statistics were calculated for the following pairs to determine significance; M0–M3, M1a–M2a, M7–M8 (see Results for additional information).

LRT values were calculated as twice the difference between the likelihood values for each of the different models, and compared to the Chi-squared values for one degree of freedom.

Preparation of the Macroarrays

Stamped membranes were placed on LB (Luria Bertani) agar containing 12.5 mg/L Chloramphenicol and incubated overnight at 37°C. To fix the BAC DNA to the membranes, cells were first lysed by incubation in an 0.5 N NaOH/1.5 M NaCl solution for 7 mins, followed by incubation in 1.5 M NaOH/0.5 M Tris Cl solution for a further 7 mins. Membranes were then allowed to air dry for 1 hour before DNA was fixed using a 0.4 N NaOH solution for 20 mins and washed in 5× SSPE solution for 7 mins. Membranes were then left to air dry for 24 hours before hybridisation.

Hybridisation conditions

Arrays were soaked in fresh hybridisation solution at 65°C overnight for pre-hybridisation treatment. This solution was then drained and fresh hybridisation solution added. Membranes were incubated for 1 hour before denatured probe solutions were added. Hybridisation of the ERV1 specific probe was carried out at 65°C for 2 days before being washed. Macroarrays hybridised with the pooled probe set were treated similarly, although hybridisation time was extended to three days.

Membranes were then rinsed with 1× SSC/0.1% SDS wash solution before being washed in fresh, preheated wash solution at 65°C for 1 hour. Washed arrays were wrapped in plastic and exposed to phosphoimager screens (GE Healthcare) before being scanned using the Amersham Storm 820 system. The membranes hybridised with the ERV1 *Gammaretrovirus*-like specific probe were exposed for two days before images were scanned. For the pooled probe set, exposure was reduced to an overnight exposure. Scanned images were rotated and trimmed to remove the membrane edges prior to identification of positive clones.

Appendix IV: ImageJ macro for densitometric analysis

```
run("Despeckle");
run("Subtract Background...", "rolling=15 light");
run("Unsharp Mask...", "radius=2 mask=0.60");
run("Make Binary");
run("Despeckle");
run("Fill Holes");
run("Erode");
run("Watershed");
run("Ultimate Points");
run("Find Maxima...", "noise=0 output=[Point Selection]
exclude light");
run("Revert");
run("Measure");
```

See also attached CD for electronic copy.

Associated file:

MacroarrayDensitometry.ijm

Appendix V: ERV detection pipeline

Please see attached CD.

Associated files:

ERV_detection_pipeline.py (python script)

ERV_detection_pipeline.conf (configuration file)

Usage:

```
ERV_detection_pipeline.py –i <input fasta file> -o
<output directory> -c <configuration file> -<run mode> -n
<integer>
```

For further details use:

```
ERV_detection_pipeline.py --help
```

Appendix VI: Published Material

Please see attached CD.

Associated files:

1759-8753-3-20.pdf (published document)

1759-8753-3-20-s1.pdf (additional file 1)

1759-8753-3-20-s2.pdf (additional file 2)

Appendix VII: Submitted Material

Please see attached CD

Associated file:

130619_JHerp_Cjohnstoni.pdf (pdf version of the manuscript submitted for review)