Retroviruses in the common brush-tailed possum (*Trichosurus vulpecula*)

A thesis submitted in partial fulfilment of the requirements for the degree of

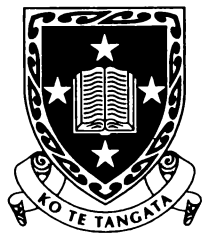
Doctor of Philosophy in Biological Sciences

at

The University of Waikato

by

Gregory John Baillie



The University of Waikato
Te Whare Wānanga o Waikato

1999

Abstract

This study details the search for, and characterisation of, retroviruses in a marsupial, the common brush-tailed possum (*Trichosurus vulpecula*). Initial efforts were directed at detecting exogenous retroviruses in possums, but the majority of the work described in this thesis involved the isolation and characterisation of a possum endogenous retrovirus.

Endogenous retroviruses were detected in possum genomic DNA by PCR amplification using degenerate primers derived from the retroviral *pol* gene. Cloning, sequencing, and analysis of these PCR products revealed the presence of several families of endogenous retroviruses in the possum genome.

Reverse transcriptase activity was detected in the blood of all possums tested using the product enhanced reverse transcriptase (PERT) assay. RT-PCR was performed on RNA isolated from possum blood plasma using the *pol*-derived degenerate primers. Cloning and sequencing of the products indicated that a homogeneous retroviral RNA species was present in the blood of possums, and that this RNA was related to, but not identical to, the endogenous possum retroviruses already detected.

A 3'-RACE approach was used to amplify the majority of the 3' end of the possum retroviral RNA. The subsequent discovery that this RNA species was derived from an endogenous retrovirus in the possum genome allowed amplification of the remainder of the genome by a combination of PCR, single primer PCR, and RT-PCR. The sequences were assembled into a contiguous sequence, the TvERV-K1 contig. In addition, a near-full-length TvERV-K1-related fragment, was amplified from possum genomic DNA, cloned, and sequenced. It was named pTvERV-K2. The TvERV-K elements are the first full-length marsupial retrovirus sequences to be reported.

Analysis of the sequences of the TvERV-K1 contig and pTvERV-K2 revealed most of the regulatory regions required for replication of a retroviral genome, as well as uninterrupted, or minimally interrupted, open reading frames

(ORFs) for the *gag*, *pro*, and *pol* genes. Only a short region of sequence with homology to the Env proteins of other retroviruses was detected. All of the TvERV-K proteins displayed highest homology to those of the simian type D retroviruses. Likewise, phylogenetic analysis using the deduced amino acid sequences of the Pro and Pol proteins, placed the possum endogenous retrovirus with the exogenous and endogenous type D retroviruses of Old World and New World monkeys. Thus, a cross-species transmission event — from marsupials to primates, from primates to marsupials, or from an unidentified source to both marsupials and primates — appears to have occurred.

There are 15–20 copies of the TvERV-K element in the possum genome, as determined by Southern hybridisation. However, integration sites appeared to be variable between possums, suggesting recent (or ongoing) retrotranspositional activity. Both PCR and Southern hybridisation analyses suggest that TvERV-K elements shorter than those from which the TvERV-K1 contig and pTvERV-K2 were derived are present in the possum genome.

The implications of these findings are discussed.

Acknowledgements

Thank you Dick Wilkins, for conceiving this project, for obtaining the funding for it, and for imparting wisdom at the most crucial times.

Thank you Cheryl O'Connor, Tim Day, Lynette Hartley, and Tony Painting for enabling and helping me to obtain possum tissues.

Thank you Raewyn Towers, for proofreading this thesis, and Chrissen Gemmill, for giving me advice regarding phylogenetic analysis and for the use of your phylogenetic software and hardware.

Thank you the people at MAF Policy, for providing the funding for this project, and for accommodating my whims.

Finally, thank you to all those people — my family, friends, flatmates, and fellow staff and students at the University of Waikato and AgResearch Ruakura — for your help and support.

Contents

List of Figures

List of Abbreviations

°C	degrees Celsius
"	inch(es)
µFD	microFarad
µg	microgram
µl	microlitre(s)
µm	micrometre(s)
µM	micromolar
A	absorbance
A_{260}	absorbance at 260nm
A_{280}	absorbance at 280nm
A_{600}	absorbance at 600nm
Amp	ampicillin
bis-acrylamide	N,N'-methylene-bis-acrylamide
bp	base pairs
BSA	bovine serum albumin
CA	capsid
cDNA	complementary DNA
Ci	Curie(s)
dATP	deoxyriboadenosine triphosphate
dCTP	deoxyribocytidine triphosphate
ddATP	dideoxyriboadenosine triphosphate
ddCTP	dideoxyribocytidine triphosphate
ddGTP	dideoxyriboguanidine triphosphate
ddTTP	dideoxyribothymidine triphosphate
dGTP	deoxyriboguanidine triphosphate
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxyribonucleotide triphosphates
dT	deoxyribothymidine
DTT	dithiothreitol
dTTP	deoxyribothymidine triphosphate
DU/dUTPase	deoxyuridine triphosphatase
EDTA	ethylenediaminetetraacetic acid
ERV	endogenous retrovirus
g	force of gravity
g_{avg}	average force of gravity
g	gram(s)
GTE	glucose, Tris, EDTA buffer
HERV	human endogenous retrovirus
hr	hour(s)
IN	integrase
IPTG	isopropyl-β-*D*-thiogalactopyranoside
kb	kilo base pairs
kV	kilovolt(s)
l	litre(s)
LB	Luria-Bertani broth
LINE	long interspersed nuclear element
LTR	long terminal repeat

M	molar
MA	matrix
mg	milligram(s)
MHR	major homology region
min	minute(s)
ml	millilitre(s)
mm	millimetre(s)
mM	millimolar
mmol	millimoles
MSD	membrane spanning domain
MYBP	million years before present
NC	nucleocapsid
ng	nanogram(s)
NJ	neighbour joining
nm	nanometre(s)
nM	nanomolar
nt	nucleotide(s)
PBS	primer binding site
PCR	polymerase chain reaction
PEG	polyethylene glycol
PEG8000	polyethylene glycol, molecular weight 8000
PERT assay	product enhanced reverse transcriptase assay
pfu	plaque forming units
pg	picogram(s)
pmol	picomoles
PPT	polypurine tract
PR	protease
rA	riboadenosine
RACE	rapid amplification of cDNA ends
RE	restriction enzyme
RER	rough endoplasmic reticulum
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RT	reverse transcriptase
RT-PCR	reverse transcription polymerase chain reaction
SA	splice acceptor
SD	splice donor
SDS	sodium dodecyl sulphate
SINE	short interspersed nuclear element
SSC	sodium chloride, sodium citrate buffer
ssDNA	single-stranded DNA
STET	sodium chloride, Tris, EDTA, Triton buffer
SU	surface protein
TAE	Tris, acetate, EDTA buffer
TBE	Tris, borate, EDTA buffer
TE	Tris, EDTA buffer
TEMED	N,N,N',N'-tetramethylethylenediamine
Tet	tetracycline
TM	transmembrane protein
TMAC	tetramethyl ammonium chloride
tRNA	transfer RNA

TSS	transformation and storage solution
U	unit(s)
UV	ultraviolet
v	volume
W	watt(s)
w	weight
X-gal	5-bromo-4-chloro-3-indolyl-β-*D*-thiogalactopyranoside

Chapter 1

General Introduction and Literature Review

1.1 Introduction

Infectious retroviruses have been isolated and characterised from a wide range of birds and mammals, and in most cases they are associated with a disease condition in their host. In addition, the genomes of all eukaryotes contain endogenous retroviruses or sequences that resemble proviruses.

Retrovirus particles have been observed in marsupial tissues, and sequences from some endogenous marsupial retroviruses have been reported. However, no pathogenic marsupial retroviruses have been isolated, no sequences from any exogenous marsupial retroviruses have been reported, and none of the endogenous marsupial retrovirus sequences reported so far are full-length. This investigation is concerned with infectious retroviruses and retrovirus-related sequences in the genome of a marsupial, the common brush-tailed possum (*Trichosurus vulpecula*).

1.2 Biology of infectious retroviruses

The biology of retroviruses has been reviewed extensively, and is only briefly summarised here. For more thorough descriptions of the various aspects of retrovirus biology, the reader is referred to the review by Varmus and Brown (1989) and the comprehensive text by Coffin *et al.* (1997a).

1.2.1 The retrovirus particle

Retroviruses are enveloped RNA viruses with diameters of 80–120nm and densities of approximately 1.16 g/ml. They are composed of about 1–2% RNA, 35% lipids, and 65% proteins (Vogt, 1997b).

A generalised retrovirus particle is shown in FIG. 1.1. At the centre of the retrovirus, two copies of the RNA genome of the retrovirus are associated with/packaged by nucleocapsid (NC) proteins. The RNA/nucleocapsid is surrounded by capsid (CA) proteins, which form the outer layer of the capsid or core. Also within the core are tRNA molecules and the retroviral enzymes integrase (IN) and reverse transcriptase (RT). Protease (PR) molecules,

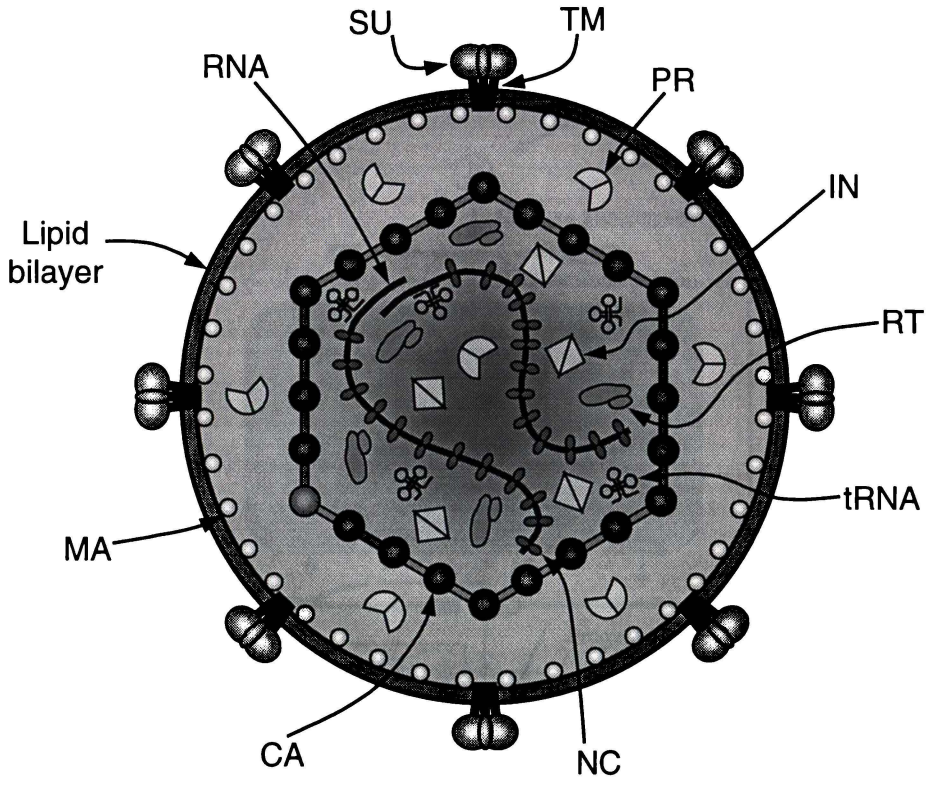


FIGURE 1.1 Mature retrovirus particle. See text for details. Adapted from Vogt (1997a).

which are responsible for proteolytic maturation of the particle after release from the host cell, are shown both inside and outside the core. A layer of matrix (MA) proteins is associated with the inside of the virus membrane. The membrane forms the outer layer of the retrovirus particle, and has the envelope proteins, which comprise transmembrane (TM) and surface (SU) domains, embedded within it (Vogt, 1997b).

1.2.2 The retroviral life cycle

The life cycle of an infectious retrovirus is shown diagrammatically in FIG. 1.2. Infection of a cell by a retrovirus begins when the SU domain of the Env protein on the virus membrane binds to a receptor protein at the surface of the cell. Following fusion of the virus and cell membranes, which is

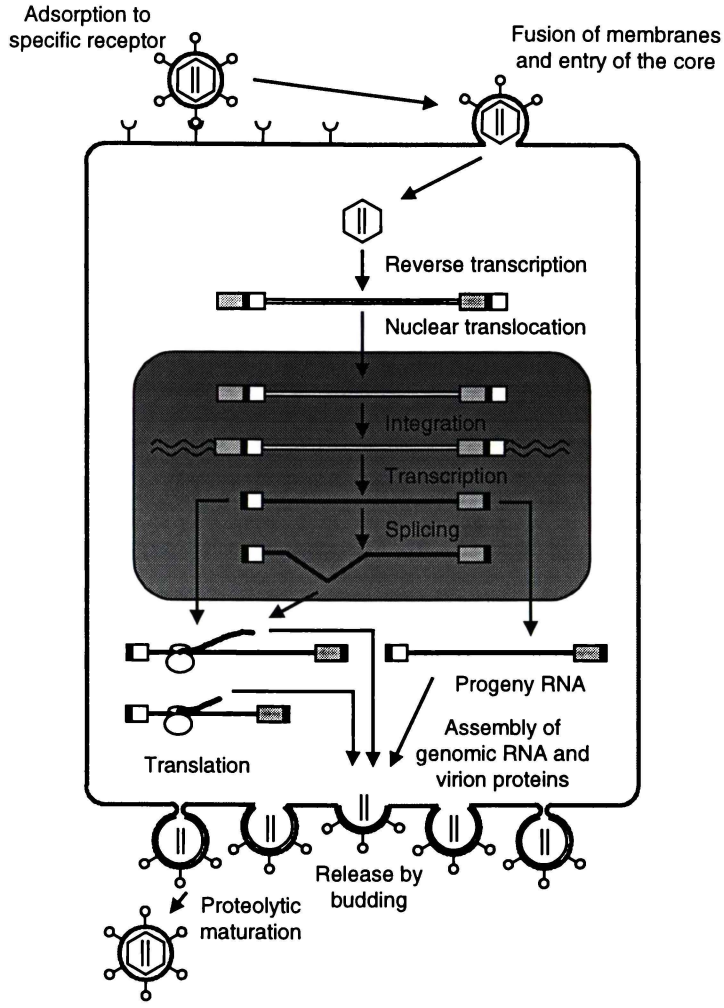


FIGURE 1.2 Retrovirus life cycle. See text for details. Adapted from Vogt (1997a).

mediated by the TM domain of Env, the retroviral core enters the cytoplasm of the cell (Hunter, 1997). The RNA genome of the retrovirus is copied into double stranded DNA by the retroviral enzyme reverse transcriptase (Telesnitsky and Goff, 1997). The DNA genome is transported to the nucleus where it is integrated into the genome of the host cell by the retroviral enzyme integrase (Kulkosky and Skalka, 1994; Brown, 1997). Many retroviruses require disassembly of the nuclear membrane during mitosis for access to genomic DNA, although HIV-1 (and perhaps other lentiviruses) are able to enter the nuclei of nondividing cells (Brown, 1997). The integrated DNA form of the retroviral genome is known as the provirus. Transcription of the provirus by host cell machinery generates full-length retroviral RNA, which resembles normal messenger RNA (mRNA) (Rabson

and Graves, 1997). A proportion of the RNA is spliced before being exported to the cytoplasm, while the remainder is exported in its unspliced form (Rabson and Graves, 1997). In the cytoplasm, some of the full-length mRNA is used as template for the synthesis of Gag, Gag-Pro, and/or Gag-Pro-Pol polyproteins (Swanstrom and Wills, 1997). The spliced mRNA is translated at the RER into the envelope (Env) polyprotein, which is transported via the Golgi apparatus to the cell surface (Hunter and Swanstrom, 1990; Swanstrom and Wills, 1997). The retroviral proteins and unspliced retroviral mRNAs assemble into immature retrovirus particles that bud from the surface of the cell, simultaneously acquiring a membrane containing Env proteins (Berkowitz *et al.*, 1996; Einfeld, 1996; Swanstrom and Wills, 1997). Following release from the cell, the immature particle undergoes maturation — retroviral protease-mediated cleavage of the Gag, Gag-Pro, and Gag-Pro-Pol polyproteins into their separate proteins — which renders it capable of infecting a new cell (Vogt, 1996; Swanstrom and Wills, 1997).

1.2.3 Retrovirus genomes and proteins

1.2.3.1 The minimum requirements

The generalised proviral DNA, genomic RNA, and spliced sub-genomic RNA of retroviruses are shown diagrammatically in FIG. 1.3. Retrovirus genomes comprise coding regions, which encode the retroviral proteins, and regulatory regions, which include sequences that are required during various stages of the retroviral life cycle. Coding and regulatory regions may overlap.

The DNA and RNA genomes are colinear, except that the proviral genome possesses long terminal repeats (LTRs) which are generated during reverse transcription of the RNA genome (Telesnitsky and Goff, 1997). The genomic RNA resembles cellular mRNA. It possesses a 5' CAP (methylated GDP attached through a 5'–5' link to the 5' terminal nucleotide) and a 3' polyadenylated tail (50–200 A ribonucleotides) (Vogt, 1997b).

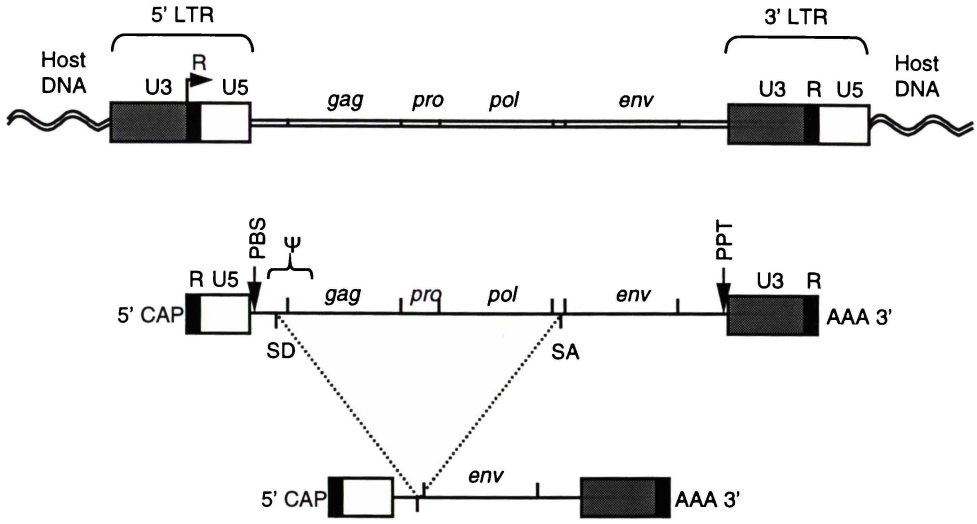


FIGURE 1.3 Generalised retroviral genome. The proviral genome (top) is transcribed into full-length mRNA (middle), a subset of which is spliced (bottom). ⌜, transcription start; CAP, 5' 7-methyl guanine CAP; AAA, poly(A) tail; SD, splice donor; SA, splice acceptor; see text for details of other abbreviations. Adapted from Vogt (1997b).

The coding regions of all replication-competent retroviruses contain *gag*, *pro*, *pol*, and *env* genes, which encode the polyprotein products Gag, Pro, Pol, and Env, respectively. Although the order of the *gag*, *pro*, *pol*, and *env* ORFs is the same, their organisation and expression differs between retroviruses (FIG. 1.4). The Pro and Pol proteins of most retroviruses are never synthesised separately — they are always translated from full-length mRNAs as Gag-Pro and/or Gag-Pro-Pol polyproteins by suppression of the termination codon of the *gag* and/or *pro* genes. The exceptions to this rule are the spumaviruses (HFV in FIG. 1.4), which express a Pro-Pol polyprotein from a spliced mRNA. In the mammalian type C retroviruses (MLV in FIG. 1.4), the *gag*, *pro* and *pol* ORFs are all in the same reading frame, but the *gag* and *pro* ORFs are separated by a stop codon which is occasionally circumvented by readthrough (termination) suppression. In the avian type C retroviruses (ALV in FIG. 1.4), the *gag* and *pro* ORFs are in the same reading frame and are not separated by a stop codon. Translation of the *pol* ORF, which is in a different reading frame from the *gag-pro* ORF, requires ribosomal frameshifting within the overlapping region of the *pro* and *pol* ORFs. A single frameshift between the *gag* and *pro* ORFs is required for expression of Gag-Pro-Pol polyproteins in the lentiviruses (HIV-1 in FIG. 1.4).

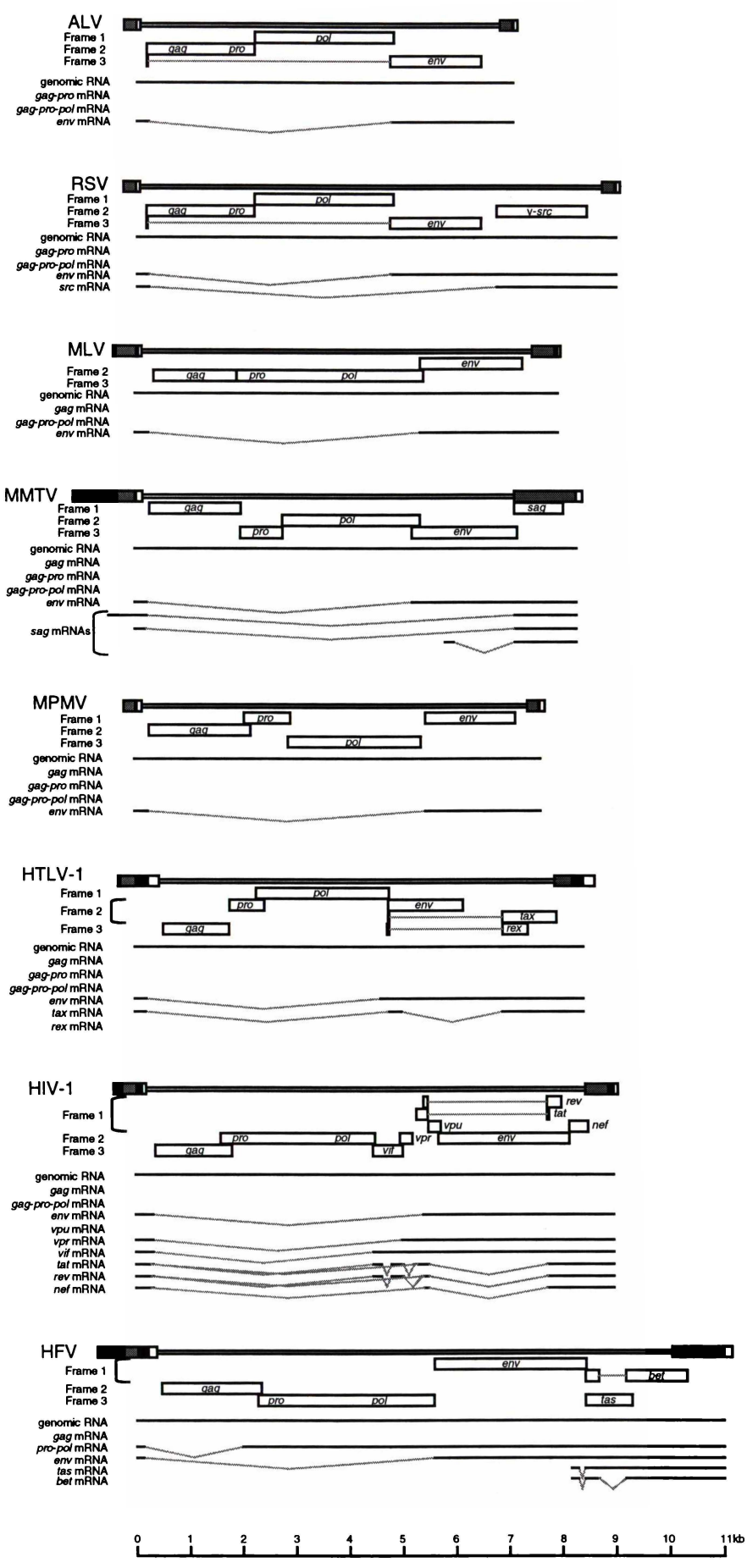


FIGURE 1.4 *see page 8 for legend.*

The type B (MMTV), type D (MPMV), and HTLV/BLV group (HTLV-1) retroviruses all require two frameshift events — one between the *gag* and *pro* ORFs, and a second between the *pro* and *pol* ORFs — for expression of Gag-Pro-Pol polyproteins (FIG. 1.4). Regardless of the mechanism of expression, retrovirus particles typically contain 10–20 times more Gag polyproteins than Gag-Pro-Pol polyproteins (Swanstrom and Wills, 1997; Vogt, 1997b). The Env polyprotein is translated from a spliced *env* mRNA for all retroviruses (Hunter and Swanstrom, 1990; Swanstrom and Wills, 1997; Vogt, 1997b).

Each of the proteins in the mature retrovirus particle (see 1.2.1 The retrovirus particle) is derived from one of the viral polyproteins (FIG. 1.5). Hence, MA, CA, and NC are derived from the Gag domain of Gag, Gag-Pro and/or Gag-Pro-Pol polyproteins, PR from the Pro domain of Gag-Pro and/or Gag-Pro-Pol, and RT and IN from the Pol domain of the Gag-Pro-Pol polyprotein (Swanstrom and Wills, 1997; Vogt, 1997b). SU and TM are generated from the Env polyprotein (Hunter and Swanstrom, 1990; Swanstrom and Wills, 1997; Vogt, 1997b). Cleavage of the Gag, Gag-Pro, and Gag-Pro-Pol polyproteins into the separate components is performed by the retroviral PR during maturation of the virus particle (see 1.2.2 The retroviral life cycle). In contrast, the Env polyprotein is cleaved by cellular proteases in the Golgi apparatus, and the resulting SU and TM components remain bound during transport to the cell surface (Hunter and Swanstrom, 1990; Swanstrom and Wills, 1997; Vogt, 1997b).

Regulatory regions of the genome are required at various stages of the retrovirus life cycle. The long terminal repeats (LTRs) of the provirus comprise U3, R, and U5 regions, so-named because of their positions in the

---

FIGURE 1.4 (*see page 7*) Organisation of the open reading frames (ORFs) and RNAs of retroviruses. The proviral DNAs of representatives of the seven recognised retroviral genera are shown, with the corresponding ORFs (represented by boxes) and the RNAs (represented by lines) used to express those ORFs shown below each provirus. Grey lines within ORFs and RNAs represent regions removed by RNA splicing. See text for further details. Full names of retroviruses are listed in Appendix 1. Adapted from Coffin (1992), Vogt (1997b), and Petropoulos (1997).

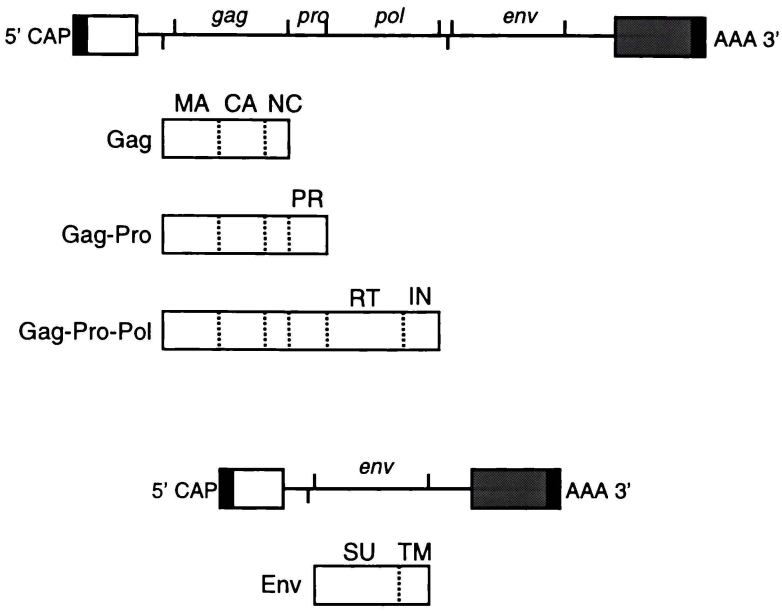


FIGURE 1.5 Retroviral polyproteins. Gag, Gag-Pro, and Gag-Pro-Pol proteins are expressed from full-length transcripts, and are cleaved by the retroviral PR during particle maturation. Env polyproteins are expressed from spliced transcripts, and are cleaved by cellular proteases. Polyproteins are indicated by boxes, with dashed vertical lines representing cleavage sites. See text for further details.

genomic RNA. The R region is repeated at both ends of the genomic RNA, U5 is unique to the 5' end of the RNA genome, and U3 is unique to the 3' end of the RNA genome (FIG. 1.3). U3 contains most of the transcriptional control elements of the provirus, including the promoter and multiple enhancer sequences (Vogt, 1997b). Also within the LTR, in U3 or R, is the polyadenylation signal, which directs host cell machinery to cleave and polyadenylate retroviral RNAs (Guntaka, 1993; Rabson and Graves, 1997). Splice donor (SD) and splice acceptor (SA) sites direct splicing of viral RNAs (Rabson and Graves, 1997), and frameshift or termination suppression signals allow the synthesis of Gag-Pro and/or Gag-Pro-Pol polyproteins (Levin *et al.*, 1993; Swanstrom and Wills, 1997). Packaging signals (Ψ), also known as encapsidation signals (E), discriminate genomic retroviral RNAs from spliced retroviral RNAs and cellular RNAs for packaging into particles (Berkowitz *et al.*, 1996; Swanstrom and Wills, 1997). The primer binding site (PBS) and polypurine tract (PPT) are required for the priming of minus-strand and plus-strand DNA synthesis, respectively, during reverse

transcription (Telesnitsky and Goff, 1997). Finally, inverted repeats at the 5' end of U3 and the 3' end of U5 function during processing of viral DNA ends by IN and integration of viral DNA into the genome of the host cell (Kulkosky and Skalka, 1994; Brown, 1997).

1.2.3.2 Simple, complex, and acutely transforming retroviruses

Retroviruses can be defined as simple, complex, or acutely transforming, depending on the proteins they encode, and the manner of expression of the proteins.

Simple retroviruses possess all of the regulatory regions, and encode all of the proteins (Gag, Pro, Pol, and Env), required for replication (see below). Although simple retroviruses may also encode proteins in addition to Gag, Pro, Pol, and Env, these proteins do not directly affect retrovirus replication and their expression is from singly-spliced mRNA, properties that distinguish them from the accessory proteins encoded by complex retroviruses (see below). The avian type C (ALV in FIG. 1.4), mammalian type C (MLV in FIG. 1.4), type B (MMTV in FIG. 1.4), and type D (MPMV in FIG. 1.4) retroviruses (see 1.2.4 Retrovirus classification) are all simple retroviruses.

The genomes of complex retroviruses resemble those of the simple retroviruses in that they possess all of the regulatory regions required for replication of the retroviral genome, and encode the Gag, Pro, Pol, and Env polyproteins. However, complex retroviruses also possess genes, known as accessory genes, in addition to *gag*, *pro*, *pol*, and *env*. Expression of the accessory genes often requires alternative or multiple splicing of transcripts, and the gene products affect transcription, processing, and translation of retroviral RNA, as well as virus particle assembly (Cullen, 1992; Coffin *et al.*, 1997b; Vogt, 1997b). Three of the currently recognised genera of retroviruses (see 1.2.4 Retrovirus classification) are classified as complex retroviruses: lentiviruses (HIV-1 in FIG. 1.4); spumaviruses (HFV in FIG. 1.4); and viruses of the HTLV/BLV group (HTLV-1 in FIG. 1.4). The fish retroviruses, yet to be officially granted the status of genus, are also classified as complex retroviruses.

Some members of the avian type C and mammalian type C genera (see 1.2.4 Retrovirus classification) carry oncogenes, and are known as acutely transforming retroviruses because of their ability to rapidly induce malignancies in their hosts (Rosenberg and Jolicoeur, 1997). The oncogenes are often included at the expense of one or more of the essential retroviral genes (although RSV, as shown in FIG. 1.4, possesses an oncogene in addition to the full complement of retroviral genes), and render the retroviruses incapable of replicating by themselves. However, acutely transforming retroviruses do possess the *cis*-acting regulatory elements required for replication of their genomes, and can replicate if they are complemented (see 1.2.8 Complementation and pseudotypes) by other retroviruses, known as a helper viruses (Vogt, 1997b). The majority of known acutely transforming retroviruses were isolated from the tumours in which they arose, and because of their low frequency of horizontal transfer and the normally lethal consequences of their infection, they do not contribute significantly to the incidence of retrovirus-induced tumours (Rosenberg and Jolicoeur, 1997).

1.2.4 Retrovirus classification

The systems used to classify retroviruses have changed as the degree to which retroviruses have been characterised has increased. Initially, retroviruses were classified in terms of the host from which they were isolated and the disease they caused in that host. The application of electron microscopic techniques to retrovirology resulted in a morphological classification system based on several criteria: the site of assembly of the virus particle within the host cell; the shape, size, and frequency of glycoprotein surface projections in the virus particle; and the shape and position of cores in mature particles (Bernhard, 1960; Teich, 1982; Coffin, 1992a; Nermut and Hockley, 1996). The advent of molecular biological techniques, however, has enabled a classification system based on the genome organisation of retroviruses and phylogenetic analysis of nucleotide and amino acid sequences (particularly those of the *pol* gene). Some of the terminology of the earlier systems has been incorporated into the current system.

The current predominant classification system is based on the nucleotide and amino acid sequences (mainly of reverse transcriptase) and the genetic organisation (presence and nature of additional viral genes [accessory genes], *gag-pro-pol* relationship, sizes of various regions and proteins, type of tRNA primer) of retroviruses. Seven retroviral genera are currently recognised: mammalian type C viruses (prototype MLV); avian type C viruses (the ASLV, prototype RSV); type B viruses (prototype MMTV); type D viruses (prototype MPMV); lentiviruses (prototype HIV-1); viruses of the HTLV/BLV group (prototype HTLV-1); and spumaviruses (prototype HFV) (Coffin, 1992a) (FIG. 1.6). The retroviruses of flies and fish have not been officially classified, but they are likely to constitute new genera (Vogt, 1997b). There is virtually no detectable sequence identity between distantly related retroviruses at the nucleotide level (Coffin, 1992a). However, amino acid sequences of conserved proteins (mainly reverse transcriptase) show identity at approximately one third to two thirds of sites between retroviruses of different genera, and greater than two thirds between retroviruses within a genus (Vogt, 1997b).

1.2.5 Distribution in animals

Retroviruses were first isolated from chickens suffering from either erythroleukemia (Ellerman and Bang, 1908, historic reference) or fibrosarcomas (Rous, 1911, historic reference). The first mammalian retrovirus to be characterised was mouse mammary tumour virus (MMTV) (Bittner, 1936). The majority of the known infectious retroviruses are from birds and mammals, and most were discovered during investigations of the disease they induced in their host. Retrovirus-induced diseases include sarcomas, a wide range of hematopoietic tumours, carcinomas, neurological diseases, immunodeficiencies, anaemia, arthritis, wasting syndromes, and osteopetrosis (Rosenberg and Jolicoeur, 1997).

Retroviruses are found in animals other than the placental mammals and birds. Disease-associated retroviruses have been observed in reptiles, fish, and molluscs (reviewed in Poulet *et al.*, 1994). Retroviral particles have been reported in amphibians (Masahito *et al.*, 1995), and endogenous

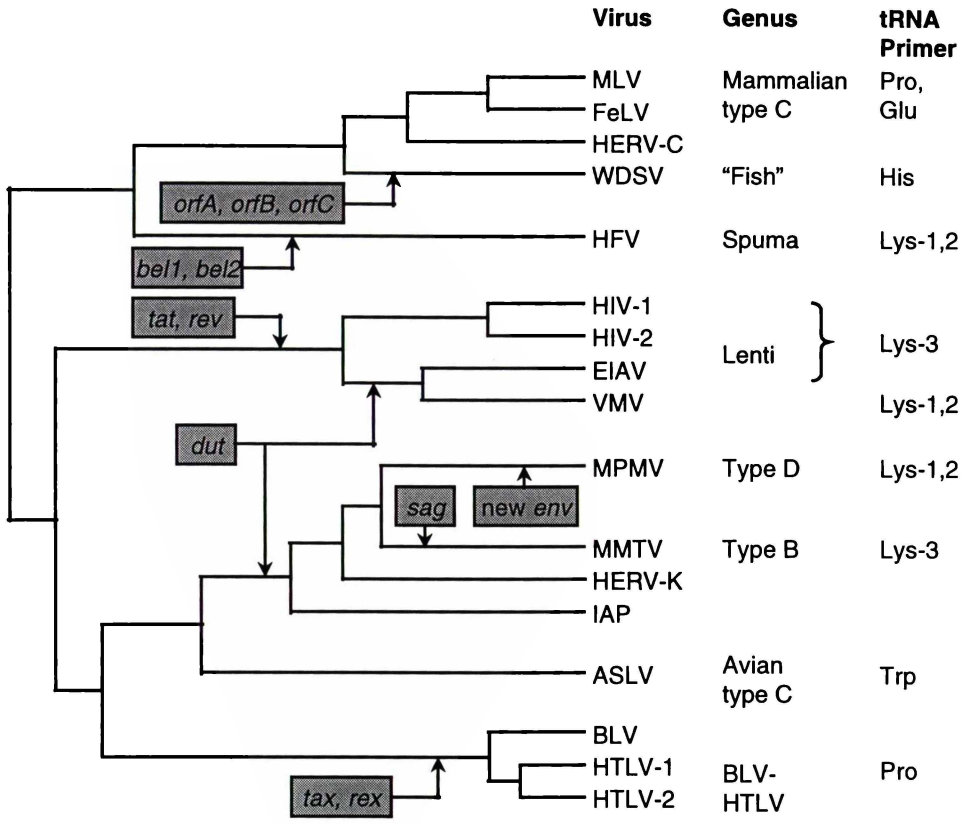


FIGURE 1.6 The retroviral genera. The tree is constructed from a comparison of the *pol* gene sequences of the retroviruses. Other events in the evolution of the retroviruses (acquisition of new/additional genes) are superimposed on the tree. The retroviral genera, and the tRNA used to prime minus-strand synthesis are shown to the right. Full names of retroviruses are listed in Appendix 1. Adapted from Vogt (1997b).

retroviral sequences are clearly present in amphibian genomes (Tristem *et al.*, 1996; Herniou *et al.*, 1998). Some members of the Ty3-*gypsy* family of retrotransposons in *Drosophila* have recently been shown to possess *env*-like genes, and are capable of infectious replication (reviewed in Boeke and Stoye, 1997).

A given animal species can host retroviruses from several genera. Humans, for example, are susceptible to infection by a spumavirus (HFV), HTLV-group viruses (HTLV-1, HTLV-2), and lentiviruses (HIV-1, HIV-2). Non-human primates host spumaviruses (SFV), lentiviruses (SIV), HTLV/BLV-type retroviruses (STLVs), mammalian type C viruses (BAEV, GaLV), and type D retroviruses (MPMV, SRV-1, SRV-2). Lentiviruses (VMV) and type B/D (JSRV) retroviruses have been isolated from sheep. Mice support the

replication of MLV-type (eg. MoMLV) and B-type (MMTV) retroviruses. An MLV-type retrovirus (REV-A) has also been found in chickens, as have avian type C retroviruses (eg. RSV) (Petropoulos, 1997).

1.2.6 Cross-species transmission

Cross-species transmission is the transmission of a retrovirus that arose within one animal species, or has been associated with that species for a significant period of time, to another animal species. Cross-species transmission appears to be a regular, and ongoing, feature of retrovirus biology. The most studied cross-species transmission events are those from primates to humans. Thus, it is now apparent that HIV-1 was transmitted to humans from chimpanzees (Gao *et al.*, 1999), that HIV-2 was transmitted to humans from sooty mangabeys (Gao *et al.*, 1992), and that the human T-cell lymphotrophic viruses (HTLVs) also had primate origins (Goubau *et al.*, 1996). Recently, transmission of simian foamy viruses (SFVs) from primates to humans has also been observed (Heneine *et al.*, 1998). In addition, the observation that the reticuloendotheliosis viruses (REVs) in birds are more closely related to the mammalian type C retroviruses than they are to the avian type C retroviruses (McClure *et al.*, 1988), suggests that cross-species transmission can occur between hosts that are more distantly related than humans are to the other primates (Doolittle *et al.*, 1989).

1.2.7 Retrovirus mutation and variation

Infectious retroviruses, like other RNA viruses, accumulate mutations at high rates (Katz and Skalka, 1990; Telesnitsky and Goff, 1997). Various processes during retrovirus replication contribute to this high mutation rate.

Retroviruses are transcribed by cellular RNA polymerase II, which lacks an editing function. Thus, errors introduced during transcription are not repaired (Telesnitsky and Goff, 1997).

Mutations can also be introduced during reverse transcription of the retroviral genome. The reverse transcriptase enzyme lacks proofreading activity

(Battula and Loeb, 1976; Roberts *et al.*, 1988), and is capable of extending mismatched nucleotides (see Pulsinelli and Temin, 1994 and references therein), which together result in high numbers of nucleotide substitution mutations. The process of reverse transcription is also likely to be responsible for many other mutations, including duplications, frameshifts, deletions, and deletions with insertions (for reviews see Katz and Skalka, 1990; Bebenek and Kunkel, 1993; Temin 1993; Telesnitsky and Goff, 1997).

Recombination is a common source of variation in retroviruses (for review see Hu *et al.*, 1993). Recombination can occur between two distinct retroviral RNAs that have been copackaged into a single virion (for review see Linial and Blair, 1982), and can generate a recombinant virus with altered biological properties. Two models for homologous recombination have been proposed. The forced copy choice (Coffin, 1979) and copy choice (Xu and Boeke, 1987) models involve transfer between plus-strand RNA templates during minus strand DNA synthesis, and appear to be the predominant mechanism of recombination. The strand-displacement assimilation model (Junghans *et al.*, 1982), which may account for some recombination events, proposes that a plus-strand DNA is displaced by the extension of another plus-strand initiated further upstream, and that the displaced strand hybridises to a complementary region on the other minus strand DNA in the particle. Recombination during reverse transcription is a feature of most models for the mechanism of acquisition of oncogenes by retroviruses (Swain and Coffin, 1992; Telesnitsky and Goff, 1997).

Host DNA polymerases and repair enzymes, which are responsible for copying the provirus along with the genomic DNA of the host cell during cell division and may also be involved in filling gaps or completing plus-strand viral DNA synthesis, may introduce mutations (Telesnitsky and Goff, 1997). However, the error rates of cellular DNA polymerases are far lower than those of RNA polymerase II and reverse transcriptase, and they are therefore unlikely to contribute significantly to the mutation rate of retroviruses. Indeed, it appears that proviruses accumulate mutations at about the same rate as cellular genes or pseudogenes (Temin, 1989). Nevertheless, endogenous retroviruses, which have entered the germline of

their host (see 1.3 Biology of endogenous retroviruses), will gradually accumulate mutations during DNA replication over hundreds, thousands, or millions of generations.

1.2.8 Selective forces which act on retroviruses

Mutations introduced by the mechanisms described above provide the basis for the evolution of retroviruses. New variants may display significant differences in their biological activity, such as increased virus replication, reduced replication, tissue-specific replication, altered tissue tropism, ability to replicate in different host, evasion of immune system, drug resistance, or oncogenesis (Katz and Skalka, 1990; Coffin, 1993). It should be noted that retroviruses are subject to selective forces exerted by the host, and it is this host-virus interaction that determines the ultimate replicative success of the variants. Also, replicative success in the short term does not necessarily correspond with retrovirus survival in the long term (Katz and Skalka, 1990; Coffin, 1993).

1.2.9 Complementation and pseudotypes

Even if it contains a defective gene, a retrovirus can undergo replication if it is expressed in the same cell as another retrovirus that possesses an intact copy of that gene. Thus, one retrovirus might encode the Gag, Pro, and Pol proteins of a virus particle, whilst the Env proteins are encoded by a different virus. Similarly, the RNA genome of a retrovirus that is transcribed but not translated may be packaged into particles formed by the protein products of (an)other retrovirus(es). This process is known as complementation or phenotypic mixing, and the resulting virus particles as pseudotypes (Linial and Blair, 1982; Varmus and Brown, 1989). Phenotypic mixing may also occur between nondefective retroviruses (Linial and Blair, 1982). Complementation is an important feature of acutely transforming viruses (see 1.2.2.2 Simple, complex, and acutely transforming retroviruses). It is also the basis for the use of retroviruses as gene transfer vectors, whereby the retroviral vector contains the gene of interest and possesses all of the retroviral sequences required in *cis* for replication (ie. packaging signal, PBS,

LTRs, PPT), but does not encode viral proteins. The proteins required for virus particle formation are expressed in *trans* from transcripts which lack the sequences (most significantly the packaging signal) required for replication (Miller, 1997).

1.2.10 Detection of retroviruses

As mentioned above, the majority of known infectious retroviruses were initially detected as etiologic agents of disease. A number of techniques are used to characterise disease-associated retroviruses, including electron microscopy, histology, immunological studies, nucleic acid hybridisation studies, reverse transcriptase assays, and, more recently, PCR analyses. Some of these techniques can be used to search for retroviruses in the absence of clinical symptoms.

Propagation of retroviruses in cell cultures and purification of retroviruses based on their density are both invaluable tools in researching retroviruses. It is often difficult to obtain large amount of retrovirus from naturally or experimentally infected animals. Consequently, many workers have developed cell culture systems that express the large amounts of retrovirus that are required for biochemical and genetic analyses (see Teich, 1982; Weiss, 1982). Retroviruses have a characteristic density of ~1.16 g/ml in sucrose. Thus, density gradient centrifugation allows the purification of retroviruses away from other cellular components in cell-culture supernatants or tissue extracts (for example see Robinson *et al.*, 1965). Retrovirus particles can also be purified and concentrated using filtration and/or low-speed centrifugation to remove cells followed by ultracentrifugation to pellet virions (for example, see Schlom and Spiegelman, 1971). The detection of retrovirus components (using the techniques described below) in the appropriate fraction(s) of density gradient fractionated samples is normally considered strong evidence for the presence of a retrovirus.

1.2.10.1 Transformation of cells in culture

Acutely transforming retroviruses can be detected using cell culture systems. The focus-forming assay was developed by Temin and Rubin (1958), and allowed the detection and titration of ALVs in cell-cultures by observing foci of transformation. Almost all retroviruses carrying oncogenes induce transformation of fibroblastoid cells in culture, and transformation assays have been developed for many such viruses (for reviews see Weiss, 1982; Rosenberg and Jolicoeur, 1997). However, transformation assays are difficult to establish and maintain, it takes time for transformation to occur, and only a limited range of retroviruses is detectable. Transformation assays are therefore not suitable for the rapid screening of large numbers of samples, or for detection of a broad range of retroviruses.

1.2.10.2 Cytopathic effect

Some retroviruses do not induce transformation of cells but do induce cytopathic effects. Such retroviruses can be detected using cell-culture systems or by screening host tissues for abnormal histology (for reviews see Teich, 1982; Weiss, 1982). The spumaviruses, which cause foamy vacuolisation of cells in the absence of any apparent disease, were discovered in this way (Enders and Peebles, 1954; Rustigian *et al.*, 1955).

1.2.10.3 Electron microscopy

Electron microscopy can be used to detect retroviruses (Bernhard, 1960). Direct detection of virus particles in infected tissues is possible if large numbers of virions are present. Alternatively, purification or concentration of virus from tissue specimens or cell-culture supernatants may improve the chances of detection (Morgan-Capner and Pattison, 1985). Incorporation of immunological reagents into the electron microscopy technique (immunoelectron miscroscopy) can also allow partial characterisation of retroviruses (Teich, 1982) and, as previously discussed, morphology in electron micrographs formed the basis of one system of retrovirus classification (see 1.2.4 Retrovirus classification). However, electron

microscopy is an inefficient method for screening large numbers of samples for the presence of retroviruses. It is more commonly used in association with other methods to confirm retrovirus association with a disease state.

1.2.10.4 Immunological techniques

Immunological techniques are also used to detect retroviruses. Antibodies raised against retroviral antigens (normally Gag proteins) can be used to detect retrovirus antigens in tissues, cell extracts, cell-culture supernatants, and purified retrovirus preparations (Teich, 1982; Weiss, 1982). Alternatively, the presence of antiviral-antibodies in the blood of an animal can be detected. Additional characterisation of retroviruses can be achieved by combining immunological methods with other techniques, such as SDS-polyacrylamide gel electrophoresis (ie. Western blotting). However, immunological methods are limited by their specificity; they will only detect retroviral antigens (or antiviral antibodies) that are closely related to those used to generate the antibodies (or antigens). Although they are often used for diagnostic purposes (for example see Hardy, 1991), they are of limited use when screening for a broad range of retroviruses.

1.2.10.5 Reverse transcriptase assays

Reverse transcriptase is an essential component of infectious retroviral particles (see Skalka and Goff, 1993; Telesnitsky and Goff, 1997), and reverse transcriptase activity can indicate the presence of an infectious retrovirus. Reverse transcriptase activity was first detected in lysed virus suspensions using viral RNA as the template (Baltimore, 1970; Temin and Mizutani, 1970). Reverse transcriptase assays were subsequently developed which detected the ability of biological extracts to reverse transcribe retroviral RNA, non-retroviral heteropolymeric RNA templates, and synthetic homopolymeric RNA templates (Temin and Baltimore, 1972; Sarngadharan *et al.*, 1978). For many years, such reverse transcriptase assays were an invaluable tool in the detection and characterisation of retroviruses of many animals (for review see Teich, 1982), including humans (Poiesz *et al.*, 1980; Yoshida *et al.*, 1982; Barré-Sinoussi *et al.*, 1983; Gallo

et al., 1984; Levy *et al.*, 1984). However, interpretation of reverse transcriptase assay results was complicated by the early observation that some cellular DNA polymerases were able to display weak reverse transcriptase activity on some templates, especially poly(rA) templates with oligo(dT) primers (Temin and Baltimore, 1972; Sarngadharan *et al.*, 1978). They were also insensitive by today's standards, and required large-scale preparation and concentration of retroviruses (Sarngadharan *et al.*, 1978).

PCR-based RT assays have been developed recently that are approximately 10^6 times more sensitive than the early assays, and are capable of detecting as few as 10 retrovirus particles (Silver *et al.*, 1993; Pyra *et al.*, 1994; Heneine *et al.*, 1995; Maudru and Peden, 1997; Arnold *et al.*, 1998). These assays use heteropolymeric RNA templates, which are less efficiently utilised by cellular DNA polymerases (Sarngadharan *et al.*, 1978). However, because of the sensitivity of the assays, some DNA polymerases can still display sufficient activity to be detected, yielding false positive results (Silver *et al.*, 1993; Pyra *et al.*, 1994; Heneine *et al.*, 1995; Maudru and Peden, 1997). The main advantages of reverse transcriptase assays are that they are able to detect reverse transcriptases of a wide range of retroviruses (Heneine *et al.*, 1995), and that they are relatively insensitive to minor sequence variations in retroviruses which can cause problems in PCR approaches (see below).

1.2.10.6 Specific sequence detection by nucleic acid hybridisation

The retrovirus life cycle includes stages where the genome exists as single stranded RNA in virions and as double stranded DNA integrated into the genome of the host cell (see 1.2.2 The retroviral life cycle). Retroviruses can be detected and characterised by searching for retroviral nucleic acids in either of these stages. Early methods relied on the detection of RNA of the appropriate size and in particles of the appropriate density in the supernatants of metabolically labelled, infected cell cultures (for example see Robinson *et al.*, 1965). Methods involving the simultaneous detection of retroviral RNA and reverse transcriptase were also developed (Schlom and Spiegelman, 1971). Later, liquid hybridisation techniques were used to

detect and characterise retroviral nucleic acids as DNA proviruses in host cell genomic DNA (for example see Temin, 1964), or as transcripts present either in cellular RNA preparations or as RNA genomes in retrovirus particles (for example see Ringold *et al.*, 1975). Subsequent hybridisation approaches utilised Southern blotting (for example see Keshet and Temin, 1978), Northern blotting (for example see Adkins *et al.*, 1982), and *in situ* hybridisation techniques (for example see Kaufman *et al.*, 1979; Rein *et al.*, 1982) to detect and characterise retroviruses. Retroviruses that are distantly related to that used as the probe can be detected by decreasing the stringency of hybridisation, although there are limits to such approaches (Teich, 1982).

1.2.10.7 PCR detection

Recently, methods based on the polymerase chain reaction have largely superseded hybridisation-based approaches for the detection and characterisation of known and unknown retroviruses. Again, retroviral nucleic acids can be detected as proviruses, in cellular RNA, or in purified viruses. A narrow range of viruses can be detected by using primers specific to subsets of retroviruses. Although RT-PCR using specific primers can detect as few as 10 copies of RNA, prior knowledge of the virus sequence is required and PCR can be dramatically affected by relatively small changes in template sequence (Böni *et al.*, 1996). Searches for broad ranges of retroviruses usually utilise degenerate primers derived from highly conserved regions of the *pol* gene, which is the most highly conserved gene amongst retroviruses (McClure *et al.*, 1988). Such primers have recently been used to PCR-amplify endogenous retroviral sequences from human genomic DNA (Shih *et al.*, 1989), to partially characterise unknown retroviruses isolated from cell-culture supernatants (Donehower *et al.*, 1990; Perron *et al.*, 1997), and to detect endogenous retroviral transcripts in human RNA (Herrmann and Kalden, 1994). Amplification using degenerate primers is considerably less sensitive than that using specific primers (Donehower *et al.* [1990] report a detection limit of 5×10^5 virus particles). However, the overwhelming advantage of PCR approaches is that the PCR products can be cloned and

sequenced, aiding identification or classification of detected viruses (see 1.2.4 Retrovirus classification).

1.3 Biology of endogenous retroviruses

1.3.1 Retroviruses can enter the germline

Retroviruses appear to be able to infect germline cells in much the same way as they do somatic cells. Initial entry into the germline almost certainly involves extracellular infection by an exogenous virus, and investigations of germline entry using ecotropic MLVs in mice suggest that, at least in this system, the oocyte is the most likely target for infection (Lock, 1988). Provirus integration in the genome of a germline cell enables transmission of this provirus to offspring in a Mendelian fashion. Such vertically inherited proviruses are known as endogenous retroviruses (Coffin, 1982; Boeke and Stoye, 1997). The consequences of residency in the germline are that the provirus is present at the same location in the genomic DNA of every cell of the host. By contrast, exogenous retroviruses are horizontally transmitted (or vertically transmitted by extracellular infectious routes) and integrate at different locations in a subset of the cells within the host.

The genome of a given animal species may contain several distantly related groups of endogenous retroviruses. For example, mice have been shown to harbour endogenous counterparts of the exogenous mammalian type C (MLV-related) and type B (MMTV) retroviruses. The mouse genome also contains numerous endogenous retroviruses that are not closely related to any known exogenous viruses, including intracisternal A-particle (IAP) elements, VL30s, murine retrovirus-like repeat sequences (MuRRS), and others (for reviews see Coffin, 1982; Boeke and Stoye, 1997). The human genome contains several groups of distantly related endogenous retroviruses which have been divided into two large classes: class I elements have homology with mammalian type C retroviruses; class II elements have highest homology with IAPs and type B retroviruses (Wilkinson *et al.*, 1994).

1.3.2 Amplification within the germline

Endogenous retroviruses are present in the genomes of vertebrates in widely varying numbers. Some endogenous retroviruses are present in only a single copy (eg. HRES-1 in humans [Perl *et al.*, 1989]), whilst others are present in over one thousand copies (eg. rodent intracisternal A-particle elements, IAPs [Kuff and Lueders, 1988]). Amplification within the germline can occur either by extracellular reinfection of the germline or by intracellular retrotransposition within germline cells (Boeke and Stoye, 1997). In addition, moderate numbers of full-length retroviruses are sometimes accompanied by several-fold more solitary LTRs. For example, the human genome contains approximately 50 full-length copies of the endogenous retrovirus HERV-K (Ono, 1986), and approximately 25000 solitary HERV-K LTRs (Leib-Mösch *et al.*, 1993). Solitary LTRs are thought to arise by recombination between the two LTRs of a full-length retroelement, resulting in the deletion of the sequence internal to the LTRs (for examples see Seperack *et al.*, 1988; Mager and Goodchild, 1989).

Excessive proviral amplification is likely to be deleterious to the host cell/organism, and mechanisms probably exist to limit the expansion of proviral numbers. These mechanisms may include, but are not limited to: receptor interference by endogenously expressed Env proteins (which block reinfection of the cell); hypermethylation of newly acquired proviral DNA, thereby preventing its expression; inhibition of expression of proviruses in the reproductive tract via tissue-specific silencers; inactivation of infectivity by deletions and mutations; and excision of coding sequences by homologous recombination between the proviral 5' and 3' LTRs, leaving a solo LTR (Löwer *et al.*, 1996; Boeke and Stoye, 1997; Sverdlov, 1998).

1.3.3 Ancient vs recent

Endogenous retroviruses are generally regarded as genetic fossils that provide a record of the retroviral infections an animal species has endured during its evolutionary history. In this respect, endogenous retroviruses are

classified as ancient or recent, depending on the length of time they have resided in the germline of the host (Coffin, 1993; Boeke and Stoye, 1997).

Ancient endogenous retroviruses are defined as those that were present in the germline of the host before speciation. Consequently, they are present at the same location in the genome of all members of the host species, and sometimes related species. Ancient endogenous retroviruses have been found in all well-studied vertebrate genomes. Although they are clearly of retrovirus origin, they are often only distantly related to exogenous retroviruses in their host species. Their genomes are typically scarred by numerous mutations, a result of the length of time they have resided within the host genome, and they usually display little or no biological activity (Coffin, 1982; Coffin, 1993; Boeke and Stoye, 1997).

Recent endogenous retroviruses are those that entered the germline post-speciation. They are therefore present at different locations in the genomes of different individuals (or populations) or may be absent from the germline in some individuals (or populations). They are normally very closely related to exogenous viruses that infect the same species. Recent endogenous retroviruses can retain extensive biological activity, and may even express infectious virus particles (Coffin, 1982; Coffin, 1993; Boeke and Stoye, 1997).

1.3.4 Biological activity of endogenous retroviruses

The range of biological activities associated with endogenous retroviruses is extreme, from none at all to competency for infection. The majority of individual endogenous retroviruses are transcriptionally silent. Nevertheless, transcripts of endogenous proviruses are frequently detected, although these transcripts are often translationally defective. Transcription may be ubiquitous, or it may be restricted to certain tissues, developmental stages, or physiological states (Coffin, 1982; Wilkinson *et al.*, 1994; Boeke and Stoye, 1997). In some cases, one or a subset of proteins may be expressed from an endogenous provirus. An example of this is provided by the human endogenous element ERV-3, which possesses numerous mutations in the

gag and *pol* genes, but has an intact *env* gene, whose product is expressed in a subset of human tissues (for review see Wilkinson *et al.*, 1994; Patience *et al.*, 1997). Endogenous retroviruses may also express all of the proteins required for the formation retrovirus particles, as exemplified by the HERV-K family of human endogenous retroviruses, that produce particles which fail to undergo maturation and are not infectious (Tönjes *et al.*, 1996). Recent endogenous retroviruses (see 1.3.3 Ancient vs recent) may also encode fully functional, infectious retroviruses, which can be expressed spontaneously or in response to physiological or external stimuli (see Coffin, 1982).

Analysis of the biological activity of individual endogenous viruses is complicated by functional complementation. Endogenous retroviruses can complement, and be complemented by, other endogenous and exogenous retroviruses. Complementation of exogenous viruses by endogenous retroviruses was first observed as the ability of some chicken cell lines to provide Env proteins for Rous sarcoma virus defective in the *env* gene (Weiss and Payne, 1971). The ability of IAPs defective in Gag to retrotranspose within the genome of host cells may provide an example of complementation between endogenous viruses (Lueders and Kuff, 1989). Particles released from human mammary carcinoma cell-line T47-D possess reverse transcriptase activity but contain transcripts defective in the *pol* gene, again suggesting complementation between endogenous elements (Seifarth *et al.*, 1995). An extreme example of the complementation of endogenous elements by exogenous retroviruses is provided by mouse VL30s. These endogenous elements do not appear to encode any functional proteins, but do possess all the *cis*-acting sequences required for replication, and are able to extracellularly retrotranspose using proteins provided in *trans* by exogenous retroviruses (reviewed in French and Norton, 1997).

1.3.5 Endogenous retroviruses can escape the germline

Endogenous retroviruses, which do not rely on transcription or reverse transcription for their replication, and generally undergo far fewer cycles of replication per unit time than do exogenous retroviruses, evolve much more slowly than do exogenous retroviruses (Doolittle *et al.*, 1989). It appears that

retroviruses spend the majority of their time as slowly evolving endogenous elements. These long periods of germline existence are interrupted by only brief periods of exogenous replication and rapid evolution (Doolittle *et al.*, 1989; Coffin, 1993). In this way, endogenous retroviruses might be reservoirs of "potential infectious agents that may on rare occasions be triggered to infect and cause disease in other organisms" (Wilkinson *et al.*, 1994). The putative natural history of endogenous and exogenous retroviruses is shown diagrammatically in FIG. 1.7. If an exogenous retrovirus is excessively pathogenic, it may overwhelm its host population, and consequently bring about its own extinction. Alternatively, an exogenous retrovirus may become extinct if it fails to spread sufficiently within its host community, or if it is successfully neutralised by the immune system of its host. Infection of the germline of its host may permit long-term survival, and slower evolution, as an endogenous retrovirus. Endogenous retroviruses can escape from the germline by infecting other organisms, thereby resuming an exogenous lifestyle. Alternatively, endogenous retroviruses

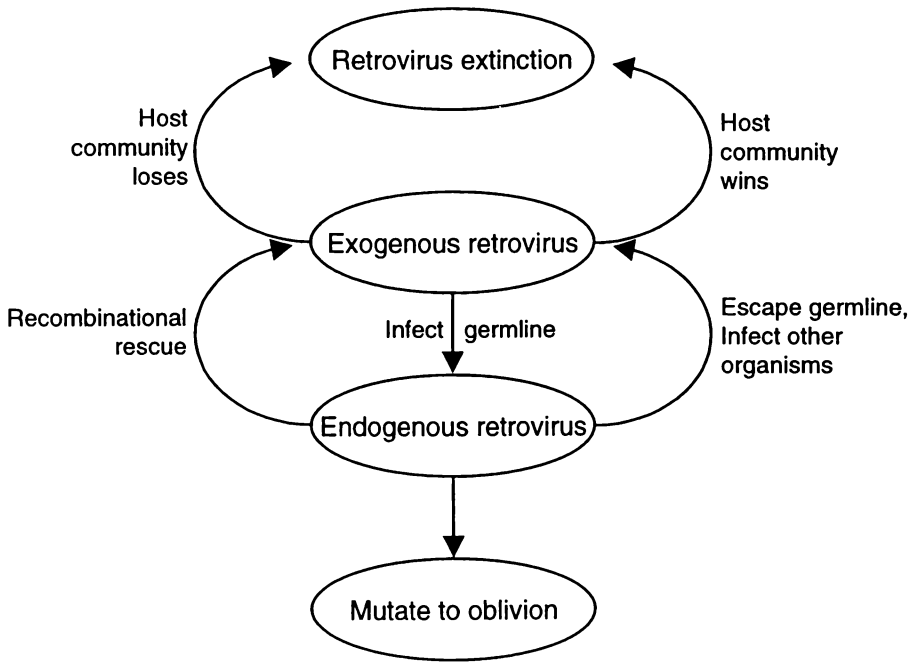


FIGURE 1.7 Postulated natural history of exogenous and endogenous retroviruses. See text for details. Adapted from Doolittle and Feng (1992).

may be recovered by recombination with other retroviruses, producing new exogenous retroviruses with altered biological properties. However, a prolonged endogenous existence will result in the accumulation of inactivating mutations and, eventually, a permanent loss of biological activity (Doolittle, 1989; Feng and Doolittle, 1992).

Several occurrences of cross-species transmission of endogenous retroviruses have been observed. Over two decades ago, cross-species transmission of mammalian type C retroviruses was recognised when closely related endogenous retroviruses in Old World primates and cats were discovered (reviewed in Todaro, 1975). Additional transfers from rats to cats and from rodents to pigs were also noted (Todaro, 1975). A recent study, also with mammalian type C retroviruses, detected close relationships between endogenous retroviruses of echidnas (monotremes) and the exogenous avian reticuloendotheliosis viruses (REVs), and endogenous viruses of the koalas (marsupials) and the exogenous gibbon ape leukemia virus (GaLV) from gibbons (placental mammals) (Martin *et al.*, 1999). Both the high levels of homology of the retroviruses, and the observation that neither the REVs or GaLV have been found in the germlines of their respective hosts, suggest that both transfer events occurred recently (Martin *et al.*, 1999). Overall, it appears that cross-species transmission events involving endogenous retroviruses are rare (Martin *et al.*, 1999), although the possibility that they may occur has prompted discussion of the potential risks of xenotransplantation (for example see Brown *et al.*, 1998).

1.3.6 Implications of presence of retroviruses in the genome/germline

Early reports estimated that sequences contributed by reverse transcription constituted up to 5% of the genomes of humans and mice (Baltimore, 1985). Recent estimates put the level much higher. LTR-retrotransposons (retrotransposons that resemble retroviruses in that they possess LTRs and encode Gag and Pro proteins, but differ from retroviruses in that they lack *env* genes) and endogenous retroviruses are thought to make up ~5% of the human genome, long interspersed nuclear elements (LINEs) ~15%, and all remaining sequences generated by reverse transcription (mainly short

interspersed nuclear elements [SINEs] and retrotranscripts, both of which depend on reverse transcriptase encoded by other retroelements for retrotransposition) another ~10 to 15% (Smit, 1996; Jurka, 1998). It has been suggested that retroelements (and other repetitive elements) are "selfish" DNA elements, which arose and have proliferated within eukaryotic genomes merely because they were able to, and because they exerted no (or only slight) negative phenotypic effects on their hosts (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Nevertheless, it is clear that the presence of at least some of these elements does have implications for their hosts. Their presence also has a number of scientific and technical implications for researchers looking for retroviruses (both exogenous and endogenous).

1.3.6.1 Implications for the host

There are a number of potential implications of the presence of retroviruses in the genome of the host. Some of the implications depend on the biological activity of the elements, others do not.

Retrotransposition of endogenous and exogenous retroviruses (and other retroelements) can have a number of possible effects on the host. Retrotransposition can result in insertional inactivation of genes, if insertion occurs within the coding region or regulatory region of a gene. Such insertional mutations are likely to be recessive (Favor and Morawetz, 1992). The dilute (*d*) coat colour (Jenkins *et al.*, 1981) and hairless (Stoye *et al.*, 1988) mutations in mice are both examples of germline mutations which have been caused by retrovirus insertions into genes (see also Coffin, 1993). Insertional activation of genes may also occur, whereby the LTRs of a provirus can direct or affect transcription of an adjacent gene. In contrast to recessive insertional inactivation mutations, activation mutations are often dominant. Insertional inactivation of a tumour suppressing gene or activation of a proto-oncogene can induce neoplastic proliferation of the cell (for review see Favor and Morawetz, 1992). Proviruses located within, or in close proximity to, cellular genes may also supply other regulatory signals, such as

splice sites and polyadenylation signals (Wilkinson *et al.*, 1994; Boeke and Stoye, 1997).

Endogenous retroviruses could enable the retrotransposition and/or duplication of other sequences within the genome. Defective endogenous retroviruses and/or LTR retrotransposons may retrotranspose using functional reverse transcriptase (and other proteins) encoded by other endogenous retroviruses. It appears likely that the reverse transcriptases of LINEs are responsible for the retrotransposition abilities of SINEs and for the duplication events that give rise to pseudogenes and retrogenes (Dhellin *et al.*, 1997), although endogenous retroviruses are another potential source of the requisite RT activity (Linial, 1987). In addition, endogenous retroviral LTRs, if they are located upstream of pseudogenes before or after pseudogene formation, may enable them to remain or become transcriptionally active (Wilkinson *et al.*, 1994). Reverse transcriptase, encoded by endogenous retroviruses or other retroelements, might also mediate gene conversion, whereby reverse transcripts of genes are made available for recombination with their genomic DNA homologues (Coffin, 1993).

Expression of endogenous retroviruses may have consequences regarding infection with exogenous retroviruses. Endogenous retroviruses may protect the host from infection by exogenous retroviruses by blocking infection of host cells or ablating target cells. The *Fv4* and *Fv1* loci are endogenous retroviruses in mice. Expression of the *env* gene of the *Fv4* provirus prevents entry by related retroviruses through a process known as receptor interference — the endogenously expressed Env protein binds to the cell-surface receptor, making unavailable for binding by exogenous virus (for reviews see Gardner *et al.*, 1991; Best *et al.*, 1997). The *Fv1* locus encodes a Gag protein (Best *et al.*, 1996) that appears to block infection at a post-entry stage (for review see Best *et al.*, 1997). Also in mice, expression of the superantigen encoding (*sag*) gene of endogenous MMTVs induces the ablation of responsive T-cells, and prevents subsequent infection by MMTVs with the same *sag* gene (Coffin, 1992b). However, expression of endogenous retroviral proteins during the establishment of immune tolerance

may reduce the immunological response to a later infection by a related exogenous retrovirus (Crittenden *et al.*, 1982; 1984). A final example of interaction between endogenous and exogenous retroviruses is the observation that recombination between endogenous viruses and exogenous viruses can occur (for examples see Weiss *et al.*, 1973; Evans and Cloyd, 1985) and can even restore competency to defective exogenous viruses (Martinelli and Goff, 1990; Murphy and Goff, 1994). Similarly, recombination can occur between endogenous retroviruses and retroviral gene-therapy vectors, and is one of the major concerns with retroviral-vector gene-transfer procedures (Miller, 1997; Chong *et al.*, 1998).

Endogenous retroviruses, or their proteins, may induce diseases in their hosts. In mice, two examples of endogenous retrovirus-induced disease have been documented. In one case, a series of recombination events between several endogenous retroviruses generates a retrovirus (mink cell focus-forming virus, MCF) that induces thymic lymphomas, by insertional activation of a proto-oncogene, in some strains of mice (Stoye *et al.*, 1991). In the other case, the replication of endogenous MMTV proviruses results in induction, by insertional mutagenesis, of mammary tumours in some strains of mice (for review see Boeke and Stoye, 1997). Potential roles of endogenous retroviruses in autoimmune diseases have been proposed, but remain contentious (for reviews see Krieg *et al.*, 1992; Wilkinson *et al.*, 1994; Urnovitz and Murphy, 1996; Boeke and Stoye, 1997). Also, the possibility that ERV proteins could be adopted, by the host, to serve a specific function has also been presented (eg. HERV-R Env proteins in the syncytiotrophoblast layer of the human placenta) (Wilkinson *et al.*, 1994).

Finally, the mere presence of endogenous retroviruses may affect genome structure. Repetitive sequences, including ERVs and their solitary LTRs as well as other retroelements, are thought to serve as foci for recombination events. Homologous but unequal crossing over can occur between related sequences at distinct loci on the same, or even different, chromosomes. Such major chromosomal changes are likely to be important genetic factors driving speciation (Boeke and Stoye, 1997).

1.3.6.2 Implications for detection of retroviruses

Because they often retain partial or full biological activity, endogenous retroviruses can complicate searches for exogenous retroviruses. Endogenous proviruses and their transcripts are frequently detected by Southern hybridisation and northern hybridisation, respectively, using probes derived from known retroviruses (Coffin, 1982; Wilkinson *et al.*, 1994). Endogenous viral antigens related to exogenous retroviral proteins are also seen. Endogenous retroviruses may express virus particles, which may be detected using electron microscopy, and may also possess biochemical and physical properties similar or identical to those of exogenous viruses. They may even possess functional enzymes, such as reverse transcriptase, and be competent for infection. Thus, for all intents and purposes, endogenous retroviruses may be detected using all of the techniques used to search for exogenous retroviruses. Approaches that seek to simultaneously detect characteristics of fully functional retroviruses, such as detection of reverse transcriptase activity and/or virus RNA in sucrose density gradient fractions of the appropriate density, should increase the likelihood of detecting viruses that are actually infectious.

1.4 The common brush-tailed possum (*Trichosurus vulpecula*)

1.4.1 Possum evolution

The common brush-tailed possum (*Trichosurus vulpecula*) is an Australian marsupial belonging to the Phalangerid family (Order Marsupialia, Family Phalangeridae) (Ganslosser and Etter-Ganslosser, 1990). The earliest marsupial fossils are from lower Cretaceous (~75 million years before present [MYBP]) North America. It is most probable that the first marsupials appeared in either North or South America, and that their spread to other continents, including Australia, occurred at a time when these continents were still joined (Thenius, 1990). Those marsupials that entered Australia subsequently evolved into a variety of forms (Thenius, 1990). The closest relatives to the brush-tailed possum are the other members of the family Phalangeridae, the scaly-tailed possums (genus *Wyulda*) and the cuscus

(genus *Phalanger*). The oldest Phalangeridae are from the middle Miocene (~14 MYBP) of Australia (Thenius, 1990).

The order of divergence of the three mammalian orders (Monotremata [egg-laying mammals], Marsupialia, Eutheria [placental mammals]) is under question. Recent attempts to resolve the issue using molecular data have not reached a consensus. One group, using mitochondrial DNA sequences, has proposed an early divergence of the eutherians away from the other mammals (~130–143 MYPB), followed by the separation of the monotreme and marsupial orders (~116–126 MYBP). The split between the South American and Australian marsupials was estimated to have occurred ~75 MYBP (Janke *et al.*, 1997). Another group, using sequences of the neurotrophins from monotremes, marsupials, and eutherians, suggest a closer relationship between the marsupial and eutherian mammals to the exclusion of monotremes (Kullander *et al.*, 1997).

1.4.2 Introduction into New Zealand

In the late 1800s and early 1900s, the possum was introduced into New Zealand from Australia to establish a fur trade. The first successful liberation in New Zealand was made in 1858, with most importations occurring between 1890 and 1900. The total imports numbered about 200–300 Australian and Tasmanian individuals. The subsequent spread of possums was accelerated by additional liberations of the New Zealand-bred progeny of the original introductions between 1890 and 1940 (Cowan, 1990). Possum numbers have steadily increased since — an estimated 60 million possums now occupy more than 90% of New Zealand (Cowan and Tyndale-Biscoe, 1997).

1.4.3 Possums are pests

Possums are now regarded as a serious ecological pest in New Zealand, and are a major economic threat. They destroy native forests by browsing on plants (Cowan, 1990), and compete with native birds for food, disturb their nesting and prey on chicks and eggs (see Cowan, 1996, and references

therein). They also damage agricultural and forestry plantations, as well as erosion-control plantings, at a financial cost of NZ$30–60 million per annum (Cowan, 1996). The greatest economic threat that possums pose, however, is to the dairy and beef industries — possums serve as a sylvatic reservoir of bovine tuberculosis (Morris and Pfeiffer, 1995).

1.4.4 Possum biocontrol

Current methods for controlling possum numbers include poisoning, trapping, and shooting. However, the costs of these methods, as well as public concerns over the use of poisons, particularly aerially-distributed sodium monofluoroacetate (1080), have brought calls for alternative methods of possum control (Cowan, 1996).

Since 1993, biocontrol of possums has been recognised as the key long-term goal of New Zealand's National Science Strategy for the control of possums and bovine tuberculosis. The most humane, and publicly accepted, method of biocontrol is likely to be one that prevents reproduction. Ideally, the blocking agent would be delivered to possums using a possum-specific transmissible vector (Jolly, 1993; Cowan, 1996; Cowan and Tyndale-Biscoe, 1997). Research is currently being conducted in several areas of possum biology, including general and reproductive physiology, and sociobiology (Lynch, 1998). Searches for, and characterisation of, possum pathogens and parasites are also being conducted (Lynch, 1998). Although viruses have been observed (Mackintosh *et al.*, 1995; Rice and Wilks, 1996; O'Keefe *et al.*, 1997; O'Keefe and Wickstrom, 1998; Meers *et al.*, 1998), no efforts have been made to use them as vectors for possum control.

1.4.5 Retroviruses in possums and other marsupials

There have been few reports of retroviruses in marsupials. Hamilton *et al.* (1979) reported the detection, by electron microscopy, of budding A-type retrovirus particles, possibly of type D retrovirus origin, in two cell lines derived from *Sminthopsis crassicaudata* (fat-tailed dunnart). Also using electron microscopy, Canfield *et al.* (1988) detected particles, which they

considered most likely to be oncoviruses (mammalian type C, avian type C, mammalian type B, or mammalian type D), in the bone marrow of a leukaemic koala (*Phascolarctos cinereus*). Similarly, Worley *et al.* (1993) reported observing retrovirus particles in neoplastic tissues of koalas. Although they also claimed to have amplified a portion of the retroviral genome using degenerate primers from the retroviral *pol* gene, the sequence has not been reported (Worley *et al.*, 1993). Syncytia formation and Mn^{2+}-dependent reverse transcriptase activity were observed in cultures of possum peripheral blood mononuclear cells (PBMCs), but no virus particles were detected using thin section electron microscopy (Meers *et al.*, 1998).

More recently, molecular biology techniques have been used to detect endogenous retrovirus sequences in the genomes of marsupials. Amplification of a retroviral element termed kangaroo endogenous retroviral element-1 (KERV-1) was associated with expansion of chromosomal centromeres and genome-wide undermethylation in a hybrid produced from the mating of two kangaroo species, *Wallabia bicolor* (swamp wallaby) and *Macropus eugenii* (tammar wallaby) (O'Neill *et al.*, 1998). A two kilobase fragment from KERV-1 displayed greatest similarity to the *pol* regions of a human endogenous retrovirus, HERV-K10, and the Y-chromosome-linked mouse intracisternal A particle, the MIAP-Y virus. Using PCR and primers derived from conserved regions within protease and reverse transcriptase genes of retroviruses, Martin *et al.* (1997) generated an 873 bp fragment from rock wallaby (*Petrogale gadmani*) genomic DNA. Phylogenetic analysis grouped this fragment most closely to the human endogenous retrovirus HERV-I (and HERV-I-related sequences), and more distantly with members of the MLV genus. Southern hybridisation using the rock wallaby fragment revealed the presence of related (presumably endogenous) retroviruses in the genomic DNA of several different families of marsupials, including the common brush-tailed possum (Martin *et al.*, 1997). Using similar primers, Herniou *et al.* (1998) amplified and sequenced retrovirus-related fragments from the stripe-faced dunnart (*Sminthopsis macroura*) and the brush-tailed possum. Two fragments were cloned from the stripe-faced dunnart. Phylogenetic analysis placed one of the fragments in a polytomy which included the type B and type D retroviruses, HERV-K elements, avian type C

viruses, and rodent intracisternal type A particle (IAP) elements. The second stripe-faced dunnart fragment grouped closely with ERV-9, and was more distantly related to the MLV group. The possum fragment grouped with endogenous retroviruses found in humans (HERV-L) and mice (MuERV-L), and showed little similarity to either of the dunnart sequences, or that of the rock wallaby described above (Herniou *et al.*, 1998). Finally, in their study of the distribution of murine leukemia virus-related endogenous retroviruses in vertebrates, Martin *et al.* (1999) isolated fragments from the koala (*Phascolarctos cinereus*) and the American opossum (*Monodelphis* ssp.). Interestingly, the koala sequence was highly homologous to that of an exogenous gibbon retrovirus, gibbon ape leukemia virus (GaLV), suggesting a recent cross-species transmission event (Martin *et al.*, 1999).

1.5 Aims of this study

The aims of this study were to detect and characterise retroviruses in the common brush-tailed possum using molecular biology techniques, and assess their potential for use in the biological control of possums. These studies should complement those of others (Rice and Wilks, 1996; Meers *et al.*, 1998) who are using more traditional virology methods, such as cell-culture and electron microscopy, to isolate viruses from possums. Initial attempts were focussed on the detection of exogenous retroviruses, as it was anticipated that biologically active retroviruses would have more potential for use as biocontrol vectors. Thus, investigations using PCR with degenerate primers were performed to determine whether endogenous retrovirus sequences were present in the possum genome. Reverse transcriptase assays and RT-PCR analyses on possum blood plasma extracts were used to detect and characterise biologically active retroviruses. Short (~130bp) clones generated using degenerate primers allowed the design of specific primers which were used to amplify a larger (~4.5kb) fragment from blood serum RNA. The sequence obtained from possum blood was subsequently found to be derived from, or closely related to, an endogenous retrovirus in the possum genome. Subsequent work focussed on the isolation of full-length copies of this endogenous element, with the aim of characterising, for the first time, a full-length marsupial retroviral

genome. To this end, a near-full-length copy of this element was amplified from possum DNA, sequenced, and analysed. The likely origins of this retrovirus and the implications of its presence in the possum genome are discussed.

Chapter 2

Materials and Methods

2.1 Possum tissue samples

Blood and tissue samples were obtained from possums which were routinely euthanased by staff at the Animal Welfare and Behaviour Unit, AgResearch Ruakura, Hamilton. Possums were euthanased by intraperitoneal injection of 4 ml of 300 mg/ml sodium pentobarbital (Pentobarb 300, Chemstoch Animal Health Ltd., Auckland). Blood samples were retrieved by cardiac puncture using a 0.9x25mm (20 G 1") Vacutainer needle (Becton Dickinson) and 9 ml EDTA Vacuette tubes (Greiner Labortechnik), and were stored on ice until processed. Other tissue samples were taken using a sterile scalpel, and were placed in sterile plastic bags. They were immediately frozen by immersion in liquid nitrogen, and subsequently stored at –70°C.

2.2 Standard reagents, solutions, and media

All standard reagents, solutions, and media were prepared as described by Sambrook et al (1989) and Ausubel *et al.* (1998) using double-distilled and deionised water (dH_2O) and chemicals and biochemicals from major suppliers (Ajax Laboratory Chemicals, BDH, Boehringer Mannheim, Life Technologies, Sigma, United States Biochemicals). Molecular biology reagents were purchased from Boehringer Mannheim, Life Technologies, and Promega. Disposable plasticware was manufactured by Axygen Scientific, Falcon, Nalgene, NUNC, Sarstedt, and Scientific Specialties Incorporated (SSI).

2.3 Standard molecular biology procedures

2.3.1 DNA precipitation

Unless otherwise stated, ethanol precipitation of DNA was performed according to Sambrook *et al.* (1989). To the DNA solution, 3 M sodium acetate (pH 5.2) was added to a final concentration of 0.3 M and the solution was mixed. Two cumulative volumes of ethanol were added, and the solution was mixed and then incubated at 4°C or room temperature for at

least 15 min. The DNA precipitate was then pelleted by centrifugation at 14000 x *g*.

2.3.2 Restriction enzyme digestion of DNA

Restriction enzymes, and appropriate reaction buffers, were purchased from Boehringer Mannheim and New England Biolabs. Restriction enzyme digestion was performed according to the manufacturers' instructions and standard protocols (Sambrook *et al.*, 1989).

2.3.3 Agarose gel electrophoresis

SeaKem LE agarose (FMC) and SeaPlaque low-gelling temperature agarose (FMC) were used for agarose gel electrophoresis. Agarose gels were typically dissolved and electrophoresed in 1xTAE in the presence of 0.5 µg/ml ethidium bromide. Occasionally, ethidium bromide was omitted from the gel and running buffer, and the gel was stained after electrophoresis in 1xTAE containing 0.5–1.0 µg/ml ethidium bromide. DNA in ethidium bromide stained gels was visualised using a Fotodyne Model 3-3102 UV transilluminator or Gibco BRL UV transilluminator TFX-35M (Life Technologies). Gels were photographed using a Polaroid Land camera (Kodak) and Polaroid 667 film or using a COHU High Performance CCD camera and Scion Image (Scion Corporation) image capture and analysis software on a personal computer.

2.4 Nucleic acid isolation

2.4.1 DNA isolation

DNA was isolated from possum tissues using the proteinase K/SDS and phenol/chloroform extraction method described by Sambrook *et al.* (1989), which was modified from the method of Blin and Stafford (1976). Usually, 100–200 mg of tissue was ground in liquid nitrogen using a mortar and pestle, and the powder scraped into a 10 ml tube (Falcon) containing 3 ml DNA extraction buffer (10 mM Tris·HCl [pH 8.0], 0.1 M EDTA [pH 8.0], 20

µg/ml RNase A, 0.5% SDS). The contents were mixed by inverting the tube, and incubated at 37°C for 1 hr. Following addition of proteinase K to a final concentration of 100 µg/ml, the tube was incubated at 50°C for 3 hr. The solution was then extracted three times with an equal volume of Tris equilibrated phenol (pH 8.0), each extraction entailing mixing for 10 min, centrifugation at 5000 x *g*, 15 min, room temperature, and transfer of the aqueous phase to a new tube using a 1 ml autopipette tip with 3 mm of the tip cut off. The final aqueous phase was transferred to a new tube, and the DNA precipitated by addition of a 0.2 volume of 10 M ammonium acetate and 2 cumulative volumes of 96% ethanol. The contents were mixed, and the DNA precipitate transferred to a 1.5 ml microcentrifuge tube using a toothpick. The DNA was washed twice with 70% ethanol, and allowed to air-dry, before being resuspended in TE buffer (pH 8.0).

2.4.2 RNA isolation

RNA was extracted from possum tissue using either the one-step method of Chomcynzki and Sacchi (1987), or using TRIzol® LS reagent (Gibco BRL).

The method of Chomcynzki and Sacchi (1987) is also described in Ausubel *et al.* (1998). To 1 ml of denaturing solution (4 M guanidinium thiocyanate, 15 mM sodium citrate, pH 7.0, 0.5% sarcosyl, 0.1 M 2-mercaptoethanol) in a 13 ml culture tube (Sarstedt), 100 µl of liquid sample or approximately 100 mg of ground tissue sample was added. The mixture was passed through an 18-gauge hypodermic needle 20 times using a 3 ml syringe. The following were added sequentially, with thorough mixing after each addition: 100 µl 2 M sodium acetate (pH 4.0); 1 ml water saturated phenol; 200 µl 49:1 chloroform:iso-amyl alcohol. The final suspension was shaken vigorously for 10 sec, and then incubated on ice for 15 min. Following centrifugation at 10000 x *g*, 20 min, 4°C, 1 ml of the upper aqueous phase was transferred to a 1.5 ml tube containing 1 µl 10 mg/ml yeast tRNA (Sigma), mixed, and divided into two equal volumes in separate 1.5 ml tubes. To each tube, 500 µl of 100% propan-2-ol was added, the contents mixed, and the tubes incubated at –20°C for 30 min. The tubes were then centrifuged at 10000 x *g*, 10 min, 4°C, and the supernatants discarded. The pellets were

resuspended in 500 µl 75% ethanol (in DEPC treated water), and incubated at room temperature for 10–15 min. Following centrifugation at 1000 x *g*, 10 min, 4°C, and removal of the supernatants, the pellets were allowed to air-dry. The pellets were resuspended in DEPC-treated TE (pH 8.0) and combined, and stored at –70°C.

RNA isolation using TRIzol® LS reagent (Gibco BRL), a mono-phasic solution of phenol and guanidine isothiocyanate, is a modification of the one-step RNA isolation procedure of Chomczynski and Sacchi (1987) described above. To 750 µl of TRIzol® LS reagent (Gibco BRL) in a 1.5 ml microcentrifuge tube, 250 µl of sample was added, and the solution mixed by pipetting several times with a 1 ml autopipette. Following incubation at room temperature for 5 min, 200 µl chloroform was added, the solution mixed by shaking the tube vigourously for 15 sec, and the tube incubated at room temperature, 5–15 min. After centrifugation at 11685 x *g*, 15 min, 5°C, the upper aqueous phase was transferred to a 1.5 ml microcentrifuge tube containing 1 µl of 10 mg/ml tRNA (Sigma). 500 µl propan-2-ol was added, the contents were mixed thoroughly, and the tube was incubated at room temperature, 10 min. The tube was then centrifuged at 11685 x *g*, 10 min, 5°C, and the supernatant discarded. The pellet was washed by adding 1 ml 75% ethanol (in DEPC treated water), mixing thoroughly, and centrifuging 7283 x *g*, 5 min, 5°C. The supernatant was removed, and the pellet either air- or vacuum-dried. The RNA pellet was resuspended in 10 µl DEPC-treated double distilled, deionised water by heating the tube to 55–60°C for 10 min.

2.5 DNA quantitation

Three methods were used for measuring or estimating DNA concentration.

Spectrophotometric determination of the DNA concentration was performed as described by Sambrook *et al.* (1989). The absorbance (A) of the DNA solution at 260 nm (A_{260}) and 280 nm (A_{280}) were measured using a Shimadzu UV-250 UV-visible spectrophotometer. The A_{260} was used to calculate DNA concentration (A_{260} of 1 corresponds to approximately 50

ug/ml double-stranded DNA), and the A_{260}/A_{280} ratio gave an estimate of the purity of the DNA (an A_{260}/A_{280} ratio of 1.8 corresponding to pure DNA).

DNA concentration was also measured using a DyNAquant 200 Fluorometer (Hoefer) and the H33258 (Hoescht) dye. PCR product, plasmid, and genomic DNA samples were processed as described by the supplier, and concentrations were determined relative to a calf thymus DNA (Sigma) standard of known concentration.

Finally, concentrations of some PCR products were estimated by visually comparing the intensity of the band of interest, in an ethidium bromide stained agarose gel, with a DNA mass ladder (Life Technologies or New England Biolabs).

2.6 Reverse transcription

Reverse transcription was performed using Superscript™ II RNase H– Reverse Transcriptase (Gibco BRL) according to the manufacturer's instructions. Typically, the RNA template and primer (2 pmol of gene specific primer or 500 pmol of poly[A] primer) were denatured by incubation at 70°C for 10 min, placed immediately on ice, and then centrifuged briefly. Other reaction components were then added to give a final reaction volume of 20 µl containing: 50 mM Tris·HCl (pH 8.3); 75 mM KCl; 3 mM $MgCl_2$; 10 mM dithiothreitol (DTT); 500 µM each dNTP (Boehringer Mannheim); and 200 U Superscript™ II reverse transcriptase. When RNasin RNase inhibitor (Promega) was used, it was incorporated at a final concentration of 1.2 U/µl. Reverse transcription was allowed to proceed by incubating the reaction at 42–50°C for 45–60 min, after which the reverse transcriptase was inactivated by incubation at 70°C, 15 min. RNA complementary to the cDNA was removed by adding 1 U *E. coli* RNase H (Boehringer Mannheim), mixing, and incubating at 37°C, 30 min. Generally, 1–5 µl of this reaction was used for subsequent amplification by PCR.

2.7 Polymerase Chain Reaction (PCR)

2.7.1 PCR primer design and synthesis

PCR primers were designed using the Oligo 4.05 Primer Analysis Software (National Biosciences, Inc.) for Macintosh. Oligonucleotides were ordered from Life Technologies. Primers were resuspended and diluted in either TE buffer (pH 8.0) or sterile dH_2O.

Melting temperatures of primers were calculated using several methods, although temperatures based on the nearest neighbour method of Breslauer *et al.* (1986) gave the most consistent results, particularly with long distance PCR. The sequences of all primers, along with their nearest neighbour melting temperatures and other relevant information, are listed in Appendix 2.

2.7.2 Standard PCR

The polymerase chain reaction (PCR) was carried out using AmpliTaq® DNA polymerase (Perkin Elmer) or Taq DNA polymerase (Boehringer Mannheim) according to the manufacturers' instructions. Typically, 50 µl reactions were performed and contained 100 mM Tris-HCl, pH 8.3 (20°C), 50 mM KCl, 1.0–3.0 mM $MgCl_2$, 0.2 mM each dNTP, 1 µM each primer, 1.0–2.5 U AmpliTaq® or Taq DNA polymerase, and 10–500 ng genomic DNA template (or equivalent molar amount of other templates). Where tetramethyl ammonium chloride (TMAC) was used, it was incorporated at a final concentration of 60 mM according to Chevet *et al.* (1995).

Reactions were performed in 200 µl PCR tubes and were either overlaid with Nujol mineral oil (Perkin Elmer) or parafin wax beads (Bayer), or were carried out in thermal cyclers with heated lids. Reactions were set up on ice, and were transferred to thermal cyclers that had been preheated to the initial denaturing temperature. Thermal cyclers used were Progene (Techne), PTC-100™ or PTC-200™ Programmable Thermal Cyclers (MJ Research, Inc.), and GeneAmp® PCR Systems 2400 and 9600 (Perkin Elmer).

Maximum ramping speeds were used for those cyclers where ramping speeds were optional. Cycling conditions (denaturation, annealing, and extension temperatures and times, as well as cycle numbers) varied depending on the particular reaction, and the details are listed in Chapters 3, 4, and 5.

2.7.3 Long distance PCR

Long distance PCR was performed using the eLONGase™ system (Gibco BRL) or Expand™ Long Template system (Boehringer Mannheim) according to the manufacturers' instructions.

Reactions using the eLONGase™ system contained 60 mM Tris-SO_4 (pH 9.1), 18 mM $(NH_4)_2SO_4$, 1.0–2.0 mM $MgSO_4$, 200 µM each dNTP, 200 nM each primer, 1.0–2.0 U eLONGase™ enzyme mix, and 10–500 ng genomic DNA template (or equivalent molar amount of other templates).

Three buffer systems can be used with the Expand™ Long Template system. Reactions contained buffer (pH 9.2), $(NH_4)_2SO_4$ as the salt, 1.75–2.25 mM $MgCl_2$, 350–500 µM each dNTP, 300 nM each primer, 1.75–3.5 U Expand™ Long Template enzyme mix; and 10–500 ng genomic DNA template (or equivalent molar amount of other templates).

Reactions were prepared and cycling performed as described above (2.7.2 Standard PCR).

2.7.4 Purification of PCR products

Purification of PCR products, to remove reaction components and primer-dimers, was performed using the Wizard® PCR Preps DNA purification system (Promega) or the High Pure™ PCR product purification kit (Boehringer Mannheim) according to the manufacturers' instructions.

2.7.5 Isolation of DNA from agarose slices

A number of methods were used to isolate DNA from agarose slices. In all cases, exposure of the DNA band(s) of interest to damaging UV light was minimised by one of two methods. Normally a Glad® oven bag (mylar) was placed between the UV light source and the gel and the band of interest was quickly excised using a sterile scalpel blade. Alternatively, the DNA of interest was electrophoresed in two adjacent lanes. Following electrophoresis, the gel was cut between the lanes, and one of the lanes was exposed to UV and the positions of the edges of the gel and the band(s) of interest were marked on a plastic bag. The unexposed lane was aligned with the markings on the plastic bag and the band(s) of interest were excised using a sterile scalpel blade.

DNA was extracted from agarose slices using the freeze-squeeze method followed by ethanol precipitation, by phenol/chloroform extraction followed by ethanol precipitation, or using an Agarose Gel DNA Extraction Kit (Boehringer Mannheim).

For the freeze-squeeze method (D. Musgrave, pers. comm.), the agarose slice was initially frozen at –20°C. The slice was then placed in the corner of a parafilm pouch, and squeezed between the thumb and forefinger. An autopipette was used to transfer the liquid squeezed from the gel to a 1.5 ml tube. The parafilm pouch was then placed between two freezer packs (–20°C or –70°C) until the agarose slice was frozen. The agarose was then squeezed again, and liquid removed. This process was repeated until liquid could no longer be squeezed from the agarose. DNA in the solution was concentrated by precipitation with sodium acetate and ethanol.

For phenol/chloroform extraction of (usually low melting point) agarose slices, the agarose slice was weighed in a 1.5ml tube, and 5–10 volumes of TE buffer (pH 8.0) were added. The agarose was melted by heating the tube to 65°C for 10 min with occasional mixing. The solution was then allowed to cool to room temperature, and was extracted once with an equal volume of phenol (pH 8.0), once with phenol:chloroform (pH 8.0), and once with

chloroform. DNA in the final aqueous phase was transferred to a new tube, and was ethanol precipitated.

Extraction of DNA from agarose slices using the Agarose Gel DNA Extraction Kit (Boehringer Mannheim) was performed according to the manufacturer's instructions.

2.8 Cloning PCR products

Cloning of PCR products generated from possum retroviral DNA and RNA was covered by the Environmental Risk Management Authority (ERMA) permit number GM098/UW002 (*E. coli* modified by retroviral sequences active in possums).

2.8.1 Ligation

Two general protocols for cloning of PCR products were used: restriction-enzyme digestion of PCR products and cloning using cohesive termini; and cloning using a T-overhang vector.

2.8.1.1 Cloning using cohesive termini

In some cases, PCR primers were designed which contained restriction enzyme cleavage sites. PCR products generated using these primers were ethanol precipitated, resuspended in restriction enzyme reaction buffer, and digested with the appropriate restriction enzyme(s). The digested fragment was purified by agarose gel electrophoresis, and was ligated into plasmid vector digested in the same way.

Restriction enzyme digested DNA was cloned into the multiple cloning site of pBluescript® II KS+ or pBluescript® II SK+ (Stratagene). Ligation reactions were performed using T4 DNA ligase supplied by Boehringer Mannheim or Promega according to the manufacturers' instructions. Typically, a 10 µl ligation reaction would contain 10–100 ng vector DNA and a 1:3–3:1 molar ratio of insert:vector DNA. For Boehringer Mannheim T4 DNA ligase, the

reaction contained 66 mM Tris·HCl (pH 7.5), 5 mM $MgCl_2$, 1 mM DTT, 1 mM ATP, and 1 U of T4 DNA ligase. For Promega T4 DNA ligase, the reaction contained 30 mMTris·HCl (pH 7.8), 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP, and 3 U T4 DNA ligase. Ligation reactions were incubated at 4–16°C for 2 hr – 4 days. Ligase was heat inactivated at 65°C for 10 min before ligation products were used for transformation.

2.8.1.2 Cloning using a T-overhang vector

Cloning using T-overhang vectors was performed using T-tailed pBluescript® II KS+ prepared as described by Marchuk *et al.* (1991), or using pGEM®-T vector (Promega).

T-tailed pBluescript® II KS+ was prepared essentially as described by Marchuk *et al.* (1991). 5–10 µg of pBluescript® II KS+ was digested using 40 U of *Eco*RV (Boehringer Mannheim) in a 50 µl reaction. The digested DNA was purified by phenol:chloroform extraction and was ethanol precipitated. *Eco*RV digested DNA was T-tailed in a 100 µl reaction containing 2.0 mM dTTP, 7.5 U Taq DNA polymerase (Boehringer Mannheim), 2.5 mM $MgCl_2$, 100 mM Tris·HCl (pH 8.3), 50 mM KCl, by incubating at 70°C for 2 hr. T-tailed vector was then extracted once with phenol:chloroform and once with chloroform, and ethanol/sodium acetate precipitated. Vector was resuspended in TE (pH 8.0) and quantitated using a DyNAquant 200 Fluorometer (Hoefer).

Normally, increasing the final extension time of the PCR thermal cycling to 30 min was sufficient to generate enough A-tailed PCR product for cloning into a T-overhang vector. In some cases, however, forced addition of adenine residues to the 3' ends of PCR products was necessary. In these cases, the PCR product was purified using the Wizard® PCR Preps purification system, and A-tailed under the same conditions as the T-tailing reaction described above, except using dATP instead of dTTP. The A-tailed PCR product was purified again using the Wizard® PCR Preps purification system, and ligated into a T-overhang vector.

Ligation conditions using T-overhang vectors were as described above for cohesive termini ligations.

2.8.2 Transformation

2.8.2.1 Preparation and transformation of chemically competent cells

Chemically competent cells were prepared using the method of Chung and Miller (1993). XL1-Blue MRF' cells (Stratagene) were streaked on an LB + tetracycline agar plate (prepared by spreading 20 µl of 12.5 mg/ml tetracycline on an LB agar plate), and incubated overnight at 37°C. The following day, 10 ml of LB broth containing 12.5 µg/ml tetracycline in a 50 ml Falcon tube was inoculated with one colony from the LB + tetracycline plate, and was incubated overnight at 37°C with shaking at 225–250 rpm. The following day, 125 µl of this overnight culture was inoculated into 125 ml of LB broth containing 12.5 µg/ml tetracycline in a 250 ml conical flask, and was incubated at 37°C with shaking at 225–250 rpm. The absorbance at 600 nm (A_{600}) of this culture was monitored. When the A_{600} reached 0.3–0.4, the culture was transferred to a 250 ml centrifuge bottle (Beckman), and centrifuged at 1000 x *g*, 10 min, 4°C. The supernatant was decanted, and the pellet was resuspended in 12 ml of 1xTSS (transformation and storage solution: LB broth containing 10% [w/v] PEG8000, 5% [v/v] DMSO, and 20 mM Mg^{2+}; prepared as described by Chung and Miller [1993]), and then transferred to a 50 ml Falcon tube. After incubation on ice for 5 min, the cell suspension was separated into 100 µl aliquots in 1.5 ml microcentrifuge tubes on ice, snap frozen in a dry ice/methanol bath, and stored at –70°C.

Chemically competent cells were thawed on ice just prior to use. 1–5 µl ligation reaction product was added to the thawed cells and mixed by gently flicking the tube. The tube was incubated on ice for 45 min, heat-shocked at 42°C for 1 min, and placed on ice for 3 min. 900 µl of SOC was then added, and the cells incubated at 37°C for 1 hr. Aliquots (100–200 µl) of cells were then plated on LB + Amp + X-gal + IPTG agar plates (prepared by spreading 100 µl 100 mg/ml Ampicillin [Sigma], 6.6 µl 120 mg/ml isopropyl-β-*D*-thiogalactopyranoside [IPTG; Sigma], and 40 µl 20 mg/ml 5-bromo-4-chloro-

3-indolyl-β-*D*-thiogalactopyranoside [X-gal; Sigma] on an LB agar plate), which were incubated at 37°C for 16–20 hr.

2.8.2.2 Preparation and transformation of electrocompetent cells

Electrocompetent *Escherichia coli* strains XL1-Blue MRF' (Stratagene) and STBL2™ (Gibco BRL) were prepared as recommended in the BioRad guide for the Pulse controller.

A colony of XL1-Blue MRF' cells from an LB + Tet agar plate was inoculated into 10 ml of LB containing 12.5 mg/ml Tet in a 50 ml Falcon tube, and incubated at 37°C overnight with shaking at 225–250 rpm. The following day, 5 ml of this overnight culture was inoculated into 500 ml of LB + Tet in a 1 l conical flask, and incubated at 37°C with shaking at 200 rpm. When the A_{600} reached 0.5–0.7, the flask was placed in an ice/water bath for 30 min, with occasional mixing. The cell suspension was then transferred to two 250 ml centrifuge bottles (Beckman), and centrifuged at 4000 x *g*, 15 min, 2°C in a Beckman JA14 rotor. The supernatants were removed, and each pellet was resuspended in 200 ml cold 10% (v/v) glycerol by swirling. Following centrifugation at 4000 x *g*, 15 min, 2°C, and removal of the supernatants, each cell pellet was resuspended in 100 ml cold 10% (v/v) glycerol. The cell suspensions were combined, and centrifuged again at 4000 x *g*, 15 min, 2°C. After removal of the supernatant, the cell pellet was resuspended in 10 ml cold 10% (v/v) glycerol, transferred to a cold 10 ml centrifuge tube, and centrifuged at 4000 x *g*, 15 min, 2°C, in a Beckman JA20.1 rotor. The supernatant was decanted, and the cell pellet resuspended in 1 ml of cold 10% (v/v) glycerol. This cell suspension was separated into 40 µl aliquots in 1.5 ml microcentrifuge tubes on ice, snap frozen in a dry ice/methanol bath, and stored at –70°C.

Electrocompetent STBL2™ cells were prepared as above, except that they were grown in LB broth without Tet, and at 30°C.

Electroporation of electrocompetent XL1-Blue MRF' cells was performed using a Gene Pulser and Pulse Controller (Bio-Rad). Aliquots of

electrocompetent cells were thawed on ice immediately prior to use. 40 µl of cells were transferred to a 1.5 ml microcentrifuge tube containing 1.0–2.5 µl of plasmid or ligation reaction product. The contents were mixed by gently flicking the tube, and then transferred to a cold 0.1 cm electroporation cuvette (BioRad). The cuvette was immediately transferred to the electroporation chamber, and pulsed at 25 µFD, 1.80 kV, 200 Ω. Immediately after pulsing, 960 µl of SOC was added, the cell suspension was transferred to a 10 ml tube, and incubated at 37°C, 1 hr, with shaking at 225 rpm. Aliquots (100–200 µl) of cells were then plated on LB + Amp + X-gal + IPTG agar plates, and incubated at 37°C for 16–20 hr.

Electroporation of STBL2™ cells was performed as above, except that incubation before plating was at 30°C, cells were plated on LB + Amp plates (ie. no X-gal or IPTG on plates), and plates were incubated at 30°C for 24 hr.

2.8.3 Small-scale preparation of plasmid DNA

Two main methods were used for preparation of plasmid DNA from bacterial colonies: a rapid boil miniprep method; and a modified alkaline lysis procedure. Generally, the rapid boil method was used for screening large numbers of colonies quickly, and the modified alkaline lysis procedure was used for preparing larger amounts of plasmid DNA for sequencing. All centrifugation steps below are at maximum speed (14000–20000 x *g*) in a benchtop microcentrifuge and at room temperature.

2.8.3.1 Rapid boil minipreps

Rapid boil minipreps were performed following the method of Holmes and Quigley (1981) as reported by Ausubel *et al.* (1998). A colony of interest was inoculated, using a toothpick or 200 µl autopipette tip, into 1.5–3.5 ml of 2xYT or Terrific Broth containing 80 µg/ml Ampicillin in a 10 ml culture tube, and incubated overnight at 37°C (30°C for STBL2™ cells) with shaking at 225–250 rpm. The next day, 1.5 ml of overnight culture were transferred to a 1.5 ml tube and cells were pelleted by centrifugation for 30 sec. The supernatant was aspirated, and another 1.5 ml of overnight culture added,

centrifuged, and supernatant removed. Cell pellets were resuspended in 300 µl STET (8% [w/v] sucrose, 5% [w/v] Triton X-100, 50 mM EDTA, 50 mM Tris·Cl [pH 8.0]) containing 200 µg lysozyme (Sigma) by scraping the tube across an autopipette-tip rack (Voo and Jacobsen, 1998). Tubes were then placed in a boiling water bath for 1 min. After removal from the water bath, the tubes were centrifuged for 10 min, room temperature. Pellets were removed using a toothpick, and 100 µl of 7.5 M ammonium acetate and 300 µl of 99.7–100% ethanol were added. The contents were mixed by flicking and inverting the tubes, and the tubes centrifuged for 10 min. The supernatants were decanted, the tubes centrifuged briefly, and remaining supernatants removed with an autopipette. Pellets were allowed to air-dry, and were then resuspended in 30 µl TE (pH8.0) containing 33 µg/ml RNase A (Sigma).

2.8.3.2 Modified alkaline lysis with PEG precipitation

The modified alkaline lysis procedure, with PEG precipitation was performed as described by Zhou *et al.* (1997). An overnight culture of cells were grown, and cells harvested as for the rapid boil miniprep procedure. Cells were resuspended in 120 µl GTE (50mM glucose, 25mM Tris·HCl, 10mM EDTA, pH 8.0) by scraping the tube over an autopipette-tip rack. 240 µl of freshly prepared 0.2 M NaOH, 1% SDS was added, and the contents mixed by flicking and inverting the tube until the contents were clear. 360 µl of cold 4 M potassium acetate was added, and the contents mixed by flicking and inverting the tube. The tube was then centrifuged for 3 min. The supernatant was transferred, taking care to avoid transferring any of the precipitate, to a new 1.5 ml tube containing 330 µl propan-2-ol. The contents were mixed, and the tube centrifuged for 3 min. The supernatant was decanted, the tube centrifuged briefly, and the remaining supernatant removed using an autopipette. The pellet was resuspended in 200 µl TE containing 20 µg/ml DNase-free RNase A and incubated at 37°C, 30 min. 100 µl of 40% (w/v) PEG8000 in 30 mM $MgCl_2$ was added, and the contents mixed by flicking and inverting the tube. Following centrifugation for 5 min, and removal of the supernatant with an autopipette, the pellet was washed by addition of 1 ml of cold 75% ethanol and mixing. The tube was

centrifuged for 1 min, the supernatant decanted, the tube centrifuged briefly, and the remaining supernatant removed with an autopipette. The pellet was allowed to air-dry, and was then resuspended in 40–100 µl TE (pH8.0).

2.8.4 Preparation of single-stranded DNA

Single-stranded DNA (ssDNA) for sequencing was prepared using a variation of the method of Sambrook *et al.* (1989). 5 µl of an overnight culture of interest was inoculated into 5 ml 2xYT containing 50 µg/ml Amp and $3x10^6$ pfu/ml VCSM13 helper phage (Stratagene), in a 30 ml culture tube (Sarstedt) and incubated at 37°C, 2 hr, with shaking at 225–250 rpm. Kanamycin was then added to a final concentration of 70 µg/ml, and the cultures incubated at 37°C, overnight, with shaking. The cultures were then transferred to 13 ml culture tubes (Sorstedt), and centrifuged at 3000 x *g*, 5 min, room temperature. 1.5 ml of supernatant was transferred to a 1.5 ml microcentrifuge tube, and centrifuged at maximum speed in a benchtop microcentrifuge, 5 min, room temperature (all remaining centrifugation steps are at maximum speed in a microcentrifuge, and at room temperature). 1.3 ml of supernatant was transferred to a new 1.5 ml tube, and 195 µl of 20% PEG in 2.5 M NaCl was added. The contents were mixed, and incubated on ice, 20 min. The tube was centrifuged for 5 min, the supernatant decanted, the tube centrifuged briefly, and the remaining supernatant removed with an autopipette. The pellet was resuspended in 150 µl TE (pH 8.0) by vortexing, and then extracted twice with an equal volume of phenol:chloroform (pH 8.0). The final aqueous phase was transferred to a new 1.5 ml tube, and 16.66 µl of 3 M sodium acetate (pH 5.2) and 335 µl of 99.7–100% ethanol were added. The contents were mixed by flicking and inverting the tube, which was then centrifuged for 5 min. The supernatant was removed with an autopipette, and the pellet allowed to air-dry. The pellet was resuspended in 10 µl TE (pH 8.0), and 2 µl analysed by agarose gel electrophoresis.

2.9 DNA sequencing

2.9.1 Manual DNA sequencing

Manual sequencing of ssDNA templates was performed using Sequenase T7 DNA polymerase V2.O (United States Biochemical) according to the manufacturer's instructions. 600 ng of ssDNA, 0.5 pmol of sequencing primer, and 2 µl of annealing buffer were combined in a 10 µl volume in a 1.5 ml microcentrifuge tube, incubated at 80°C for 5 min, allowed to cool to room temperature, and then placed on ice. 3.5 µl of labelling mix (1.6 µl milliQ, 1 µl 100 mM DTT, 0.4 µl 5xLabeling Nucleotide Mix [7.5 µM dGTP, 7.5 µM dTTP, 7.5 µM dCTP], 0.5 µl 1000–1500 Ci/mmol [α-35S] dATP) was added. 2 µl (3 U) of Sequenase T7 DNA polymerase in dilution buffer (10 mM Tris·HCl [pH 7.5], 5 mM DTT, 0.5 mg/ml BSA) were then added, the contents mixed using an autopipette, and incubated at room temperature for 4 min. 3.5 µl of this labeling reaction was then transferred to 2.5 µl of each of four dideoxy termination mixtures ("G", "A", "T", and "C") at 41°C (each termination mixture contained 80 µM dATP, 80 µM dCTP, 80 µM dGTP, 80 µM dTTP, and 50 mM NaCl; in addition the "G" mixture contained 8 µM ddGTP, the "A" mixture 8 µM ddATP, the "T" mixture 8 µM ddTTP, and the "C" mixture 8 µM ddCTP), mixed using an autopipette, and incubated at 41°C for a further 5min. The reactions were then placed at room temperature, and 4 µl stop buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol FF) was added to each. Sequencing reactions were denatured at 75°C for 5 min, and placed immediately on ice before 3.5 µl was loaded to a polyacrylamide gel.

Manual sequencing gels were run on an IBI Baserunner system. Acrylamide gels were prepared by mixing 14 ml 18% acrylamide solution (19:1 acrylamide:bis-acrylamide), 6 ml 10xTBE, 40 ml 7 M urea, 550 µl fresh 10% ammonium persulphate, and 55 µl TEMED (N,N,N',N'-tetramethylethylenediamine), and pouring immediately. The gel was allowed to polymerise for 2 hr, before being prerun in 1xTBE for 30 min at 60 W. Sequencing reactions were loaded to the gel, which was then run in 1xTBE at 60 W for 2.5–3 hr. The gel was then transferred to Whatman 3MM paper,

dried on a vacuum gel-drier, and exposed to X-OMAT AR autoradiography film (Kodak) at –70°C for 20–60 hr.

2.9.2 Automated DNA sequencing

Plasmid DNA and PCR products were sequenced using an ABI PRISM® 377 DNA Sequencer and cycle sequencing reagents (PE Applied Biosystems). Rhodamine dye terminator, dRhodamine dye terminator, BigDye™ terminator, and dGTP BigDye™ terminator cycle sequencing chemistries (PE Applied Biosystems) were used. Typically, a 10 µl sequencing reaction contained 4 µl Terminator Ready Reaction Mix (contains dye-labeled terminators, deoxynucleotide triphosphates [dATP, dCTP, dITP, dUTP], AmpliTaq® DNA polymerase, FS, with thermally stable pyrophospatase, $MgCl_2$, Tris·HCl [pH 9.0]), template DNA (100–300 ng dsDNA, or 1–5 ng/100 bp PCR product), and 1.6 pmol sequencing primer. Thermal cycling was performed in 500 µl reaction tubes in an Omnigene thermal cycler (Hybaid), or in 200 µl reaction tubes in a GeneAmp® PCR System 2400 or 9600 thermal cycler (Perkin Elmer) or PTC-100™ Programmable Thermal Cycler (MJ Research). The following cycling conditions were used: 25 cycles of denaturation at 96°C for 10 sec; primer annealing at 45–60°C for 5 sec; extension at 60°C for 4 min. After cycling, extension products were purified using the ethanol/sodium acetate precipitation method recommended by PE Applied Biosystems: 1 µl of 3 M sodium acetate (pH 4.6–5.2) and 25 µl 95% ethanol were added, and the contents mixed by flicking the tube. After incubation on ice for 10 min, the tube was centrifuged at maximum speed in a microcentrifuge for 30 min, and the supernatant was removed using an autopipette. The DNA pellet was washed with 125 µl 70% ethanol, and the tube centrifuged at maximum speed for 5 min. The supernatant was removed, and the DNA pellet vacuum- or air-dried. The pellet was resuspended in 2–3 µl loading buffer (prepared by mixing five volumes of deionised formamide with one volume of 25 mM EDTA (pH 8.0), 50 mg/ml blue dextran), denatured by incubating at 90°C, 2 min, placed immediately on ice. 1–1.5 µl was loaded to a 4% 19:1 or 4.25% 29:1 polyacrylamide (Bio-Rad) gel or 5% LongRanger gel (FMC) prepared and run as described by the manufacturers.

2.10 Southern blotting and hybridisation

2.10.1 Southern blotting

Prior to Southern blotting, agarose gels were soaked in transfer solution (1.5 M NaCl, 0.5 M NaOH) for 2 x 20 min. Gels were blotted essentially according to the method of Sambrook *et al.* (1989). Five sheets of 'beermat' (B&F Paper Ltd.), cut to the same width and length as the gel, were soaked in transfer solution, and placed in a pile on a flat surface. One sheet of Whatman 3MM paper, also cut to the same size as the gel and also soaked in transfer solution was placed on top of the pile. The gel was placed on top of the pile and overlaid with a piece of Hybond N+ nylon membrane (Amersham) soaked in transfer solution, followed by one piece of presoaked Whatman 3MM, one piece of presoaked 'beermat', and one piece of dry 'beermat'. Finally, a stack of paper towels approximately 5 cm thick was placed at the top of the pile. A glass plate or gel tray was placed on top of the paper towels, and the gel was left to blot overnight. Following blotting, the membrane was soaked in 2xSSC for 5 min and excess liquid blotted onto 3MM. DNA was cross-linked to the membrane using a UV crosslinker (Gibco BRL UV Crosslinker BLX-254) ($1.2x10^5$ µJoules) or by exposing the membrane to a UV transilluminator for 10–15 sec. Membranes were stored sealed in plastic bags.

2.10.2 Preparation of probes

The gag hybridisation probe was generated by PCR from pTvERV-K2. Approximately 250 pg of pTvERV was amplified in a standard PCR reaction using 50 pmol of each of TvP05F and TvP07R and 1.5 mM Mg^{2+} (Cycling conditions: Cycle 1: 94°C/40 sec; 60°C/30 sec; 72°C/1 min 30 sec; Cycles 2–20: 94°C/10 sec; 60°C/30 sec; 72°C/1 min 30 sec). The PCR product (1434bp) was purified using the Wizard® PCR Preps purification system.

The *gag* probe was radiolabelled using the *redi*prime II random primer labelling system (Amersham) according to the manufacturer's instructions. Approximately 25 ng of *gag* probe in 45 µl of deionised water was denatured

at 94°C for 5 min, and subsequently added to a *redi*prime II reaction tube. After thorough mixing, the contents were transferred to a PCR tube containing 5 µl Redivue 3000 Ci/mmol [α-^{32}P] dCTP (Amersham), mixed, and incubated at 37°C for 30 min. The probe was then denatured by adding 15 µl 4 M NaOH, and added to the hybridisation solution (see below).

2.10.3 Prehybridisation, hybridisation, and washing

Prior to hybridisation, membranes were prehybridised in 20 ml Church and Gilbert hybridisation solution (0.5 M $NaHPO_4$, 1 mM EDTA, 7% SDS) (Church and Gilbert, 1984) in a Hybaid bottle for 30–60 min at the hybridisation temperature. The Church and Gilbert solution was then discarded and replaced with another 10–15 ml of Church and Gilbert solution. The denatured probe was added, and hybridisation was allowed to occur at 68°C in a rotary hybridisation oven (Hybaid) overnight.

After hybridisation, the hybridisation solution was either discarded or stored in a 50 ml Falcon tube at 4°C for later re-use. The membrane was washed briefly in 2xSSC, 0.1% SDS at room temperature. The membrane was then washed for 2 x 15 min in 0.2xSSC, 0.1% SDS at room temperature for low stringency washes, followed by 2 x 15 min in 0.2xSSC, 0.1% SDS at 42°C for moderate stringency washes, and 2 x 15 min in 0.2xSSC, 0.1% SDS at 68°C for high stringency washes. The membrane was then sealed in a plastic bag, and exposed to XAR-5 film (Kodak) at –70°C, using an intensifying screen, for an appropriate period of time (normally at least 24 hr).

2.11 Product enhanced reverse transcriptase (PERT) assay

Reverse transcriptase activity in possum blood samples was tested using the product enhanced reverse transcriptase (PERT) assay of Pyra *et al.* (1994).

2.11.1 Sample preparation

Possum blood samples, collected in EDTA Vacuette (Greiner Labortechnik) tubes, were centrifuged at 16000 x *g* for 10 min, room temperature. The upper aqueous layer (plasma) was then filtered through a 0.2 µm filter (Millipore), and centrifuged in a 1.5 ml polyallomer tube (Beckman; P/N 348349) in a TLA-100.3 rotor in a Beckman TL-100 Benchtop Ultracentrifuge at 45000 rpm (71215 x g_{avg}) for 90 min, room temperature. The supernatant was removed with an autopipette, and the pellet resuspended in 30 µl buffer A (50 mM KCl, 25 mM Tris·HCl [pH 7.5], 5 mM dithiothreitol, 0.25 mM EDTA, 0.02% Triton X-100, 50% glycerol). Samples were stored at –20°C.

2.11.2 Reverse transcription

For each reaction, 0.3 µg (0.28 pmol) of bacteriophage MS2 RNA (Boehringer Mannheim) and 9 pmol of primer RT-1 (see Appendix 2) in a 1.4 µl volume were denatured by incubating at 95°C, 5 min, annealed by incubating at 37°C, 10 min, and then placed on ice for 5 min. This primed template (1.4 µl) was then added to 23.6 µl reverse transcription buffer, to give a 25 µl reaction containing (in addition to the template and primer) 56 mM Tris·HCl (pH 8.3), 56 mM KCl, 9 mM $MgCl_2$, 11.2 mM dithiothreitol, 1 U/µl of RNasin RNase inhibitor (Promega), 0.13 µg/µl BSA (Boehringer Mannheim), 0.4% (v/v) Triton X-100, and 1 mM each dNTP (Boehringer Mannheim). The reaction was overlaid with 50 µl Nujol mineral oil (Perkin Elmer), and 3 µl of pretreated sample in buffer A (see above) was added. Reverse transcription was allowed to proceed at 37°C for 5 hr, following which the RNasin was inactivated by incubation at 95°C for 7 min.

2.11.3 Product amplification

To the RT reaction mixture, 75 µl of a PCR reaction mixture containing 14 pmol of RT-1, 25 pmol of primer RT-2 (see Appendix 2), 8 ng of RNase A, 2.5 U AmpliTaq® DNA polymerase (Perkin Elmer) in 37.5 mM KCl, 10 mM Tris·HCl (pH 8.3), and 0.01% gelatin were added. This mix was incubated for 30 min at 37°C to ensure MS2 RNA degradation by RNase A; then 25

cycles consisting of 94°C/30 sec denaturation, 55°C/100 sec annealing, and 72°C/110 sec extension were performed on a GeneAmp® PCR system 9600 thermal cycler.

2.11.4 Detection of products

Ten µl of PERT assay product was electrophoresed on a 1.5% agarose gel (1xTAE, 0.5 µg/ml ethidium bromide).

2.12 Chromosome walking using single primer PCR and sequencing

Single primer PCR followed by sequencing were performed using modifications of the method of Screaton *et al.* (1993), with either Taq DNA polymerase (Boehringer Mannheim) or eLONGase™ enzyme mix (Gibco BRL). For Taq DNA polymerase, a reaction containing 100 ng genomic DNA, 20 pmol primer, 100 mM Tris·HCl (pH 8.3), 50 mM KCl, 2.5 mM $MgCl_2$, 250 µM each dNTP, and 2 U Taq DNA polymerase (Boehringer Mannheim) was cycled using the following conditions: initial denaturation at 95°C, 5 min, followed by 60 cycles of 95°C for 30 sec; 45°C for for 30 sec, and 72°C for 2 min. Parallel reactions were performed with annealing temperautres of 55°C and 60°C. When using eLONGase™ enzyme mix (Gibco BRL), each reaction contained 100 ng genomic DNA, 20 pmol primer, 60 mM Tris·SO_4 (pH 9.1), 18 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 1 mM $MgCl_2$, 200 µM each dNTP, and 1 U eLONGase enzyme mix. The cycling conditions were identical to those for the Taq reactions, except that extension was at 68°C for 5 min per cycle.

After cycling, amplification products were combined, purified using the Wizard® PCR Preps purification system, and eluted in 50 µl distilled, deionised water. 5.5 µl of this combined single primer PCR product was sequenced using 1.6 pmol of internal primer and BigDye™ terminator chemistry on an ABI PRISM® 377 DNA Sequencer.

2.13 DNA sequence analysis

2.13.1 Restriction enzyme sites, conceptual translation, and open reading frame analysis

Identification of restriction enzyme sites, conceptual translation, and analysis of open reading frames (ORFs) were performed using tools included in the DNA Strider™ 1.2 application. Conceptual translation was also performed using the Translate tool (http://www.expasy.ch/tools/dna.html) at the ExPASy Molecular Biology Server (http://www.expasy.ch).

2.13.2 BLAST searches

Searches of GenBank nucleic acid and protein sequence databases were performed using the ungapped BLAST (Altschul *et al.*, 1990) and gapped BLAST 2.0 (Altschul *et al.*, 1997) servers at the National Centre for Biotechnology Information (NCBI) WWW site (http://www.ncbi.nlm.nih.gov/). BLASTn searches use query nucleotide sequences to search nucleotide sequence databases. BLASTp searches use query amino acid sequences to search protein sequence databases. BLASTx searches translate an input nucleotide sequence in all six reading frames, and use the deduced amino acid sequence to search protein sequence databases.

2.14 Phylogenetic analysis

2.14.1 Sequence sources

The full names, accession numbers, and source references of all retroviral sequences used for sequence comparisons and phylogenetic analyses are listed in Appendix 1.

2.14.2 Sequence alignments

Alignments of nucleotide and amino acid sequences were either performed manually using a previously published alignment as a guide, or were

performed using the CLUSTAL X windows interface (Thompson *et al.*, 1997) for the CLUSTAL W sequence alignment program (Thompson *et al.*, 1994). Nucleotide sequence alignments using CLUSTAL X were normally performed using the default parameters for gap opening (10.0), gap extension (0.05), and transition weighting (0.5). Amino acid sequence alignments were also performed using the default gap opening and extension parameters, and the BLOSUM series protein weight matrix (Henikoff and Henikoff, 1992).

2.14.3 Construction of phylogenetic trees

Alignments generated using CLUSTAL X were used to construct phylogenetic trees. Distance-based trees were constructed using the Neighbour-joining (NJ) approach (Saitou and Nei, 1987) within the CLUSTAL X application (Thompson *et al.*, 1997). Bootstrap analysis, with 1000 replicates, was performed on all NJ trees. Maximum parsimony trees were constructed using PAUP* Version 4.0b2 (Swofford, 1998), and the optimal tree was found using the branch-and-bound approach (Hendy and Penny, 1982). Both NJ and maximum parsimony trees were displayed and manipulated using TreeView V1.5.2 (Page, 1998).

Chapter 3

Preliminary Investigations: PCR From Genomic DNA, Reverse Transcriptase Assays on Blood Plasma, and RT-PCR from Blood Plasma RNA

3.1 Introduction

3.1.1 Retrovirus detection

A number of methods can be used to detect and characterise retroviruses. Traditional methods include monitoring cells in culture for retrovirus-induced transformation or cytopathic effect, electron microscopy to detect virus particles, immunological detection of virus antigens or antiviral antibodies, and nucleic acid hybridisation techniques. Although many of these methods are still used for the detection of known retroviruses, they are usually only applied to cases in which clinical symptoms of retroviral disease are present, and they are generally of limited use for the detection of broad ranges of retroviruses (see 1.2.10 Retrovirus detection).

3.1.1.1 Detection of retroviruses by PCR using degenerate primers

Recently, PCR using degenerate primers derived from conserved regions of retroviral genomes has been used to detect and characterise retroviruses. The primers used in this study, polF and polR, were derived from two highly conserved regions of the RT protein (FIG. 3.1). They were identical to those used by Herrmann and Kalden (1994) to PCR-amplify human endogenous retroviruses and their transcripts, and were almost identical to those shown to amplify a region from the reverse transcriptase gene of all retroviruses tested (Shih *et al.*, 1989). Endogenous retroviruses are present in the genomes of all vertebrates (see 1.3 Biology of endogenous retroviruses), and are invariably amplified using primers derived from conserved retroviral regions (Shih *et al.*, 1989; Wichman and Van Den Bussche, 1992; Tristem, 1996; Herniou *et al.*, 1998). In addition, the use of such primers to RT-PCR amplify RNA from vertebrate cells often results in the detection of transcripts of endogenous retroviruses (Herrmann and Kalden, 1994; Lefebvre *et al.*, 1995; DeHaven *et al.*, 1998). It has also been possible to characterise retroviruses which have been purified away from cellular nucleic acids, such as those isolated from cell culture supernatants (Donehower, *et al.*, 1990; Perron *et al.*, 1997).

```
HTLV-1  GTACTACCCCAAGGG TTTAAA -78bp- CTTCAG TACATGGATGAC
         V  L  P  Q  G   F  K          L  Q   Y  M  D  D

HTLV-2  GTCCTTCCACAGGGG TTTAAA -78bp- GTCCAA TACATGGATGAC
         V  L  P  Q  G   F  K          V  Q   Y  M  D  D

HIV-1   GTGCTTCCACAGGGA TGGAAA -78bp- TATCAA TACATGGATGAT
         V  L  P  Q  G   W  K          Y  Q   Y  M  D  D

Pr-RSV  GTCTTGCCCCAAGGG ATGACC -78bp- TTGCAT TATATGGATGAT
         V  L  P  Q  G   M  T          L  H   Y  M  D  D

Mo-MLV  AGACTCCCACAGGGT TTCAAA -78bp- CTACAG TACGTGGATGAC
         R  L  P  Q  G   F  K          L  Q   Y  V  D  D

BLV     GTCCTACCTCAAGGC TTCATT -78bp- GTGTCC TATATGGACGAT
         V  L  P  Q  G   F  I          V  S   Y  M  D  D

EIAV    TGTTTACCACAAGGA TTCGTG -78bp- TATCAA TATATGGATGAT
         C  L  P  Q  G   F  V          Y  Q   Y  M  D  D

MMTV    GTTTTGCCCCAGGGT ATGAAA -78bp- GTGCAT TACATGGATGAC
         C  L  P  Q  G   F  V          Y  Q   Y  M  D  D

MPMV    GTTTTACCACAAGGT ATGGCC -78bp- ATACAT TACATGGATGAC
         V  L  P  Q  G   M  A          I  H   Y  M  D  D

5' ctcgaattcGTNYTNCCNCARGG 3'                3' ATRTACCTRCTRttcgaactc 5'
        polF primer                                  polR primer
```

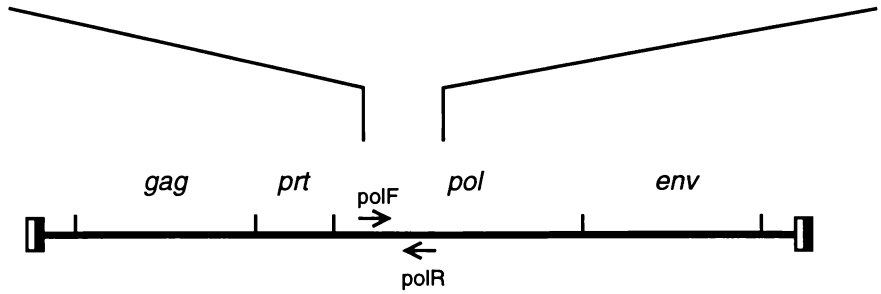


FIGURE 3.1 Degenerate PCR primers derived from highly conserved regions within the *pol* gene. Nucleotide and amino acid sequences from the 5' end of the *pol* gene (and the N-terminal region of the reverse transcriptase protein) of several retroviruses are aligned. PCR primers polF and polR, shown below the aligned sequences, were used in this study, and are identical to those used by Herrmann *et al.* (1994). Primer nucleotides in lowercase letters represent noncomplementary 5' extensions that contain recognition sequences for *Eco*RI (polF) and *Hin*dIII (polR). N, Y, and R in the primer sequences indicate all four nucleotides, pyrimidines, and purines, respectively. Full names of retroviruses are listed in Appendix 1. Adapted from Shih *et al.* (1989) and Donehower *et al.* (1990).

3.1.1.2 Reverse transcriptase assays

Ultrasensitive PCR-based reverse transcriptase assays have been developed recently (Silver *et al.*, 1993; Pyra *et al.*, 1994; Heneine *et al.*, 1995; Maudru and Peden, 1997; Arnold *et al.*, 1998). Such assays can detect the reverse transcriptase activity of as few as 10 particles of a wide

range of retroviruses. Consequently, retroviruses purified from biological tissues and fluids can now be detected, a situation which was not previously possible with traditional reverse transcriptase assays (see Heneine *et al.*, 1995; Yamamoto *et al.*, 1996). In this study, I used the product enhanced reverse transcriptase (PERT) assay (Pyra *et al.*, 1994), which is outlined in FIG. 3.2.

3.1.2 Phylogenetic analysis

Phylogenetic analysis of retroviral nucleic acid and protein sequence data can be used to estimate the evolutionary history of a retrovirus, and also constitutes one criterion for the classification of retroviruses (Myers, 1997; and see 1.2.4 Retrovirus classification). The phylogenetic relationships of distantly related retroviruses is normally determined using the sequence of the polymerase domain of the RT protein. The polymerase domain is the most highly conserved protein region amongst retroviruses (McClure *et al.*, 1988), and possesses a number of highly conserved motifs and residues that

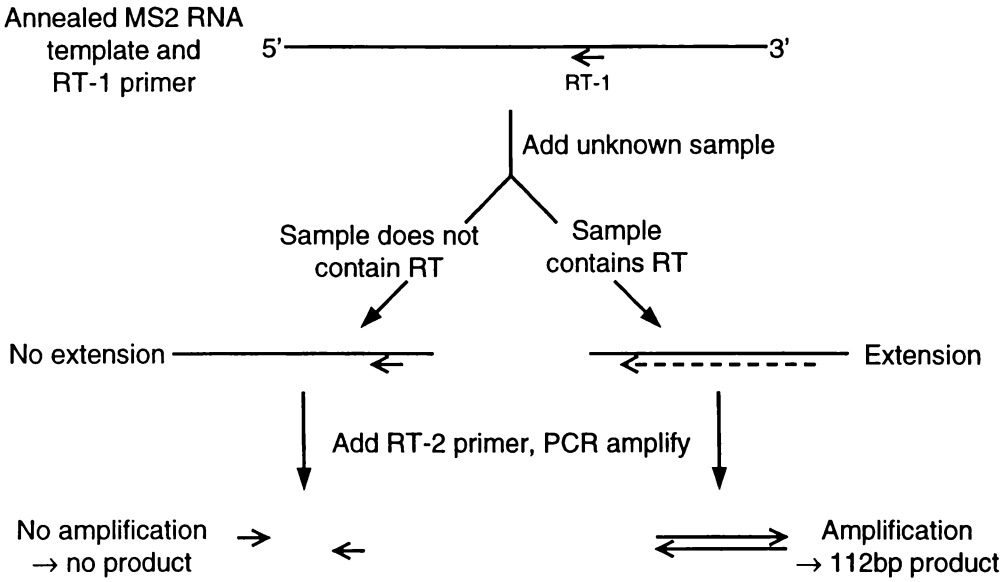


FIGURE 3.2 The product enhanced reverse transcriptase (PERT) assay of Pyra *et al.* (1994). The unknown sample is added to annealed MS2 RNA template and RT-1 primer. If the sample contains reverse transciptase (RT) activity, the MS2 template is reverse transcribed, and the DNA copy is subsequently amplified using RT-1 and RT-2 primers to produce a 112bp product. If the unknown sample lacks RT activity, the MS2 RNA template is not reverse transcribed, and no product is generated by the PCR step.

aid alignment (Doolittle *et al.*, 1989; Xiong and Eickbush, 1990; McClure *et al.*, 1993). Extended regions of sequence can be used for the analyses of more closely related retroviruses, as long as homologous regions can be identified and aligned. Although the evolutionary relationships inferred from phylogenetic analyses using other retroviral proteins are normally similar (or identical) to those derived using the RT protein, they occasionally differ. Such differences normally reflect recombination of portions of the retroviral genome between distantly related retroviruses. For example, the simian type D retroviruses, have a *pol* gene related to those of the type B and avian type C retroviruses, but an *env* gene more closely related to those of the mammalian type C retroviruses (York *et al.*, 1992).

3.1.2.1 Sequence alignment

The first step in phylogenetic analysis is the alignment of the sequences. Closely related sequences, with readily identified regions of homology, can often be aligned manually. However, alignment of large numbers of sequences, and alignment of distantly related sequences are usually performed using computer programs. A number of computer programs are currently available for alignment of nucleotide and amino acid sequences (see Hillis *et al.*, 1996). Although these programs differ in their approaches, they were all written with the same aim in mind — to correctly align two or more sequences (nucleotide or amino acid) so that the nucleotide or amino acid at each position is aligned with the homologous nucleotides or amino acids of the other sequences.

Alignment programs assign scores to aligned sequences. Sequences that are 100% identical and perfectly aligned have the maximum possible score. Sequences that differ (ie. they have nucleotide substitutions or insertions/deletions relative to each other) or are imperfectly aligned have a score that is below the maximum. Thus, alignment programs align two or more sequences such that their alignments produce the highest score. Nucleotide or amino acid substitutions, or introduction of gaps into sequences (representing insertion or deletion events in the history of a sequence), can usually be penalised at the discretion of the user. Nucleotide

or amino acid substitutions may be assigned the same cost regardless of the type of substitution. Alternatively, weight matrices may be applied according to the type of nucleotide substitution (transitions may be penalised less than transversions) or amino acid substitution (substitution with a chemically similar amino acid may be penalised less than substitution with one that is chemically very different). Similarly, introduction of new gaps into sequences can be assigned a different (normally higher) penalty than extension of a pre-existing gap (see Hillis *et al.*, 1996).

3.1.2.2 Tree construction

The second step in phylogenetic analysis involves inference of phylogenetic relationships based on the aligned sequences. Again, numerous computer programs are available for phylogenetic analysis, and again, they all differ in their approaches (see Swofford *et al.*, 1996). The methods can be divided into two general types: distance-based methods; and character-based methods. Distance-based methods, such as the neighbour joining (NJ) method (Saitou and Nei, 1987), calculate the pairwise distances between all of the sequences in an alignment, and construct trees based on the pairwise distances. Weight matrices, similar to those described above, can be factored into the pairwise distance calculations. In a strict sense, distance methods are phenetic rather than cladistic — that is, sequences are grouped based merely on their overall similarities (Myers, 1997). Character-based methods, such as maximum parsimony (Fitch, 1971) and maximum likelihood (Felsenstein, 1981), attempt to determine phylogenetic relationships by analysing each character in the alignment separately. Hence, character-based methods are truly cladistic — their trees represent (or attempt to represent) the actual evolutionary pathways leading to the current sequences (Myers, 1997). Character-based methods are generally regarded as being superior to distance-based methods for elucidating evolutionary relationships (Swofford *et al.*, 1996; Myers, 1997). Character-based methods are far more computationally intensive than distance-based methods, however, and, when large numbers of sequences are being analysed, distance-based methods often provide good, although less accurate, estimates of the true phylogenies.

3.1.2.3 Bootstrap analysis

Bootstrap analysis (Felsenstein, 1985), a method of statistical resampling, is often used to estimate the confidence that can be placed in a given tree. When performing a bootstrap analysis, a program generates a number of pseudosamples (normally 100–1000) by randomly selecting data points in an alignment. The number of data points in each pseudosample is normally less than or equal to the number of data points in the original alignment. The data points are replaced after each sample is taken (sampling with replacement) so that some data points in the original alignment will not be represented in the pseudosample, whereas other data points are present twice or more. For each pseudosample, a new tree is generated, and the frequency with which various branching patterns are observed within these trees is noted (Lake and Moore, 1998). Thus, bootstrap values adjacent to nodes represent the number of times the sequences branching from those nodes were grouped together in the pseudosample trees (Swofford *et al.*, 1996).

3.2 Results

3.2.1 Detection of retrovirus-related sequences in the genomes of possums

3.2.1.1 PCR using degenerate primers from the pol *gene*

PCR, using the polF and polR primers, was performed on genomic DNA isolated from possum blood, liver, and mammary cells. The results of a typical PCR are shown in FIG. 3.3. A product of approximately 130bp was generated for all possum genomic DNA tested. This corresponds to the size expected (135bp) of the product of amplification of the reverse transcriptase gene using the polF and polR primers (see FIG. 3.1 and Shih *et al.*, 1989).

3.2.1.2 Cloning and sequencing of PCR products

The product of PCR from the mammary gland DNA of one possum (FIG. 3.3, possum 1, lane M) was cloned into pBluescript II KS(+). In total, 12 clones

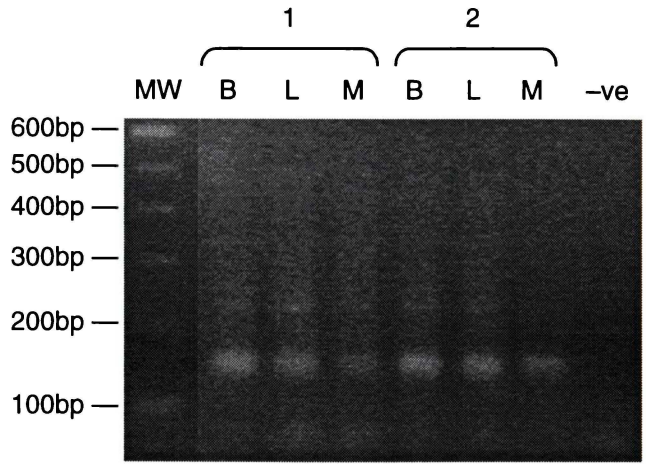


FIGURE 3.3 PCR-amplification of possum genomic DNA using polF and polR primers derived from the *pol* gene. Genomic DNA (500ng) from blood (B), liver (L), and mammary (M) cells of two possums (1, 2) was PCR amplified in the presence of 1.25 U AmpliTaq (Perkin Elmer), 2.5 mM Mg^{2+}, 60 mM TMAC using 50 pmol each of polF and polR primers (see FIG. 3.1) and the following cycling conditions: Cycles 1–3: 94°C/1 min; 37°C/1 min; 72°C/2 min; Cycles 4–40: 94°C/30 sec; 50°C/30 sec; 72°C/1 min; final extension at 72°C/10 min. The PCR products were electrophoresed on a 2% agarose gel. MW: molecular weight standard (sizes shown at left). –ve: negative (water) PCR control.

were isolated which contained retrovirus-related sequences. The nucleotide and deduced amino acid sequences of those clones are shown in FIG. 3.4.

Of the twelve clones, eight (TvPol1, TvPol2, TvPol7, TvPol8, TvPol9, TvPol10, TvPol11, and TvPol12) possessed uninterrupted open reading frames over their entire lengths. Of the remaining four clones, clone TvPol6 contained an in-frame stop codon, clone TvPol7 required the addition of one nucleotide to maintain the open reading frame, and clones TvPol3 and TvPol4 both required the addition of two nucleotides to maintain the open reading frame, and also contained in-frame stop codons (FIG. 3.4). When the primer sequences were omitted from comparisons, the nucleotide sequences of two of the clones, TvPol7 and TvPol8, were identical. Two more clones, TvPol9 and TvPol10, were identical at the amino acid sequence level.

Searches (BLASTp) of the Genbank database were performed using the deduced amino acid sequences in FIG. 3.4. The sequences corresponding to

```
          polF.
TvPol1    GTACTGCCTCAGGGCATGGCCAACAGCCCCACGGTATGTCAGATGTACGTAGCAGCCATC
           V  L  P  Q  G  M  A  N  S  P  T  V  C  Q  M  Y  V  A  A  I
TvPol2    GTTTTGCCCCAGGGTATGAAAAATAGTCCCACCTTATGTCAATGGTATGTTCAACAAGTT
           V  L  P  Q  G  M  K  N  S  P  T  L  C  Q  W  Y  V  Q  Q  V
TvPol3    GTATTACCGCAAGGGATGAAAAATAGTCCTGCCATGTGTCAACAATATGTTGCACAAGCT
           V  L  P  Q  G  M  K  N  S  P  A  M  C  Q  Q  Y  V  A  Q  A
TvPol4    GTGTTGCCGCAGGGGATGAAAAATAGTCCCACCATGTGTCAACAATATGTTGCACAAGCT
           V  L  P  Q  G  M  K  N  S  P  T  M  C  Q  Q  Y  V  A  Q  A
TvPol5    GTGTTGCCGCAGGG?ATGAAAAATAGTCCCACTATGTGTCAACAATATGTTGCACAAGTT
           V  L  P  Q  G  M  K  N  S  P  T  M  C  Q  Q  Y  V  A  Q  V
TvPol6    GTGTTGCCTCAGGGGATGAAAAATAGTCCCACCATGTGTCAACAGTATGCTGCACAAACT
           V  L  P  Q  G  M  K  N  S  P  T  M  C  Q  Q  Y  A  A  Q  T
TvPol7    GTTCTCCCGCAGGGATATCTGCATAGCCCTACTATTTGTCATAGGACTGTAGCTCAGCAT
           V  L  P  Q  G  Y  L  H  S  P  T  I  C  H  R  T  V  A  Q  H
TvPol8    GTGCTGCCGCAGGGATATCTGCATAGCCCTACTATTTGTCATAGGACTGTAGCTCAGCAT
           V  L  P  Q  G  Y  L  H  S  P  T  I  C  H  R  T  V  A  Q  H
TvPol9    GTACTTCCGCAAGGATATATTCATAGCCCAACTATTTGCCACAGGATTGTAGCTGAGCAT
           V  L  P  Q  G  Y  I  H  S  P  T  I  C  H  R  I  V  A  E  H
TvPol10   GTGCTACCTCAGGGATATATTCATAGCCCAACTATTTGCCACAGGATTGTAGCTGAACAT
           V  L  P  Q  G  Y  I  H  S  P  T  I  C  H  R  I  V  A  E  H
TvPol11   GTATTACCTCAGGGATATCTGTATAATCCAACTATTTGTCATAGGATTGTGGCTGAATAT
           V  L  P  Q  G  Y  L  Y  N  P  T  I  C  H  R  I  V  A  E  Y
TvPol12   GTGCTACCGCAGGGGTTTCGAGATAGCCCTCATTTATTTGGCCAGGCTTTGGGAAAAGAT
           V  L  P  Q  G  F  R  D  S  P  H  L  F  G  Q  A  L  G  K  D

                                                          polR
TvPol1    ATACAGCCAGTGAGAGAACAATTCCCCGGCGCCATGATTATACACTACATGGACGAC
           I  Q  P  V  R  E  Q  F  P  G  A  M  I  I  H  Y  M  D  D
TvPol2    ATTCGCCCTGTCCACGAGCGCTATCCAACTGTGCTTTTAATACATTACATGGACGAC
           I  R  P  V  H  E  R  Y  P  T  V  L  L  I  H  Y  M  D  D
TvPol3    TTACAACCTATAA??GAGAAATATCCCTAAGCTTATCTCATTCATTATATGGACGAC
           L  Q  P  I  X  E  K  Y  P  *  A  Y  L  I  H  Y  M  D  D
TvPol4    TTACAACCTATAA??GAGAAATATCCCTAAGCTTATCTCATTCATTACATGGACGAC
           L  Q  P  I  X  E  K  Y  P  *  A  Y  L  I  H  Y  M  D  D
TvPol5    TTACAACCTATAAGAGAGAAATATCCCAAAACTTATCTCATTCATTACATGGACGAC
           L  Q  P  I  R  E  K  Y  P  K  T  Y  L  I  H  Y  M  D  D
TvPol6    TTACAAGCTATAAGAGAGAAATAGCCAAAAGCTTATCTCACTCATTACATGGACGAC
           L  Q  A  I  R  E  K  *  P  K  A  Y  L  T  H  Y  M  D  D
TvPol7    TTGGATGAATTAGAA------CTGCCTGGTATACAGCTTACCCACTACATGGACGAC
           L  D  E  L  E  -  -  L  P  G  I  Q  L  T  H  Y  M  D  D
TvPol8    TTGGATGAATTAGAA------CTGCCTGGTATACAGCTTACCCACTATATGGACGAC
           L  D  E  L  E  -  -  L  P  G  I  Q  L  T  H  Y  M  D  D
TvPol9    TTGGATGAATTAAAG------CTACCTGGTGTACAGCTTACACATTATATGGACGAC
           L  D  E  L  K  -  -  L  P  G  V  Q  L  T  H  Y  M  D  D
TvPol10   TTGGATGAATTAAAG------TTACCTGGTGTACAGCTTACACATTACATGGACGAC
           L  D  E  L  K  -  -  L  P  G  V  Q  L  T  H  Y  M  D  D
TvPol11   TTGGATAAATTAAAG------TCACCTGGTGTGCAGCTTACCCACTACATGGACGAC
           L  D  K  L  K  -  -  S  P  G  V  Q  L  T  H  Y  M  D  D
TvPol12   TTAAGGGACCTGAAA------TTAACTGACAGTAGCTTAATTCAGTATATGGACGAC
           L  R  D  L  K  -  -  L  T  D  S  S  L  I  Q  Y  M  D  D
```

FIGURE 3.4 Nucleotide and deduced amino acid sequences of retrovirus-related clones from possum genomic DNA. The products of amplification using polF and polR primers on genomic DNA from possum mammary gland (see FIG. 3.3, possum 1, lane M) were purified, digested with *Hin*dIII and *Eco*RI restriction enzymes (which cut within the polF and polR primers, respectively), and cloned into pBluescript II KS(+) (Stratagene) as described in Chapter 2 – Materials and Methods. Twelve clones containing retrovirus-related inserts (as determined by manual sequencing and BLASTx searches) were sequenced using an ABI PRISM® 377 automated DNA sequencer. The deduced amino acid sequences are shown below the nucleotide sequences. The regions covered by polF and polR are indicated. x indicates any amino acid; * indicates a stop codon; - indicates a gap introduced for alignment; ? indicates an unknown nucleotide (coded as N for translation) added to correct frameshift.

the polF and polR primers were removed for the database searches, so that only the sequences between the primers were used. Alignments of both nucleotide and amino acid sequences were also performed using the CLUSTAL X program (see 2.14.1 Sequence alignments). Together, the database search results and the alignments allowed the clones to be divided into three groups, as shown in FIG. 3.5. All three groups possessed similarity to marsupial sequences previously reported by Herniou *et al.* (1998). The first group comprised clones TvPol1–TvPol6, which were highly similar to the HERV-K family of endogenous retroviruses in humans, the rodent intracisternal A-type particles (IAPs) of rodents, and the type D retroviruses. This group also bore strong homology to the RV-stripe faced dunnartI sequence reported by Herniou *et al.* (1998) (FIG. 3.5). The second group

```
Group 1
                polF                                        polR
TvPol1          VLPQGMANSPTVCQMYVAAIIQPVREQFPGAMIIHYMDD
TvPol2          VLPQGMKNSPTLCQWYVQQVIRPVHERYPTVLLIHYMDD
TvPol3          VLPQGMKNSPAMCQQYVAQALQPIXEKYP*AYLIHYMDD
TvPol4          VLPQGMKNSPTMCQQYVAQALQPIXEKYP*AYLIHYMDD
TvPol5          VLPQGMKNSPTMCQQYVAQVLQPIREKYPKTYLIHYMDD
TvPol6          VLPQGMKNSPTMCQQYAAQTLQAIREK*PKAYLTHYMDD
RV sfdI               MANSPILCQ??VDKVLAPVRNLYPNVYMLH
HERV-K10              MLNSPTICQTFVGRALQPVREKFSDCYIIH
CHIAP                 MANSPTICQLYVQKALEPVRKQFTSMIMIH
MPMV                  MANSPTLCQKYVATAIHKVRHAWKQMYIIH

Group 2
                polF                                        polR
TvPol7          VLPQGYLHSPTICHRTVAQHLDELELP-GIQLTHYMDD
TvPol8          VLPQGYLHSPTICHRTVAQHLDELELP-GIQLTHYMDD
TvPol9          VLPQGYIHSPTICHRIVAEHLDELKLP-GVQLTHYMDD
TvPol10         VLPQGYIHSPTICHRIVAEHLDELKLP-GVQLTHYMDD
TvPol11         VLPQGYLYNPTICHRIVAEYLDKLKSP-GVQLTHYMDD
RV cpossum            DLHSPTICHRIVAEHLDKLKLA-GIQLTH
MuERV-L               YINSPALCHNLVRRDLDRLDLPQSITLVH
HERV-L                YINSPALCHNLIQSDLDHFSLPQDITLVH

Group 3
                polF                                        polR
TvPol12         VLPQGFRDSPHLFGQALGKDLRDLKLTDSSLIQYMDD
RV sfdII              FRDSPHLFGQALAKDLRDLELKGSCLLQ
ERV-9                 FRDSPHLFGQALAKDLGHFSSPGTLVLQ
HERV-H                FRDNPHYFSQALSHDLLSFHSSASHLIQ
```

FIGURE 3.5 Three groups of retrovirus-related clones. The amino acid sequences in FIG. 3.4 were used to perform BLASTp searches and alignments as described in the text, which allowed the clones to be divided into three groups. RV sfdI, RV sfdII, and RV cpossum correspond to the RV stripe-faced dunnartI, RV stripe-faced dunnartII, and RV common possum sequences reported by Herniou *et al.* (1998). Details of remaining retrovirus sequences are listed in Appendix 1. x indicates any amino acid; * indicates a stop codon; - indicates a gap introduced for optimal alignment; ? indicates unknown amino acids due to frameshifts.

comprised clones TvPol7–TvPol11, which were most closely related to MuERV-L, a murine endogenous retrovirus with similarity to the human endogenous element HERV-L (Bénit *et al.*, 1997). This group also had strong homology to the RV-common possum sequence reported by Herniou *et al.* (1998) (FIG. 3.5). Database searches using the amino acid sequence of TvPol11 internal to the polF and polR primers failed to detect any similarity to any retrovirus-related sequences. However, comparison of their sequences shows that clone TvPol11 is clearly related to the other clones from group 2 (FIG. 3.5). The differences in amino acid sequences adjacent to the polF primer between clone TvPol11 and the clones from group 2 may account for the anomalous database search results. The final group comprised only one clone, TvPol12, which showed most similarity to the human elements ERV-9 and HERV-H, and is very similar to RV-stripe faced dunnartII (Herniou *et al.*, 1998) (FIG. 3.5).

3.2.2 Detection of reverse transcriptase activity in possum blood

Particulate material was prepared from filtered blood plasma as described in Chapter 2 – Materials and Methods. Reverse transcriptase activity in these preparations was detected using the product enhanced reverse transcriptase (PERT) assay (Pyra *et al.*, 1994) (see 2.11 Product enhanced reverse transcriptase [PERT] assay). Results of the PERT assay on blood samples from four possums are shown in FIG. 3.6. The expected amplification product from samples containing RT activity is 112bp. As shown, reverse transcriptase activity appears to be present in the blood of all four possums. Indeed, the blood of all possums subsequently tested contained reverse transcriptase activity as determined by the PERT assay.

3.2.3 Detection of retroviral RNA in possum blood

3.2.3.1 RT-PCR using degenerate primers from pol *gene*

The results of the PERT assay suggested that retroviral particles were present in the blood of all possums. RT-PCR, using primers polF and polR, was conducted on RNA isolated from filtered possum blood plasma in an

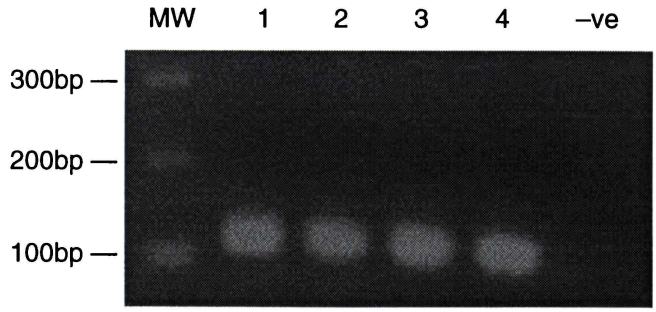


FIGURE 3.6 Reverse transcriptase activity in possum blood plasma. Blood plasma was collected from four possums (1–4), and filtered as described in Chapter 2 – Materials and Methods. Particulate material in the filtered plasma was pelleted by ultracentrifugation, resuspended and assayed for reverse transcriptase activity using the product enhanced reverse transcriptase (PERT) assay (see Chapter 2 – Materials and Methods). The PERT assay products were electrophoresed on a 2% agarose gel. MW: molecular weight standard (sizes shown at left). –ve: negative (buffer A without reverse transcriptase) control.

attempt to obtain sequence information from the putative retrovirus particles. The results of RT-PCR from serum RNA of three possums are shown in FIG. 3.7. All three possums tested yielded RT-PCR fragments of the appropriate size (~130bp) for a retroviral reverse transcriptase gene. No RT-PCR product was generated in the Superscript II –ve controls, which indicates that the products were generated from RNA and not contaminating DNA.

3.2.3.2 Cloning and sequencing of RT-PCR products

The RT-PCR product generated from plasma RNA of one possum (FIG. 3.7, possum 3) was cloned into pBluescript II KS(+). Five retrovirus-related clones were isolated and sequenced. Their nucleotide sequences and deduced amino acid sequences are shown in FIG. 3.8a. All of the clones were very similar. Indeed, the only nucleotide differences occurred within the primer regions (FIG. 3.8a). The amino acid sequences of all the clones were almost identical, the only difference being a L→W change in the polF primer region of clone TvPol16 (FIG. 3.8b). The sequence differences in the primer regions may not represent real differences in the sequences of the templates, given that polF and polR are degenerate primers. None of the sequences were interrupted by stop codons or frameshift mutations.

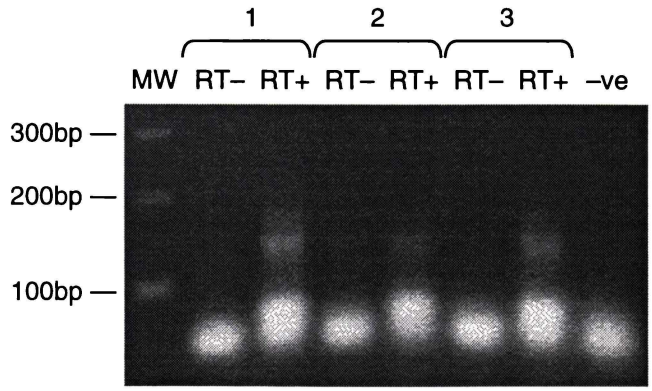


FIGURE 3.7 RT-PCR amplification of RNA from possum blood plasma using polF and polR primers from the *pol* gene. RNA was isolated from blood plasma of three possums (1–3), incubated in the presence (RT+) or absence (RT–) of Superscript II reverse transcriptase (Life Technologies), and PCR amplified in the presence of 60mM TMAC. Reverse transcription conditions were as described in Chapter 2 – Materials and Methods, with 20 pmol polR primer, and incubation at 37°C for 4hr. PCR cycling conditions: Initial RNA degradation (70ng RNase A), 37°C/30 min; Cycles 1–10: 94°C/1 min, 37°C/1 min, 72°C/1 min; Cycles 11–30: 94°C/15 sec, 55°C/30 sec, 72°C/30 sec; final extension 72°C/2 min. The PCR products were electrophoresed on a 2.5% agarose gel. MW: molecular weight standard (sizes shown at left). –ve: negative (water) PCR control).

A database search (BLASTn) was performed using the nucleotide sequence of TvPol13 internal to the polF and polR primers. No retrovirus-related sequences were detected. However, a search (BLASTp) using the amino acid sequence over the same range revealed considerable similarity to the type B/type D Jaagsiekte sheep retrovirus (JSRV), the simian type D retroviruses from Old World (MPMV, SRV-1, SRV-2) and New World (SMRV-H) monkeys, the type B retrovirus (MMTV), the endogenous human elements (HERV-K), and the rodent IAP sequences (FIG. 3.8b). They also bore significant similarities to the group 1 clones amplified from possum genomic DNA (see FIG. 3.5) and the RV stripe-faced dunnartl sequence reported by Herniou *et al.* (1998) (FIG. 3.8b).

3.2.4 Phylogenetic analysis of possum retroviral sequences

Phylogenetic analysis was performed using the amino acid sequences of the possum clones from genomic DNA and plasma RNA. The possum sequences were aligned with the corresponding regions of sequence from

(a)

```
                polF
TvPol13  GTGCTGCCTCAGGGCATGACCAACAGCCCTACCCTCTGCCAGAAGTATGTTGCCCAGACA
          V  L  P  Q  G  M  T  N  S  P  T  L  C  Q  K  Y  V  A  Q  T
TvPol14  GTGTTGCCTCAGGGCATGACCAACAGCCCTACCCTCTGCCAGAAGTATGTTGCCCAGACA
          V  L  P  Q  G  M  T  N  S  P  T  L  C  Q  K  Y  V  A  Q  T
TvPol15  GTGTTGCCCCAGGGCATGACCAACAGCCCTACCCTCTGCCAGAAGTATGTTGCCCAGACA
          V  L  P  Q  G  M  T  N  S  P  T  L  C  Q  K  Y  V  A  Q  T
TvPol16  GTCTGGCCTCAGGGCATGACCAACAGCCCTACCCTCTGCCAGAAGTATGTTGCCCAGACA
          V  W  P  Q  G  M  T  N  S  P  T  L  C  Q  K  Y  V  A  Q  T
TvPol17  GTATTGCCTCAGGGCATGACCAACAGCCCTACCCTCTGCCAGAAGTATGTTGCCCAGACA
          V  L  P  Q  G  M  T  N  S  P  T  L  C  Q  K  Y  V  A  Q  T

                                                          polR
TvPol13  ATCGACCCCTTTCGCTTACAATTTCCACAACTTTATATCATTCACTATATGGACGAC
          I  D  P  F  R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D
TvPol14  ATCGACCCCTTTCGCTTACAATTTCCACAACTTTATATCATTCACTACATGGACGAC
          I  D  P  F  R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D
TvPol15  ATCGACCCCTTTCGCTTACAATTTCCACAACTTTATATCATTCACTACATGGACGAC
          I  D  P  F  R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D
TvPol16  ATCGACCCCTTTCGCTTACAATTTCCACAACTTTATATCATTCACTATATGGACGAC
          I  D  P  F  R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D
TvPol17  ATCGACCCCTTTCGCTTACAATTTCCACAACTTTATATCATTCACTACATGGACGAC
          I  D  P  F  R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D
```

(b)

```
                polF                                    polR
TvPol13  VLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
TvPol14  VLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
TvPol15  VLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
TvPol16  VWPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
TvPol17  VLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
TvPol1   VLPQGMANSPTVCQMYVAAIIQPVREQFPGAMIIHYMDD
TvPol2   VLPQGMKNSPTLCQWYVQQVIRPVHERYPTVLLIHYMDD
TvPol3   VLPQGMKNSPAMCQQYVAQALQPIXEKYP*AYLIHYMDD
RV sfdI       MANSPILCQ??VDKVLAPVRNLYPNVYMLH
JSRV          MTNSPTLCQKFVATAIAPVRQRFPQLYLVH
MPMV          MANSPTLCQKYVATAIHKVRHAWKQMYIIH
SMRV-H        MANSPTLCQKFVAAAIAPVRSQWPEAYILH
MMTV          MKNSPTLCQKFVDKAILTVRDKYQDSYIVH
HERV-K        MLNSPTICQTFVGRALQPVREKFSDCYIIH
CHIAP         MANSPTICQLYVQKALEPVRKQFTSMIMIH
```

FIGURE 3.8 Nucleotide and deduced amino acid sequences of retrovirus related clones from possum blood plasma (a), and alignment of their amino acid sequences with those of related retroviruses (b). The product of RT-PCR on blood plasma RNA of one possum (FIG 3.7, possum 3, RT+) was purified, restriction enzyme digested, and cloned into pBluescript II KS(+) as described in the legend to FIG. 3.4. Five retrovirus-related clones were isolated, and ssDNA preparations from these clones were sequenced manually and using an ABI PRISM® 377 automated sequencer. TvPol1–TvPol3 are clones from possum genomic DNA (see Fig. 3.4). RV sfdI corresponds to the RV stripe-faced dunnartI sequence reported by Herniou *et al.* (1998). Details of remaining retrovirus sequences are listed in Appendix 1. ? indicates unknown amino acids due to frameshift.

the marsupial sequences previously reported by Martin *et al.* (1997), Herniou *et al.* (1998), and Martin *et al.* (1999), and from representatives of all of the retroviral genera. The alignment was performed using the CLUSTAL X program as described in Chapter 2 – Materials and Methods, and is shown in

Appendix 3. The alignment was used to construct a neighbour-joining tree (FIG. 3.9) using the CLUSTAL X program, which was subsequently tested by bootstrap analysis, as described in Chapter 2 – Materials and Methods.

Because of the short lengths of the sequences involved, the bootstrap support for many of the relationships was low, and this limits the confidence

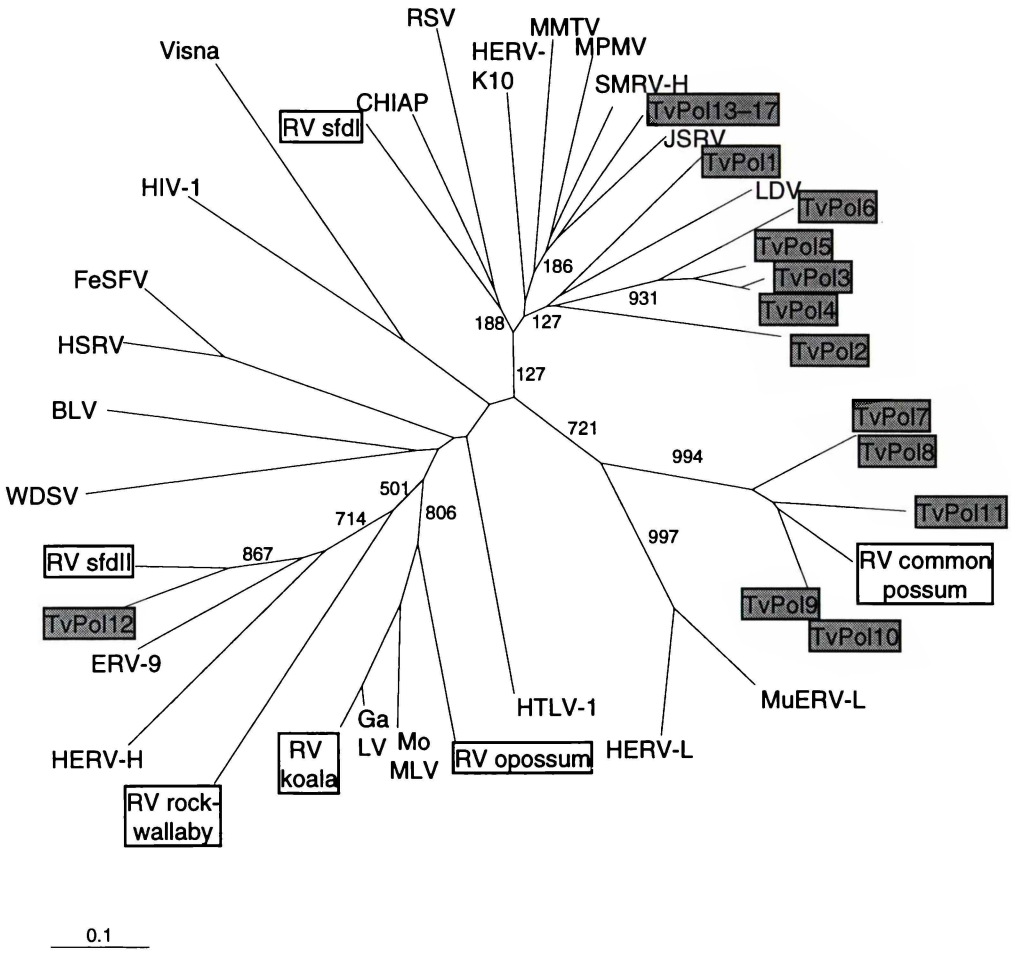


FIGURE 3.9 Neighbour-joining tree based on the amino acid sequences of a short region of the Pol protein. The possum retroviral sequences (boxed and shaded) from possum genomic DNA (TvPol1–TvPol12) and plasma RNA (TvPol13–17) were aligned with those of other marsupial retroviral sequences (boxed) and representatives of all of the retroviral genera. The alignment was performed using CLUSTAL X as described in Chapter 2 – Materials and Methods, and is shown in Appendix 3. The RV rock wallaby sequence was reported by Martin *et al.* (1997). The RV common possum, RV sfdI, and RV sfdII sequences are those from the common possum and stripe-faced dunnart reported by Herniou *et al.* (1998). The RV koala and RV opossum sequences were reported by Martin *et al.* (1999). Details of all other retroviral sequences are listed in Appendix 1. Bootstrap support (1000 replicates) for some groups are indicated. Branch lengths are proportional to the genetic distance between sequences, as indicated by the scale.

that can be placed in these relationships. However, some conclusions can be drawn. Although BLASTp searches and sequence alignments indicated that the group 1 possum clones (TvPol1–TvPol6) should be grouped together (see FIG. 3.5), the support for this group in the tree in FIG. 3.9 is very low (bootstrap value = 127). Clones TvPol3–TvPol6 appear to belong to a single group (bootstrap value = 931), whereas clones TvPol1 and TvPol2 are more distantly related, a situation which is also apparent by visual inspection of their sequences (see FIG. 3.5). The genomic clones of group 2 (TvPol7–TvPol11) form an exclusive group, with high bootstrap support (bootstrap value = 994), that also includes the previously reported sequence from the common brush-tailed possum, RV common possum (Herniou *et al.*, 1998). The support for the placement of the group 2 sequences with the HERV-L and MuERV-L elements is moderate (bootstrap value = 721). The final genomic clone, TvPol12, was clearly (bootstrap value = 867) more closely related to the RV stripe-faced dunnartII sequence (Herniou *et al.*, 1998) than it was to any other retrovirus. None of the possum genomic clones were closely related to the RV rock wallaby sequence (Martin *et al.*, 1997), the RV stripe-faced dunnartI sequence (Herniou *et al.*, 1998), or the RV koala or RV opossum sequences (Martin *et al.*, 1999) (FIG. 3.9).

Although the clones amplified from possum blood plasma RNA (TvPol13–TvPol17) are grouped with the type D retroviruses (MPMV, SMRV-H, JSRV) in the tree in FIG. 3.9, the support for this group is low (bootstrap value = 186), and the precise placement of clones TvPol13–TvPol17 remains uncertain. Nevertheless, it is clear that the clones amplified from plasma RNA are distinct from those amplified from genomic DNA (FIG. 3.9).

3.3 Discussion

It is clear that the genomes of possums, like those of all vertebrates, contain endogenous retroviruses. PCR, using primers (polF and polR) derived from conserved regions of the retroviral *pol* gene, was performed on possum genomic DNA. The PCR products were cloned, and twelve retrovirus-related clones were isolated. On the bases of sequence alignment, database

searches, and phylogenetic analysis, these clones could be separated into three groups.

The first group of possum clones (TvPol1–TvPol6) was most homologous to members of the HERV-K/type B/type D retrovirus group on the basis of protein database searches. They also grouped together in a phylogenetic analysis, although support for this group was weak. Rather, it appeared that four of the clones (TvPol3–TvPol6) were closely related to one another, whereas the other two clones (TvPol1, TvPol2) were more distantly related. These differences may reflect infection, over a period of time, of the possum germline (or the germline of its ancestors) by a HERV-K-/type B-/type D-related retrovirus that was evolving within the possum (or its ancestor). Alternatively, they could represent separate cross-species transmission events of related but distinct retroviruses from other animals. Interestingly, HERV-K belongs to a large family of endogenous retroviruses (the class II-like sequences) in humans, comprising at least ten groups (Andersson *et al.*, 1999). The groups are between 50% and 80% identical to one another on the basis of nucleotide sequence (Medstrand and Blomberg, 1993; Andersson *et al.*, 1999), and it has been suggested that they have been continuously integrating into the primate/human germline over the last 30 million years (Andersson *et al.*, 1999; see also Chapter 7). Perhaps a similar situation has occurred in possums. Herniou *et al.* (1998) reported a sequence from the stripe-faced dunnart (RV stripe-faced dunnartI) which belonged to a polytomy including the avian type C, HERV-K, type B, and type D retroviruses. However, both sequence alignments and phylogenetic analyses revealed only weak similarity between RV stripe-faced dunnartI and the possum genomic clones.

Based on alignments, database searches, and phylogenetic analysis, the second group of possum genomic clones (TvPol7–TvPol11) formed an exclusive group, which also included the only previously reported possum retroviral sequence, RV common possum (Herniou *et al.*, 1998). The phylogenetic analysis presented here, and that performed by Herniou *et al.* (1998) using a longer region of sequence, group the possum group 2 clones with the ERV-L endogenous retroviruses of humans (HERV-L) and mice

(MuERV-L). An extensive study of the distribution of ERV-L elements in vertebrates was conducted recently (Bénit *et al.*, 1999). ERV-L elements are ubiquitous and ancient within placental mammals, and have undergone recent retrotranspositional amplification within the genomes of some (mainly simian and murine) hosts (Bénit *et al.*, 1999). ERV-L-related elements were not detected in kangaroo DNA, or in the DNA of other non-mammalian vertebrates. However, Bénit *et al.* (1999) suggested that ERV-L elements might be present in the genomes of these animals, but that they were too distantly related to those in the placental mammals to be detected by Southern blotting or PCR screening. Indeed, in addition to the possum sequence already mentioned, Herniou *et al.* (1998) also isolated ERV-L-related fragments from a bird (elegant crested tinamou [*Eudromia elegans*]) and an amphibian (rhinematrid caecilian [*Epicrionops marmoratus*]). Thus, ERV-L elements appear to be widespread amongst vertebrates, not only the placental mammals. Further investigation would be required to determine the copy number of the ERV-L-related element in the possum genome, and its coding capacity.

The final group of possum genomic clones comprised only one member, TvPol12. Sequence alignments, database searches, and phylogenetic analysis all revealed considerable similarity of TvPol12 to the RV stripe-faced dunnartII sequence reported by Herniou *et al.* (1998). The dunnart sequence was previously shown to group with human endogenous retroviruses ERV9 and HERV-H (Herniou *et al.*, 1998). This relationship is supported by the phylogenetic analysis presented here. Further, the possum and dunnart sequences are more closely related to each other than they are to ERV-9 and HERV-H. This observation suggests either that the dunnart and the possum inherited, from a common ancestor, an ERV-9-/HERV-H-like retrovirus, or that such a retrovirus has been transmitted more recently between these animal species. HERV-H is an ancient integrant within the genomes of humans, Old World primates, and New World primates, and Anderssen *et al.* (1997) have suggested that its integration might have predated the simian/prosimian divergence. The results presented here, and by Herniou *et al.* (1998), suggest that it is possible that the HERV-H integration might even have predated the divergence between placental

mammals and marsupials. However, it should be emphasised that the phylogenetic analyses reported here are based on very short sequences, which limits their accuracy. Further investigation of the distribution of related retroviruses in marsupials and other vertebrates would be required to elucidate the true evolutionary history of these viruses.

A number of observations suggest that many more retrovirus-related elements remain to be discovered in the possum genome. First, only two of the sequences reported here (TvPol7 and TvPol8) were identical over the entire lengths of their nucleotide sequences (after omitting the primer sequences), and only two more (TvPol9 and TvPol10) were identical at the amino acid level. Second, Martin *et al.* (1997), using a HERV-I-related fragment isolated from the rock wallaby, detected HERV-I related sequences in the possum genome by Southern analysis. However, no HERV-I related sequence has been reported in possums. Third, MLV-related sequences appear to be ubiquitous in higher vertebrates (Herniou *et al.*, 1998), and have been amplified from genomic DNAs of both the American opossum and the koala (Martin *et al.*, 1999). It might therefore be expected that possums also possess MLV-related endogenous retroviruses.

It is clear that possums, and probably their evolutionary ancestors, have been subjected to the same kinds of retroviral infections that have affected all other vertebrates (Coffin, 1982; Wilkinson *et al.*, 1994; Boeke and Stoye, 1997). More extensive experimental analyses will be required to determine the biological activity of the possum elements. In addition, longer sequences of the possum endogenous viruses, and comparisons with related elements in the genomes of other marsupials will be required to determine the origins of these elements and their time of entry into the genome of the possum (or its evolutionary ancestors).

Reverse transcriptase activity was detected in the blood plasma of all possums tested. That the reverse transcriptase activity was filterable and pelletable, and therefore presumably particle-associated, suggested that it was of retrovirus origin. However, a positive PERT assay result is not definitive proof of the presence of an infectious retrovirus. Various DNA

polymerases in normal mammalian cells possess reverse transcriptase activity. DNA-dependent DNA polymerases, such as DNA polymerase α and DNA polymerase γ, are able to reverse transcribe synthetic (homopolymeric) RNA templates (Sarngadharan *et al.*, 1978). Although their activity on heteropolymeric templates (such as the MS2 RNA template in the PERT assay) is normally very low (or absent) (Sarngadharan *et al.*, 1978), it may be sufficient to produce a positive signal in the PERT assay (Lugert *et al.*, 1996). Lysates of mammalian cells often also possess genuine reverse transcriptases (Silver *et al.*, 1993; Pyra *et al.*, 1994; Heneine *et al.*, 1995; Böni *et al.*, 1996; Lugert *et al.*, 1996; Maudru and Peden, 1997;), which may be encoded by endogenous retroviruses, or by other retroelements (eg. LINEs) (see 1.3.6 Implications of presence of retroviruses in the genome/germline). More rigorous analysis, such as PERT assays on density gradient centrifugation fractions, would be required before the reverse transcriptase activity observed here could be categorically assigned a retrovirus-particle origin.

Although the reaction components of the PERT assay were tested for inherent reverse transcriptase activity and for amplifiable DNA (not shown), the possibility that artefacts arose from amplifiable DNA inadvertently contaminating the blood serum samples was never formally eliminated.

The *pol*-derived degenerate primers (polF and polR) were also used to RT-PCR amplify RNA isolated from filtered possum blood plasma. The RT-PCR products were cloned and sequenced, and five clones were isolated which were all identical (when primer sequences were omitted) in both nucleotide and amino acid sequences. Although the number of clones sequenced was small, the results suggest that they were generated from a homogeneous, or near-homogeneous, RNA population that is present in possum blood. The clones were very similar to type D retroviruses in primates (for example MPMV and SMRV-H) and sheep (JSRV), type B retroviruses in mice (MMTV), and related endogenous elements in humans (HERV-K) and rodents (IAPs). They were also related to, but distinct from, the group 1 clones generated from possum genomic DNA, and the RV stripe-faced dunnartI sequence reported by Herniou *et al.* (1998). The clones derived

from possum plasma RNA possessed uninterrupted reading frames over their entire lengths. Although the *pol*-related sequences of these clones are short, it is tempting to speculate that they were derived from a larger open reading frame encoding a functional reverse transcriptase protein, that which was detected using the PERT assay. Alternatively, the reverse transcriptase activity detected using the PERT assay might have been expressed from one of the endogenous elements in the possum genome, many of which also possessed uninterrupted reading frames (see FIG. 3.4).

Degenerate primers similar to polF and polR have been used to RT-PCR amplify RNA extracted from tissues of humans (Herrmann and Kalden, 1994; Lefebvre *et al.*, 1995) and other animals (DeHaven *et al.*, 1998). In all cases, more than one species of retrovirus-related sequence was detected. Two groups have used similar primers to amplify and characterise relatively pure retrovirus preparations isolated from cell culture supernatants (Donehower *et al.*, 1990; Perron *et al.*, 1997). However, to my knowledge, this is the first time such primers have been used to detect and characterise what may be a single retroviral RNA species extracted directly from biological samples, and illustrates the potential of such primers.

The conclusion that can be drawn from all of these observations is that a retrovirus-related sequence (or retroviral genome) is present in the blood plasma of all possums tested. It may be derived from a homogeneous population of retrovirus particles that are derived from an exogenous, infectious retrovirus that is endemic to possums. Alternatively, it may be expressed from an endogenous element present in the possum genome, which is related (but not identical) to endogenous elements already detected in the genome of the possum (TvPol1–TvPol6) and another marsupial (RV stripe-faced dunnartl), and which releases retrovirus particles. That the (albeit short) open reading frames of these clones are uninterrupted presents the possibility that the putative retrovirus also encodes functional reverse transcriptase, the activity of which was detected using the PERT assay.

Chapter 4

Amplification, Cloning, and Sequencing of a Full-Length Possum Endogenous Retrovirus, TvERV-K

4.1 Introduction

4.1.1 Initial 3'-RACE approach

Given that a retroviral particle-associated RNA appeared to be present in the blood of possums, an attempt was made to clone the full-length RNA. An RT-PCR-based approach, outlined in FIG. 4.1, was used initially, and was intended to exploit the internal sequence already obtained, as well as two specific structural features of retroviral RNA. First, retroviral RNAs possess poly(A) tails, and therefore an approach based on the rapid amplification of cDNA ends (3'-RACE; Frohman *et al.*, 1988) of mRNA can be used. An oligo(dT) primer, with an extension at its 5' end, is used to prime reverse transcription at the poly(A) tail of the RNA. The resulting cDNA is subsequently amplified using one primer corresponding to the 5' extension of the oligo(dT) primer, and another primer derived from the known sequence within the RNA (Frohman *et al.*, 1988). Second, retroviral RNAs possess direct repeats (R) at their 5' and 3' ends. Thus, theoretically, a primer derived from the R sequence at the 3' end of the genome can be used to amplify the unknown region between R at the 5' end of the genome and the

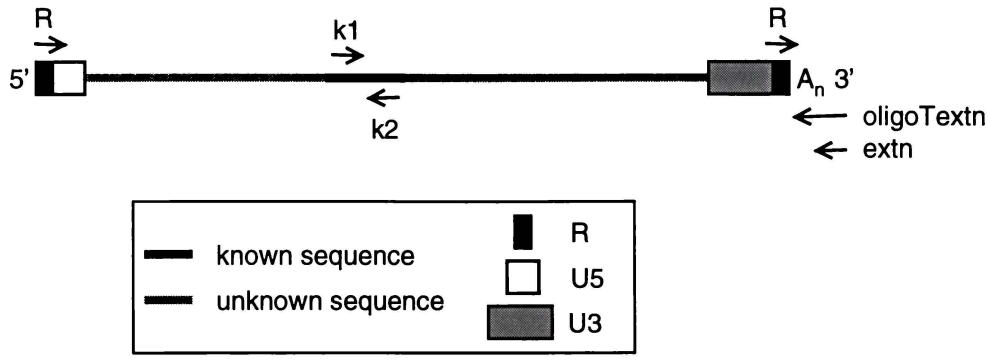


FIGURE 4.1 The strategy behind cloning full-length retroviral RNA. The RNA is reverse transcribed using oligoTextn (an oligo(dT) primer that is lock-docked at its 3' end and has an extension at its 5' end). The 3' end of the viral genome is amplified from the cDNA product using a primer from the known sequence (k1) and extn (identical to the 5' end of oligoTextn). The sequence of the PCR product will include the sequence of the R region at the 3' end of the genome. A primer derived from the R region at the 3' end of the genome can then be used, in conjunction with a primer from the known sequence (k2), to amplify the 5' end of the genome from the original cDNA copy, or from an internally-primed cDNA.

known sequence already obtained (FIG. 4.1). This latter step is, in essence, 5' RACE which circumvents many of the normal drawbacks of 5' RACE (see Frohman, 1995). Although this approach was, in theory, possible using current reverse transcription and PCR reagents, at least two factors were identified as potentially problematic. First, it was possible that the amounts of the retroviral RNA in preparations from blood plasma would be too low for successful reverse transcription and PCR-amplification. Second, the success of the approach might also depend on the sequence at the 3' end of the RNA genome (ie. the R region), and whether it was conducive to the design of a primer suitable for PCR.

4.1.2 Single primer PCR and direct sequencing of products

Traditionally, complete copies of endogenous proviruses have been isolated from the DNA of their hosts by screening genomic libraries by Southern hybridisation (Coffin, 1982; Wilkinson *et al.*, 1994). However, this approach is time-consuming and labour-intensive. Recently, a number of PCR-based techniques for obtaining unknown genomic sequence flanking known sequence have been developed. Examples are inverse PCR, novel *Alu*-PCR, vectorette-PCR, splinkerette-PCR, targeted gene-walking PCR, restriction-site PCR, capture PCR, panhandle PCR, and boomerang DNA amplification (BDA) (reviewed in Hui *et al.*, 1998). Some of these approaches are equally applicable to the cloning of unknown endogenous viral sequences flanking known viral sequences, and host genomic sequences flanking known viral sequences (see Hui *et al.*, 1998). However, although less labour-intensive than the generation and screening of a genomic library, most of these methods require multiple amplification, ligation, restriction enzyme digestion, and/or purification steps, some of which require optimisation (Hui *et al.*, 1998).

Direct sequencing of single primer PCR products was used by Screaton *et al.* (1993) to obtain unknown sequences adjacent to known sequences in YAC clones and genomic DNA. In contrast to the methods mentioned above, this approach involves only one PCR step (several PCR reactions performed simultaneously) and one purification step, followed by direct sequencing of

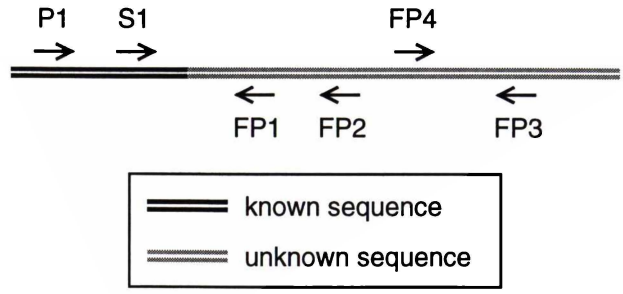


FIGURE 4.2 Direct sequencing of single primer PCR products. Single primer PCR reactions are performed using P1 as primer. FP1–FP4 represent potential mispriming sites for P1. S1 is the sequencing primer. Adapted from Screaton *et al.* (1993).

the products. The method, shown in FIG. 4.2, relies on the tendency of PCR primers to misprime (ie. prime polymerisation at imperfect complementary sequences), usually under non-stringent conditions. Parallel PCR reactions are performed at various annealing temperatures using a single primer from within the known sequence of genomic DNA (FIG. 4.2, P1). False priming in some of the reactions will result in amplification products of various lengths — some due to correct priming at the true complementary site (P1) and false priming at adjacent sites (FIG. 4.2, FP1–FP3), and some due to false priming at both ends (eg. FIG. 4.2, FP3 with FP4). Sixty cycles of amplification are performed, to ensure sufficient amplification of misprimed products. The reaction products are pooled, purified, and sequenced. Those fragments that contain the true complementary site at one end will also contain sequences complementary to the sequencing primer (FIG. 4.2, S1), and will serve as templates in the sequencing reaction. The original method described by Screaton *et al.* (1993) employed manual sequencing, whereas I have adapted the method for use with an ABI PRISM® 377 automated sequencer, and, in one case, using a long-template PCR system for the PCR.

4.2 Results

4.2.1 3'-RACE to amplify the 3' end of the retroviral genome

RNA from filtered blood plasma of two possums was reverse transcribed using a lock-docked (Borson *et al.*, 1992) oligo(dT) primer with a 5' extension

(oligoTextn). The resulting cDNA was PCR-amplified using a primer (TvP11F) derived from the sequence of clone TvPol13 and a primer (extn) corresponding to the 5' extension of the oligo(dT) primer. A smear of RT-PCR products was visible after agarose gel electrophoresis, with prominent ~2.8kb, ~3.3kb, and ~4.5kb fragments in the product from one of the possums (FIG. 4.3, lane 1).

The ~4.5kb RT-PCR product (FIG. 4.3, lane 1) was gel purified and reamplified using the TvP11F and extn primers (Cycling conditions: Initial denaturation: 94°C/30 sec; Cycles 1–20: 94°C/30 sec, 50°C/30 sec, 68°C/5 min; final extension 30min at 68°C [to maximise the addition of nontemplated A at 3' ends]). The reamplified ~4.5kb product was gel-purified again, cloned into T-tailed pBluescript® II KS(+), and transformed into chemically

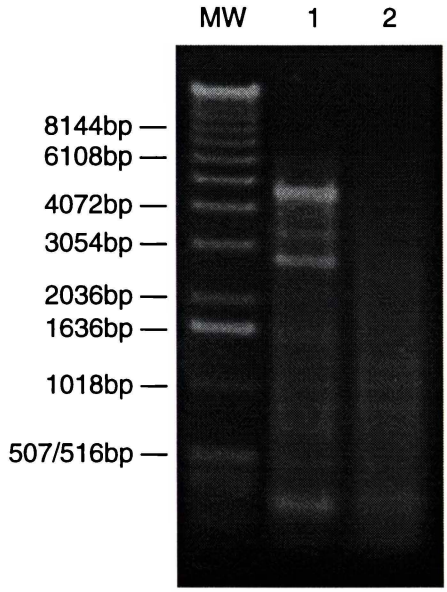


FIGURE 4.3 RT-PCR amplification of possum blood plasma RNA using TvP11F and oligoTextn/extn. RNA was isolated from 250 µl of the filtered blood plasma of two possums (lanes 1 and 2) using TRIzol® LS reagent (Life Technologies), resuspended in 10 µl DEPC-treated milliQ. A 2.5 µl aliquot of RNA was reverse transcribed at 50°C for 1 hr using 200 U Superscript™ II reverse transcriptase (Life Technologies) and 500 ng of oligoTextn as primer. The RNA strands of resulting RNA/DNA hybrids were digested with 1 U of RNase H (Boehringer Mannheim). 7.5 µl of the reverse transcription reactions were amplified in 50 µl reactions using 10 pmol of each of TvP11F and extn primers, 2 U eLONGase™ enzyme mix (Life Technologies), and 1.8 mM Mg^{2+}. Cycling conditions: Cycle 1: 94°C/1 min, 50°C/30 sec, 68°C/6 min; Cycles 2–35: 94°C/30 sec, 50°C/30 sec, 68°C/6 min. 15 µl of each reaction was electrophoresed in 0.6% agarose.

competent XL1-Blue MRF' *E. coli* cells. Seven clones containing inserts of the appropriate length were isolated. Of these, two were sequenced with M13 –21 and M13 Rev primers using an ABI PRISM® 377 automated sequencer. The clones were identical over the accurate range. Database searches using the nucleotide sequences revealed similarity of the TvP11F-proximal sequences to retroviral *pol* sequences.

One clone, pTvRV3', was sequenced in its entirety by primer walking (ie. by designing a primer from the end of the sequence obtained using the previous primer), using the primers TvP14F, TvP15F, TvP16F, and TvP17F successively in the 3' (forward) direction and TvP21R, TvP20R, TvP19R, and TvP18R successively in the 5' (reverse) direction (see Appendix 4). The sequence of pTvRV3' is shown as nucleotides 3882–8579 of the TvERV-K1 contig in Appendix 4. Clone pTvRV3' was 4762bp in length (including primers). BLASTn and BLASTx searches using the nucleotide sequence of clone TvRV3' revealed considerable homology to the *pol* gene of type D retroviruses at its 5' end. However, no homology to the *env* gene of any retrovirus was detected.

The 3' end of the pTvRV3' sequence lacked sequence motifs which are normally found at the 3' ends of retroviruses. In particular, no obvious polypurine tract (PPT; normally found immediately upstream of U3), TATA box (within U3), or polyadenylation signal (AATAAA; normally found within U3 or R) were present. Consistent with this observation, attempts at RT-PCR-amplifying the 5' end of the putative retrovirus from plasma RNA using a primer (TvP24F) designed from the 3' end of pTvRV3' in conjunction with a primer (TvP13R) designed from the 5' end (and in the opposite direction) were unsuccessful (data not shown). It appeared that the 3' end of pTvRV3' was not the true 3' end of the retrovirus, and that TvP24F was therefore not within the R region.

4.2.2 Clone pTvRV3' was derived from an endogenous possum retrovirus

The lack of a large ORF at the 3' end of pTvRV3' suggested that a functional *env* gene was absent from the putative retroviral RNA, which, in turn, suggested that it might have been generated from an endogenous, rather than exogenous, retrovirus. This possibility was investigated by PCR using primers derived from pTvRV3' to amplify possum genomic DNA. As shown in FIG. 4.4, PCR from mammary and liver genomic DNA of two possums

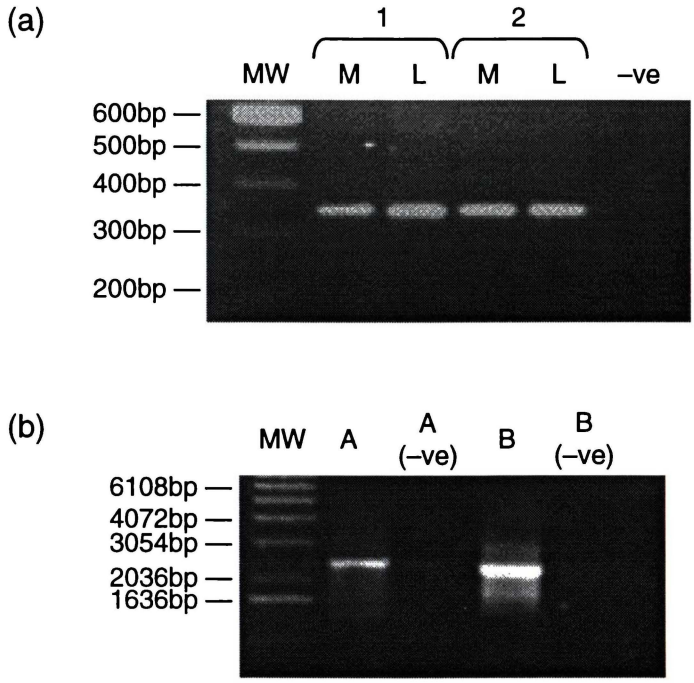


FIGURE 4.4 PCR amplification of possum genomic DNA using primers derived from pTvRV3'. (a) 500ng of genomic DNA isolated from mammary (M) and liver (L) tissues of two possums (1, 2) was PCR amplified using TvP17F and TvP18R in a standard 50 µl reaction (2 mM $MgCl_2$, 2 U Taq DNA polymerase [Boehringer Mannheim]; Cycling conditions: Initial denaturation: 94°C/30 sec; Cycles 1–40: 94°C/15 sec, 50°C/30 sec, 72°C/30 sec; final extension 72°C/10 min). 9 µl of each reaction was electrophoresed in 2.5% agarose. MW: 100bp ladder (sizes shown at left). –ve: negative (water) PCR control. (b) 275 ng of mammary genomic DNA from another possum was PCR amplified with primers TvP11F and TvP18R (A) and primers TvP17F and TvP21R (B) using eLONGase enzyme mix (Life Technologies). 50 µl reactions contained 1 U eLONGase enzyme mix, 1.8 mM Mg^{2+}, PTC-200. Cycling conditions: Cycle 1: 94°C/1 min, 60°C/30 sec, 68°C/2 min 30 sec; Cycles 2–40: 94°C/30 sec, 60°C/30 sec, 68°C/2 min 30 sec. 9 µl of each reaction was electrophoresed in 0.8% agarose. MW: 1kb ladder (sizes shown at left). A(–ve), B(–ve): negative (water) PCR controls using primer pairs in A and B, respectively.

using TvP17F and TvP18R generated fragments of the expected size (347bp based on pTvRV3') from every DNA sample. PCR from mammary genomic DNA of another possum using two overlapping primer pairs (TvP11F with TvP18R and TvP17F with TvP21R) also generated products of the expected sizes (2482bp and 2245bp, respectively, based on pTvRV3'). These results suggested that pTvRV3' was indeed derived from an endogenous virus.

Attempts at amplifying the 5' end of the putative endogenous retrovirus from genomic DNA using TvP24F and TvP13R were also unsuccessful, substantiating the possibility that TvP24F was not within the R region.

4.2.3 Direct sequencing of single primer PCR products to obtain sequences upstream and downstream of clone pTvRV3'

In an attempt to obtained sequence data upstream and downstream from the pTvRV3' sequence, a chromosomal walking technique involving the direct sequencing of single primer PCR (spPCR) products was used. The method of Screaton et al (1993), which employed manual sequencing, was modified for use with the ABI PRISM® 377 automated sequencer as described in Chapter 2 – Materials and Methods. Single primer PCR using TvP13R on genomic DNA and sequencing using TvP12R (spPCR1/TvP12R, FIG. 4.5) yielded sequence 5' to clone pTvRV3' (nt 3295–3881 of the TvERV-K1 contig, Appendix 4). Single primer PCR using TvP22F and sequencing using TvP23F (spPCR2/TvP23F, FIG. 4.5) yielded sequence 3' to clone pTvRV3' (nt 8580–8957 of the TvERV-K1 contig, Appendix 4).

The sequence of spPCR2/TvP23F was interesting in two respects. First, an A-rich region (5'-AAGAGAAGAAAAAAAAAA-3') occurred near the start of the sequence (nt 8582–8599 of TvERV-K1 contig, Appendix 4), and immediately downstream of the sequence at the 3' end of pTvRV3'. Second, when the spPCR2/TvP23F sequence was used to conduct a search of protein sequence databases (BLASTx search), a region downstream from the A-rich region was shown to encode a protein sequence with homology to the TM proteins of several retroviruses (see Chapter 5). Thus, it seemed likely that the oligo(dT)-based primer used to reverse transcribe the plasma

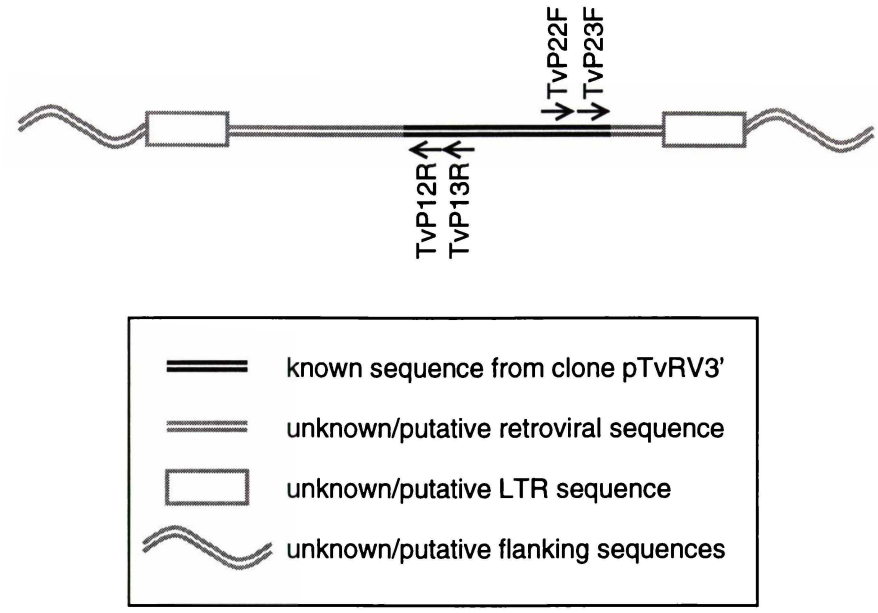


FIGURE 4.5 Single primer PCR and direct sequencing to obtain sequence upstream and downstream from clone pTvRV3' sequence. Single primer PCR was performed (as described in Chapter 2 – Materials and Methods) using primers TvP13R and TvP22F, and the PCR products sequenced with TvP12R (spPCR1/TvP12R) and TvP23F (spPCR2/TvP23F), respectively.

RNA (see 4.2.1 3' RACE to amplify the 3' end of retrovirus genome) bound to the A-rich region within the RNA, instead of (or as well as) the poly(A) tail of the RNA, and that preferential amplification of the internally primed cDNA had occurred (see 4.3 Discussion).

Amplification of oligoTextn-primed plasma RNA using TvP22F and extn generated only the 246bp product expected from oligoTextn priming at the internal A-rich region (data not shown).

4.2.4 3'-RACE to amplify the true 3' end of the retroviral genome

A primer (TvP26F) was designed from the sequence downstream from the A-rich region of spPCR2/TvP23F. RNA isolated from possum blood plasma was reverse transcribed using oligoTextn as primer. The product of reverse transcription was then amplified using TvP26F and extn. A ~800bp product was generated (FIG. 4.6).

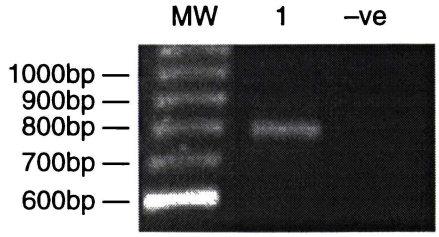


FIGURE 4.6 RT-PCR amplification of RNA from possum blood plasma using TvP26F and oligoTextn/extn. Total RNA was isolated from 250ul possum blood plasma using TRIzol reagent (Life Technologies) and resuspended in 10 µl of DEPC-treated water. 5 µl was reverse transcribed at 50°C for 1 hr using 200 U Superscript II reverse transcriptase (Life Technologies) and 500 ng of oligoTextn as primer. The RNA strands of RNA/DNA hybrids were digested with 1 U of RNase H (Boehringer Mannheim), 37°C, 30 min. 1 µl of the reverse transcription reaction was amplified in 50 µl reactions using 50 pmol of TvP26F and 50 pmol of extn, and 1.25 U Taq DNA polymerase (Boehringer Mannheim); 2 mM Mg^{2+}; PTC-200. Cycling conditions: Cycle 1: 94°C/1 min, 50°C/30 sec, 72°C/1 min; Cycles 2–40: 94°C/15 sec, 50°C/30 sec, 72°C/1 min; final extension 72°C/7 min). 9 µl of each reaction was electrophoresed in 2% agarose. MW: 100bp ladder (sizes shown at left). –ve: negative (milliQ) PCR control.

The ~800bp product (PCR800) was gel-purified and sequenced directly using TvP26F and an ABI PRISM® 377 sequencer. The resulting sequence corresponds to nucleotides 8958–9673 of the TvERV-K1 contig (Appendix 4). There were 9 nucleotide differences between PCR800 and spPCR2/TvP23F in an overlapping region of 81 nucleotides. The 3' end of PCR800 contained potential PPT and polyadenylation signal sequences (see Chapter 5), and thus it appeared that the sequence of PCR800 contained U3 and R regions, and represented the true 3' end of the putative retrovirus.

4.2.5 Single primer PCR and sequencing to determine sequence downstream of the R region

Two primers (TvP28F and TvP01F) derived from the sequence of PCR800 were used to perform single primer PCR on possum genomic DNA and sequencing of the single primer PCR product, respectively. Although this approach did not produce sequence using Taq DNA polymerase (Boehringer Mannheim), good sequence (spPCR3/TvP01; nt 9674–9818 of the TvERV-K1 contig, Appendix 4) was obtained when it was repeated using the

eLONGase™ long template PCR system (Life Technologies) and a longer extension time (5 min). The overlapping 49 nucleotides of spPCR3/TvP01 and PCR800 were identical, apart from the seven nucleotides immediately preceding the poly(A) tail of PCR800. It was assumed that the sequence of spPCR3/TvP01F beyond the R region included the U5 region of the 3' LTR of an endogenous retrovirus, and probably also genomic sequence adjacent (3') to the LTR. Comparison with LTR sequences obtained later allowed delineation of the 3' end of the LTR (see Chapter 5).

4.2.6 MuERV-L-related endogenous retroviral sequences are adjacent to the 3' LTR

Searches of nucleotide (BLASTn) and protein (BLASTx) sequence databases were conducted using the sequence downstream of the 3' LTR. A search using the deduced amino acid sequence of nucleotides 195–506 of spPCR3/TvP01F revealed considerable similarity to the product of the *pol* gene of MuERV-L, an endogenous retrovirus in mice (Bénit *et al.*, 1997) (FIG. 4.7).

4.2.7 PCR to generate the 5' end of the endogenous retrovirus

Primers TvP28F and TvP01F were both used in conjunction with TvP10R to try to amplify the 5' end of the endogenous retroviral genome from possum genomic DNA. A ~3kb product was generated using TvP01F and TvP10R, but no product was generated using TvP28F and TvP10R (FIG. 4.8). Based on analysis of the sequence of the TvERV-K1 contig (see Chapter 5), two nucleotides at the 5' end of TvP28F were predicted to lie within the putative PPT (nt 9425–9443 of the TvERV-K1 contig), two nucleotides were between the PPT and U3, and the rest of TvP28F (19 nucleotides) was expected to be within U3. Thus, amplification with TvP28 might have been unsuccessful because only part of it lay within U3 (see also 4.3 Discussion). TvP01F was another ~140bp downstream of TvP28F. Successful amplification with TvP01F and TvP10R, and sequencing of the product (see below) indicated that TvP01 almost certainly lay within U3.

```
spPCR3/TvP01F   CAGGGGCAAAGAATTATGGGAAGATATTGGTTACTAATTTGTCAGCTTTTCATGTAGAT
spPCR3/TvP01F      G  A  K  N  Y  G  K  I  L  V  T  N  L  S  A  F  H  V  D
MuERV-L Pol                                                    ...H  V  N
                                                                  *  *

spPCR3/TvP01F   GCTCACGTGGCCCTCACCATGCCAGAATGTGAGTACAATGTGCATGCTGATCAACTAGCA
spPCR3/TvP01F   A  H  V  A  L  T  M  P  E  C  E  Y  N  V  H  A  D  Q  L  A
MuERV-L Pol     A  H  Q  K  V  T  S  A  E  E  E  F  N  N  Q  V  D  K  M  T
                *  *           *        *     *     *              *

spPCR3/TvP01F   AAGACTGCTACTCAACATATTGTTCCTACTCTAAAACCAACTGATGATCTGATACTAGCA
spPCR3/TvP01F   K  T  A  T  Q  H  I  V  P  T  L  K  P  T  D  D  L  I  L  A
MuERV-L Pol     R  S  V  D  S  Q  T  L  S  P  A  I  P  -  -  -  -  V  I  A
                                                   *                    *

spPCR3/TvP01F   AAATGGGTCCATCAAACAGCTGGCCATTTAGGGGTCCAGGCCACACACCAGTGGGCATAA
spPCR3/TvP01F   K  W  V  H  Q  T  A  G  H  L  G  V  Q  A  T  H  Q  W  A  -
MuERV-L Pol     Q  W  A  H  E  Q  S  G  H  G  G  R  D  G  G  Y  P  W  A  Q
                   *     *           *  *     *                    *  *

spPCR3/TvP01F   GATCAAGGTATTAGCATCTCACACTCATTGTTAAGACAAGTAGCAGAGAAATGTCACATA
spPCR3/TvP01F   D  Q  G  I  S  I  S  H  S  L  L  R  Q  V  A  E  K  C  H  I
MuERV-L Pol     Q  H  G  L  P  L  T  K  A  D  L  A  T  A  A  A  D  C  Q  I
                      *                          *           *        *     *

spPCR3/TvP01F   TGTCAACTGGAAA
spPCR3/TvP01F   C  Q  L  E
MuERV-L Pol     C  Q...
                *  *
```

FIGURE 4.7 Alignment of the deduced amino acid sequence of nucleotides 195–506 of spPCR3/TvP01F with the *pol* gene product of MuERV-L. The MuERV-L sequence is from Bénit *et al.* (1997). * indicates identical amino acids in spPCR3/TvP01F and MuERV-L.

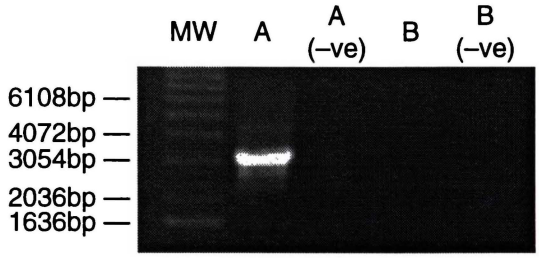


FIGURE 4.8 Amplification of the 5' end of the endogenous retrovirus from genomic DNA. 100 ng of possum mammary genomic DNA was PCR amplified with primers TvP01F and TvP10R (A) and primers TvP28F and TvP10R (B) using eLONGase enzyme mix (Life Technologies). 50 µl reactions contained 1 U eLONGase enzyme mix, 1.8 mM Mg^{2+}, GeneAmp 2400. Cycling conditions: Cycle 1: 94°C/1 min, 55°C/30 sec, 68°C/4 min; Cycles 2–35: 94°C/30 sec, 55°C/30 sec, 68°C/4 min. 8 µl of each reaction was electrophoresed in 0.8% agarose. MW: 1kb ladder (sizes shown at left). A(–ve), B(–ve): negative (water) PCR controls using primer pairs in A and B, respectively.

The ~3kb product (FIG. 4.8, lane A) was purified using the Promega Wizard® PCR purification system, diluted, and reamplified in a hemi-nested PCR using TvP01F and TvP09R (which is internal to TvP10R at the 3' end). The product of this amplification (PCR3000) was sequenced directly using TvP01F and TvP09R. Sequencing using TvP01F yielded only a short stretch of readable sequence, as did all subsequent sequencing using primers internal to TvP01F (TvP02F, TvP03F). Good sequence was obtained using TvP09R as the sequencing primer, however, and the sequence is shown as nucleotides 2687–3294 of the TvERV-K1 contig (Appendix 4). Another primer, TvP08R, was designed from the end of the sequence obtained using TvP09R.

Because sequencing from the 5' end of PCR3000 was proving unsuccessful, and in the interest of determining the remaining sequence as quickly as possible, I decided to clone the remaining 5' region of the retroviral genome. PCR3000 was reamplified using TvP03F and TvP08R (eLONGase™ system; 10 pmol each primer, 1.5 mM Mg^{2+}, 1 U eLONGase enzyme mix; Cycling conditions: Cycle 1: 94°C/1 min, 55°C/30 sec, 68°C/2 min 45 sec; Cycles 2–25: 94°C/30 sec, 55°C/30 sec, 68°C/2 min 45 sec; final extension: 68°C/6 min). The product was Wizard purified, A-tailed, ligated into pGEM®-T vector (Promega), and transformed by electroporation into electrocompetent XL1-Blue MRF' *E. coli* cells. One of the resulting clones (pTvRV5') was sequenced in its entirety with M13 –21, M13 Rev, TvP07R, and TvP06R (nt 264–2686 of the TvERV-K1 contig).

All of the relevant sequences were used to construct a contiguous sequence from the 5' LTR to the 3' LTR. The sequence of this contig, named TvERV-K1 contig (for *Trichosurus vulpecula* endogenous retrovirus, with lysine [K] tRNA primer binding site, contiguous sequence), is shown in Appendix 4.

4.2.8 Amplification and cloning of a near-full-length endogenous retrovirus

The sequence of the TvERV-K1 contig was assembled from the sequences of a variety of RT-PCR, PCR, and single primer PCR products. Because it

was possible that different parts of the sequence were generated from related yet different endogenous elements in the possum genome (see 4.3 Discussion), I thought it was necessary to clone and sequence a single element.

Primers TvP03F and TvP27R were used to PCR amplify possum genomic DNA. As shown in FIG. 4.9, amplification from mammary genomic DNA of two possums generated a product of the expected size of ~9kb for both, as well as numerous smaller products which differed between the two samples. A prominent ~7.5kb product was present in both samples, and was the major product in sample 2. One of the samples also contained products of ~8kb, ~7.2kb, ~6.7kb, ~6kb, ~5kb, and ~3.2kb (FIG. 4.9, lane 1). The second sample contained products of ~6.7kb, ~5.3kb, ~4.1kb, and ~3.2kb, in addition to the ~9kb and ~7.5kb products, and it is possible that the prominent ~7.5kb obscured other products of a similar size (FIG. 4.9, lane 2). Attempts to achieve exclusive amplification of the ~9kb band, by making changes in annealing temperature, template concentration, and buffer conditions, were unsuccessful.

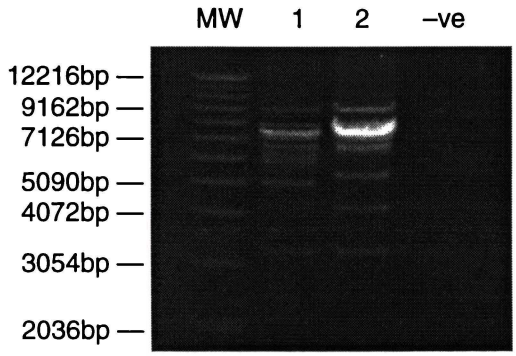


FIGURE 4.9 PCR amplification of full-length endogenous retrovirus from possum genomic DNA. 100 ng of mammary genomic DNA from two possums (1, 2) was PCR amplified with primers TvP03F and TvP27R using Expand Long Template PCR system with buffer system 3 (Boehringer Mannheim). 50 µl reactions contained 2.5 U Expand Long Template Enzyme, 2.25 mM Mg^{2+}, PTC-100. Cycling conditions: Cycle 1: 94°C/2 min 10 sec, 58°C/30 sec, 68°C/8 min; Cycles 2–10: 94°C/10 sec, 58°C/30 sec, 68°C/8 min; Cycles 11–30: 94°C/10 sec, 58°C/30 sec, 68°C/8 min 20 sec + 20 sec/cycle; final extension 68°C/8 min. 8 µl of each reaction was electrophoresed in 0.7% agarose. MW: 1kb ladder (sizes shown at left). –ve: negative (water) PCR control.

The ~9kb band of one of the PCR products (FIG. 4.9, lane 2) was gel-purified, reamplified (2.5 U Expand™ Long Template Enzyme, 2.25 mM Mg^{2+}; Cycling conditions: Cycle 1: 95°C/2 min 30 sec, 58°C/30 sec, 68°C/8 min; Cycles 2–10: 95°C/30 sec, 58°C/30 sec, 68°C/8 min; Cycles 11–30: 95°C/30 sec, 58°C/30 sec, 68°C/8 min 20 sec + 20 sec/cycle; final extension 68°C/8 min), and gel-purified again. The final product was ligated into pGEM®-T vector (Promega). Attempts at transforming the ligation reaction product into electrocompetent XL1-Blue MRF' *E. coli* cells (Stratagene) were unsuccessful (none of 30 colonies screened contained an insert of the appropriate length). Following transformation into electrocompetent STBL2™ *E. coli* cells (Life Technologies), however, one clone (of 140 colonies screened) contained an insert of the appropriate size. This clone, named pTvERV-K2 (for plasmid containing *Trichosurus vulpecula* endogenous retrovirus with lysine [K] tRNA PBS), was sequenced in its entirety using an ABI PRISM® 377 automated DNA sequencer, BigDye™ terminator chemistry, and primers M13 –21, TvP05F, TvP06R, TvP07R, TvP08R, TvP09R, TvP11F, TvP12R, TvP13R, TvP14F, TvP15F, TvP16F, TvP17F, TvP19R, TvP20R, TvP21R, TvP22F, TvP26F, and M13 Rev. Sequence with BigDye™ terminator chemistry and TvP25R terminated within the GC-rich region at nucleotides 8194–8226 of pTvERV-K2 (Appendix 5). However, this region was successfully sequenced using TvP25R and dGTP BigDye™ terminator chemistry. The sequence of pTvERV-K2 is shown in Appendix 5 and is analysed in Chapter 5.

4.3 Discussion

Retroviral RNAs possess structural features that are theoretically conducive to their characterisation using standard molecular biology procedures. I had intended to use a 3'-RACE approach to amplify the 3' end of the putative retroviral RNA in possum blood. A primer derived from the sequence at the 3' end could then have been used to amplify the 5' end of the virus, because retroviral RNAs possess identical sequences (the R region) at their 5' (immediately after the 5' CAP) and 3' (immediately before the poly[A] tail) ends. However, this approach was only partially successful. Reverse transcription was performed using an anchored oligo(dT)-based primer

(oligoTextn), which was supposed to bind (exclusively) to the poly(A) tail of the retroviral RNA. It appears that the primer also bound at an A-rich region within the RNA (~1 kb upstream from the poly[A] tail) and this internally-primed cDNA appeared to be preferentially amplified over that primed at the true poly(A) tail. Preferential amplification of the internally primed product can be explained by two (not necessarily mutually exclusive) phenomena. First, native M-MuLV reverse transcriptase and Superscript™ II reverse transcriptase (M-MuLV RT with point mutations that inactivate the RNase activity) can perform RNA and DNA strand displacement synthesis (ie. displacement of a non-template RNA or DNA strand hybridised to a DNA template). However, displacement synthesis is much slower than non-displacement synthesis (Whiting and Champoux, 1994; Kelleher and Champoux, 1998; Whiting and Champoux, 1998). According to the current model, displacement of a DNA strand from an RNA template is not expected to be required during reverse transcription of the retroviral genome. Unpublished work suggests that Superscript™ II can only displace 50–100 base pairs of DNA on an RNA template (Gary Gerard, pers.comm.; Chris Gruber, pers. comm.). Thus, it is conceivable that reverse transcription was primed (by oligoTextn) at both the internal A-rich region and the poly(A) tail of the retroviral RNA, but that those products primed at the poly(A) tail were not able to proceed efficiently past the internal A-rich region because of DNA products primed at this region. Second, even if full-length cDNAs were generated from primers bound at the poly(A) tail, the internally-primed cDNAs may have been preferentially amplified. Preferential amplification of short products over long products, where both the short and long products have the same primer-binding sequences, can occur with standard PCR reagents, although long-distance PCR reagents are reported to reduce or eliminate this effect (Wilton and Lim, 1996; Sanchez *et al.*, 1998).

No RT– or –ve PCR controls were performed in the RT-PCR experiment shown in FIG. 4.3. The fact that pTvRV3' contained the sequence of oligoTextn at its 3' end would suggest that it was amplified from the product of reverse transcription. However, the large amount of oligoTextn in the reverse transcription reaction (37 pmol) and the large amount of the reverse transcription reaction that was added to the PCR reaction (7.5 µl in a total

reaction volume of 50 µl), meant that oligoTextn was present at a higher concentration in the PCR reaction (0.28 µM) than were the PCR primers (0.2 µM). It is therefore possible that oligoTextn served as a PCR primer in the PCR reactions in FIG. 4.3, and that the PCR products resulted from amplification from contaminating genomic DNA. However, the failure to detect contaminating genomic DNA in the same RNA preparations using RT-PCR with polF and polR primers (data not shown) counts against this possibility.

An RT– control was also lacking from the RT-PCR shown in FIG. 4.6. However, that the sequence of oligoTextn is present at the end of PCR800 suggests this product was almost certainly generated from an RNA species, given that no poly(A) region is present in the provirus at this position.

Analysis of the sequence of pTvRV3' suggested that it did not represent the true 3' end of a retrovirus — it lacked recognisable PPT, TATA, and polyadenylation signal sequences.

Inspection of the open reading frames of pTvRV3' (see Chapter 5) suggested that it was not derived from a functional, infectious retrovirus. Although a large open reading frame, corresponding to the *pol* gene, was observed at the 5' end of pTvRV3', the 3' end possessed no long open reading frames. Nor were any similarities to the *env* genes (or *env* gene products) of any known retroviruses detected in the 3' sequence of pTvRV3'. Thus, it seemed likely that pTvRV3' was derived from the message of an endogenous retrovirus. That elements corresponding to pTvRV3' were present in the genome of possums was confirmed by PCR on possum genomic DNA.

The presence of pTvRV3' in the possum genome allowed the characterisation of the rest of the proviral genome by a combination of single primer PCR and sequencing, RT-PCR, and PCR approaches. A contiguous sequence, the TvERV-K1 contig, which was assembled from the partial sequences, is shown in Appendix 4, and is analysed in more detail in Chapter 5.

Single primer PCR followed by direct sequencing of the PCR products was used on several occasions. Initially, the method described by Screaton et al (1993) was followed for the PCR step, with the sequencing step modified for use with an automated DNA sequencer. On one occasion (spPCR3), I used a long template PCR system (eLONGase) and longer extension times for the PCR step. This modification resulted in good sequence using primers that had previously given poor sequence under the standard conditions. The rationale behind this approach was that, given that mispriming is expected to be a rare event (Screaton *et al.*, 1993), the use of a long template PCR system should increase the amplifiable range within which mispriming could occur and increase the yield of misprimed products.

Another interesting observation from the sequence of spPCR3 was that the sequence adjacent to the LTR of TvERV-K1 contig possessed considerable similarity to the *pol* gene of MuERV-L. Thus, it appears that the TvERV-K1 element integrated into the possum genome within another retroelement, a MuERV-L-related element (TvERV-L?). A MuERV-L-related sequence from possums was reported earlier by Herniou *et al.* (1998), and similar sequences were observed in the products of PCR from possum genomic DNA (see Chapter 3). However, the sequence in spPCR3 corresponds to a region of the *pol* gene which is 3' to that reported by Herniou *et al.* (1998) and those amplified using the polF and polR primers.

Clustering of retroelements is often observed (for reviews see Kuff and Lueders, 1988; Wilkinson *et al.*, 1994). The integration of retroelements within, or in close proximity to, other retroelements may represent similar integration site specificities of the retroelements (Wilkinson *et al.*, 1994). Although retroviruses do not appear to display sequence specificity at their sites of integration, the distribution of proviruses is often not random, and may reflect interaction with DNA-bound proteins or a preference for certain DNA topologies (Craigie, 1992). Alternatively, observations of clustering of retroelements may simply be the result of selection against retroelements that are integrated into more important sites in the genome. It is likely that integration of one retroelement into another one (which, presumably, is

already tolerated) is more likely to be tolerated by the host (Wilkinson *et al.*, 1994).

In other vertebrate genomes which have been studied, the copy numbers of different endogenous retroviruses can range from a single copy to over a thousand copies (see 1.3 Biology of endogenous retroviruses). Of those that are present in multiple copies, individual copies usually differ in sequence. Thus, given that the TvERV-K1 contig was assembled from the sequences of a number of different RT-PCR, PCR, and spPCR products, it was possible that it contained sequences from several related but different copies of TvERV-K related elements in the possum genome. This possibility was supported by the observations that, where sequences of PCR products overlapped (for example, spPCR2 and PCR800, PCR800 and spPCR3, spPCR3 and pTvRV5' [see Appendix 4]), they often differed due to nucleotide substitutions and insertions/deletions. Thus, the sequence of a single provirus was sought. PCR-amplification of possum genomic DNA using TvP03F and TvP27R generated an expected ~9.2kb product (as well as smaller products [see below]), which cloning and sequencing confirmed was closely related to the TvERV-K1 contig. The sequence of this clone (pTvERV-K2) is shown in Appendix 4, and is also analysed in more detail in Chapter 5.

All attempts to achieve exclusive amplification of the ~9.2kb product were unsuccessful. Although the smaller fragments generated by PCR using TvP03F and TvP27R may represent non-specific, non-retrovirus-related amplification products, it is also possible that they are derived from endogenous retroviruses that are related to, but that have deletions or rearrangements relative to, the TvERV-K1 contig and pTvERV-K2 sequences. In this respect, amplification from plasma RNA generated products of ~3.3kb and ~2.8kb, in addition to the ~4.5kb fragment that was cloned (pTvRV3') (see FIG. 4.3, lane 1). Although these products may also have resulted from non-specific priming, they might have been amplified from shorter TvERV-K1 contig-/pTvERV-K2-related RNA species that are expressed and packaged along with that of the RNA from which pTvRV3' was derived.

Chapter 5

Analysis of the Sequences of the TvERV-K1 Contig and pTvERV-K2

5.1 Introduction

Retrovirus genomes comprise regions that encode proteins (coding regions) and regions that are required for replication of the retrovirus but which do not necessarily encode proteins (regulatory motifs). Although these sequences may diverge markedly between retroviruses, certain motifs are absolutely required for replication and are consequently highly conserved, even between distantly related retroviruses. Such conserved regions can be used to predict the likely extent of biological activity (or inactivity) of a retrovirus in the absence of direct experimental assays. Nucleotide and amino acid sequences can also be used to delineate evolutionary relationships between retroviruses.

5.1.1 Regulatory regions

Retroviruses possess several regulatory regions that are used at various stages of replication, including the LTRs, splice donors and acceptors, frameshift or termination suppression signals, a packaging signal, and a primer binding site (PBS) and polypurine tract (PPT).

Proviral genomes possess two identical long terminal repeats (LTRs) at the time of integration into the host genome. Each LTR can be divided into three regions: U3; R; and U5 (Vogt, 1997b). The 5' end of U3 of the 5' LTR defines the 5' end of the provirus. U3 contains promoter (TATA box and CAAT motif) and enhancer sequences which direct transcription by host cell machinery (Rabson and Graves, 1997). The transcription start site defines the U3/R boundary. The R region is repeated at both ends of the retroviral RNA genome, and is required for the first strand transfer during reverse transcription (Telesnitsky and Goff, 1997). The 3' end of the 5' LTR is defined by the 5' end of the PBS. Similarly the 3' end of the PPT defines the 5' end of the 3' LTR. The site of polyadenylation within the 3' LTR defines the R/U5 boundary, and the polyadenylation signal (usually TATAAA) is located within U3 or R (Rabson and Graves, 1997). Finally, the 3' end of U5 is the 3' end of the provirus.

Pre-integration retroviral DNA is cleaved at the 3' end of each strand, and two nucleotides are removed, leaving a terminal CA dinucleotide at each end (Kulkosky and Skalka, 1994; Brown, 1997). The two nucleotides which are removed from the 5' and 3' ends of the retroviral genome correspond to the two nucleotides between the PPT and U3, and those between U5 and the PBS, respectively. Inside the removed nucleotides, retroviral LTRs normally possess inverted repeats (often imperfect) at their 5' and 3' ends (Kulkosky and Skalka, 1994; Brown, 1997).

All replication-competent retroviruses possess splice donors near their 5' ends, normally between the PBS and the start of the *gag* gene, although in some cases within the *gag* gene (Rabson and Graves, 1997). Splicing between this splice donor and a splice acceptor upstream from the *env* gene produces the Env-coding mRNA. Some retroviruses, especially the complex retroviruses, exploit additional splice donors and acceptors and alternative splicing to generate a variety of sub-genomic mRNAs (Rabson and Graves, 1997).

The expression of Gag-Pro and/or Gag-Pro-Pol proteins of retroviruses requires circumvention of translation termination codons at the ends of the *gag* and/or *pro* ORFs. Most retroviruses rely on frameshifting between the *gag*, *pro*, and/or *pol* genes during translation. Retroviral frameshifts always occur in the –1 direction, and must occur in the region where the two reading frames overlap. A seven nucleotide slippery sequence, which represents the last two codons in the original (0) reading frame as well as the first nucleotide in the new (–1) reading frame, acts as the frameshift site. Slippery sequences have the general feature 5'-X XXA AA<u>C</u>-3' or X XXU UU<u>U/A</u>, where X is a specific base, the underlined base is read in both the original and –1 reading frames, and the spacing designates codons in the original reading frame (Levin *et al.*, 1993; Swanstrom and Wills, 1997). In the mammalian type C retroviruses, the *pro* and *pol* ORFs lie in the same reading frame as the *gag* ORF, and synthesis of Gag-Pro-Pol proteins requires readthrough (termination) suppression of the termination codon at the end of the *gag* gene. This amber (UAG) termination codon is occasionally decoded by glutamine tRNA (Levin *et al.*, 1993; Swanstrom and

Wills, 1997). Secondary RNA structures, such as hairpins and pseudoknots, appear to increase the efficiency of both frameshifting and readthrough (termination) suppression (Levin *et al.*, 1993; Swanstrom and Wills, 1997).

The packaging signal (Ψ) of the retrovirus genome directs packaging of the full-length retroviral RNA genome into retrovirus particles, to the exclusion of sub-genomic retroviral mRNAs and cellular RNAs. The packaging signal is normally located near the 5' end of the genome, just downstream from the splice donor site and sometimes overlapping the 5' end of the *gag* gene (Berkowitz, 1996; Swanstrom and Wills, 1997). However, packaging signals may also be discontinuous. The secondary structure formed by the packaging signal appears to be more important than its sequence *per se* (Berkowitz, 1996; Swanstrom and Wills, 1997).

The primer binding site (PBS) is located immediately downstream of the 5' LTR, and defines the 3' end of the LTR. The PBS of all retroviruses is complementary to the 3' end of a tRNA, although the species of tRNA varies between retroviruses. Following entry into a host cell, reverse transcription of the retroviral RNA genome is primed by the 3' end of the specific tRNA which is complementary to the PBS (Telesnitsky and Goff, 1997). One system for naming endogenous retroviruses includes the species of the tRNA used as primer for reverse transcription in the name of the retrovirus (for example, the human endogenous retrovirus HERV-K has a PBS which is complementary to the 3' end of $tRNA^{Lys}_{1,2}$; the PBS of HERV-H corresponds to $tRNA^{His}$) (Wilkinson *et al.*, 1994).

The polypurine tract (PPT) is located immediately upstream of the 3' LTR, and acts as primer for plus strand DNA synthesis during reverse transcription. The actual PPT primer is generated by cleavage of the plus strand RNA by the RNase H activity of reverse transcriptase. Although the actual sequence is not conserved, the PPT, as its name suggests, is rich in purine (A and G) residues (Telesnitsky and Goff, 1997).

5.1.2 Coding regions

The sequences of retroviral open reading frames (ORFs), or, more importantly, their protein products, can shed light on the likely biological activity of a retrovirus. To be competent for replication, a retrovirus must possess complete ORFs for all of the major retroviral proteins, namely Gag, Pro, Pol, and Env. In addition, motifs that are absolutely essential for the correct functioning of those proteins must be retained for full biological activity of the virus. The Gag, Pro, Pol, and Env proteins all possess characteristic conserved motifs that are required for correct functioning, and that also aid their identification.

The Gag polyproteins of retroviruses are cleaved, during virus particle maturation, into the matrix (MA), capsid (CA), and nucleocapsid (NC) proteins (Vogt, 1996; Swanstrom and Wills, 1997). All of these proteins possess characteristic motifs. The Gag proteins of many retroviruses are myristylated at their amino-termini (ie. the N-termini of the MA proteins), and possess recognition signals for the myristylation machinery. Myristate is a 14-carbon fatty acid that is added cotranslationally to the Gag protein. The Gag proteins of some retroviruses are not myristylated, but in those that are, myristylation is believed to be an important determinant of the membrane-associated nature of these proteins (Vogt, 1997b). The CA protein contains a region, known as the major homology region (MHR), which is the most conserved region amongst distantly related retroviruses (except for the spumaviruses, which lack this region) (Vogt, 1997b). The MHR appears to be important in the very late stages of virion maturation, and after infection of a new cell (Craven *et al.*, 1995; Craven and Parent, 1996). The NC proteins of most retroviruses, excluding the spumaviruses, contain one or two CCHC motifs with the structure $CX_2CX_4HX_4C$ (where X represents any amino acid). Each CCHC motif co-ordinates a Zn^{2+} ion, and forms a "zinc-finger" (Berkowitz *et al.*, 1996; Vogt, 1997b; Rein *et al.*, 1998). The CCHC motifs are critical for retroviral RNA packaging, and may function at other stages of retrovirus replication (Berkowitz *et al.*, 1996; Vogt, 1997b; Rein *et al.*, 1998). NC proteins also commonly have many basic residues proximal to and within the CCHC motifs that are crucial for other NC functions, including RNA

maturation within the retrovirus particle, and enhancement of reverse transcription and strand-transfer during replication of the RNA genome (Berkowitz *et al.*, 1996; Rein *et al.*, 1998).

Retroviral proteases (PRs) are homodimeric aspartic proteases. They are incorporated into retrovirus particles as part of Gag-Pro and/or Gag-Pro-Pol polyproteins. They release themselves from this polyprotein context, and cleave the Gag, Gag-Pro, and/or Gag-Pro-Pol polyproteins into their separate proteins, during maturation of the virus particle (Vogt, 1996; Swanstrom and Wills, 1997). Several conserved regions, corresponding to active site residues, have been identified in retroviral PRs, the most significant of which is the Asp-Thr/Ser-Gly active site triplet (Rao *et al.*, 1991; Vogt, 1996; Swanstrom and Wills, 1997).

The Pol protein comprises reverse transcriptase (RT), which includes polymerase and RNase H domains, and integrase (IN) proteins. The Pol protein is also incorporated into retrovirus particles as part of Gag-Pro-Pol polyproteins, from which it is released during particle maturation (Vogt, 1996; Swanstrom and Wills, 1997). The polymerase domain is the most conserved region of all the retroviral proteins (McClure *et al.*, 1988), and possesses several motifs that are conserved not only in retroviral reverse transcriptases (Jacobo-Molina and Arnold, 1991; Katz and Skalka, 1994; Telesnitsky and Goff, 1997), but also in reverse transcriptases from all retroelements (Doolittle *et al.*, 1989; Xiong and Eickbush, 1990; McClure, 1993), in all RNA-dependent polymerases (Poch *et al.*, 1989), and in all nucleic acid polymerases (Delarue *et al.*, 1990). The RNase H domain of RT also possesses several conserved motifs (Davies *et al.*, 1991; Katz and Skalka, 1994; Telesnitsky and Goff, 1997). Finally, IN can be divided into three domains: an N-terminal domain of unknown function which contains a His-His-Cys-Cys zinc-finger-like motif (Khan *et al.*, 1991); a central catalytic core containing the D,D(35)E motif which includes determinants of DNA substrate specificity and residues critical for catalysis (Katzman *et al.*, 1991; Khan *et al.*, 1991; Kulkosky *et al.*, 1992); and a C-terminal domain that is poorly conserved (for reviews see Katz and Skalka, 1994; Polard and Chandler, 1995; Brown, 1997).

The Env polyprotein comprises two proteins, the surface (SU) and transmembrane (TM) proteins. Although their amino acid sequences are generally not highly conserved, these proteins possess functionally conserved regions. The Env polyprotein initially possesses an N-terminal signal sequence, which is recognised by the signal recognition particle (SRP). The SRP transfers the nascent Env polyprotein to the endoplasmic reticulum (ER), where subsequent protein synthesis is accompanied by translocation across the ER membrane. The SU proteins of most retroviruses are highly modified by N-linked glycosylation at the Asn residue in Asn-X-Ser/Thr motifs (where X is any amino acid). SU is the primary determinant of receptor recognition. The TM protein contains two characteristic hydrophobic regions. The fusion peptide is activated upon receptor binding and mediates fusion of the viral and host cell membranes. The membrane-spanning domain serves to anchor TM (and therefore the whole Env protein) in the viral membrane (for reviews see Einfeld, 1996; Hunter, 1997; and Swanstrom and Wills, 1997).

5.2 Results

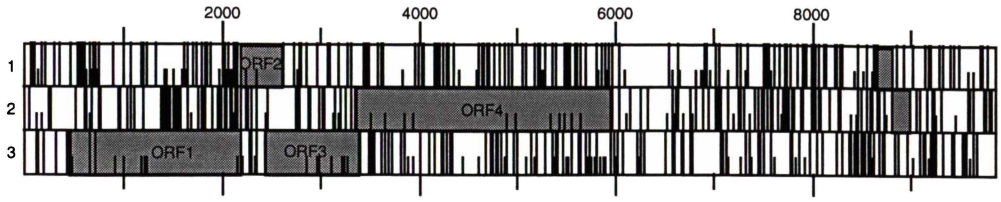
5.2.1 Alignment of the TvERV-K1 contig and pTvERV-K2

The sequences of the TvERV-K1 contig (see Appendix 4) and pTvERV-K2 (see Appendix 5) were aligned using the CLUSTAL X program as described in Chapter 2 – Materials and Methods (data not shown). The sequences contained numerous small (1–10 nt) insertions/deletions and nucleotide substitutions relative to each other. In addition, pTvERV-K2 contained a 115 nt insertion (nt 6267–6381) relative to the TvERV-K1 contig (between nt 6524 and 6525 in the TvERV-K1 contig). When this insertion was excluded, the sequences of the TvERV-K1 contig and pTvERV-K2 were 97% identical.

5.2.2 Open reading frames (ORFs)

The open reading frames (ORFs) of both the TvERV-K1 contig and pTvERV-K2 were deduced and are shown in FIG. 5.1. Both sequences possess several large ORFs within their 5' halves. ORF1 and ORF2 of the TvERV-K1

(a) TvERV-K1 contig



(b) pTvERV-K2

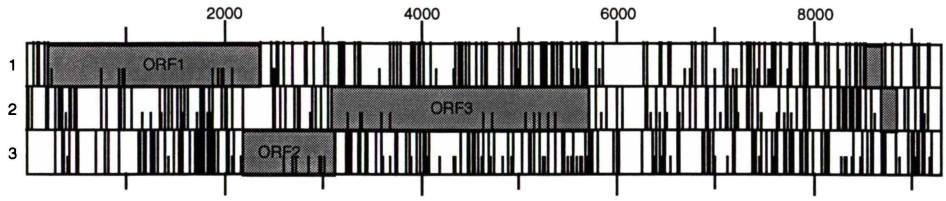


FIGURE 5.1 Open reading frame (ORF) maps of the TvERV-K1 contig (a) and pTvERV-K2 (b). The sequences of the TvERV-K1 contig and pTvERV-K2 were converted into ORF maps using DNA Strider. The shaded regions represent sequences with homology to retroviral proteins: ORFs 1 and 2 of the TvERV-K1 contig correspond the *gag* gene, ORF 3 corresponds to the *pro* gene, and ORF 4 corresponds to the *pol* gene; ORFs 1, 2, and 3 of pTvERV-K2 correspond to the *gag*, *pro*, and *pol* genes, respectively (see text for details). The shaded regions near the 3' end of each genome represent amino acid sequences with homology to retroviral transmembrane (TM) proteins. Numbers at the top of each ORF map represent the nucleotide positions as in Appendix 4 and Appendix 5. The numbers to the left represent the three reading frames. Short vertical lines within the ORF maps indicate start codons (AUG); long vertical lines indicate stop codons (TAA, TAG, TGA).

contig correspond to ORF1 of pTVERV-K2 — the frameshift from ORF1 to ORF2 of the TvERV-K1 contig is due to a single nucleotide (G) insertion at position 2210 of the TvERV-K1 contig. ORF3 and ORF4 of the TvERV-K1 contig correspond to ORF2 and ORF3 of pTvERV-K2. The ORF structures of the TvERV-K1 contig and pTvERV-K2 were compared with those of other retroviruses. The arrangement of pTvERV-K2, in particular, resembles those of the type B retrovirus MMTV, the type D retroviruses, as well as HTLV and BLV, all of which require frameshifts between both the *gag* and *pro* and the *pro* and *pol* reading frames for synthesis of the Gag-Pro-Pol polyprotein (Petropoulos, 1997; see also 1.2.3 The retrovirus genome). It will be shown below that the shaded regions comprising ORFs 1 and 2 of the TvERV-K1 contig and ORF1 of pTvERV-K2 represent the *gag* genes (see 5.2.4.1 Gag), ORF3 of the TvERV-K1 contig and ORF2 of pTvERV-K2 represent the *pro*

genes (see 5.2.4.2 Pro), and ORF4 of the TvERV-K1 contig and ORF3 of pTvERV-K2 represent the *pol* genes (see 5.2.4.3 Pol) of the TvERV-K elements. Shaded regions near the 3' ends of the genomes share amino acid sequence homology with the transmembrane (TM) proteins of some retroviruses (see 5.2.4.4 Env).

5.2.3 Regulatory regions

5.2.3.1 Long terminal repeats (LTRs)

The sequences of the TvERV-K1 contig 5' and 3' LTRs are shown in FIG. 5.2, and are compared with the U5 sequence from the 5' end of pTvERV-K2. The 5' ends of the TvERV-K1 contig 5' and 3' LTRs (to just beyond R) were both derived from the sequences of PCR800 and spPCR3 (see Appendix 4), and are consequently identical over this range. The 3' end of the TvERV-K1 contig 5' LTR was derived from the sequence of pTvRV5' (see Appendix 4) whereas the 3' end of the TvERV-K1 contig 3' LTR was derived from the sequence of spPCR3 (see Appendix 4), and these regions differ due to nucleotide substitutions and insertions/deletions. The U5 sequence from pTvERV-K2 differs from the same region of the TvERV-K1 contig 5' and 3' LTRs due to nucleotide substitutions and insertions/deletions. BLASTn searches using all of these sequences failed to detect any similarity to any retroviral LTRs.

The 5' end of the TvERV-K1 contig LTRs was assigned based on the observation that a putative PPT sequence is located adjacent to it at the 3' end of the TvERV-K1 contig (see 5.2.3.5 Polypurine tract [PPT]). The 3' end of the 5' LTR is bound by the PBS (see 5.2.3.4 Primer binding site [PBS]). The U3/R boundary (ie. the transcription start site) was assigned based on its similarity to the transcription start site in eukaryotic promoters (Breathnach and Chambon, 1981; Bucher, 1990). Amplification and sequencing of the viral RNA or primer extension analysis would be required to confirm this as the site of transcription initiation. The R/U5 border is based on the sequence immediately preceding the poly(A) tail at the 3' end of PCR800 (see 4.2.3 3'-RACE to amplify the true 3' end of the retroviral genome). The CA

```
                    IR U3
TvERV-K1 5' LTR   ---TGCCGGAGACTGGAAAATCGCCTTACCCCTCCCCTCTGCCTGACCCGCGCCCCCTCC
TvERV-K1 3' LTR   …CATGCCGGAGACTGGAAAATCGCCTTACCCCTCCCCTCTGCCTGACCCGCGCCCCCTCC
pTvERV-K2         ------------------------------------------------------------

TvERV-K1 5' LTR   CTTCGCTAGCTTTCATTTCTCGCTTAGCTCTCATTTCCTGGCTAACTCTCGTTTCTGGGA
TvERV-K1 3' LTR   CTTCGCTAGCTTTCATTTCTCGCTTAGCTCTCATTTCCTGGCTAACTCTCGTTTCTGGGA
pTvERV-K2         ------------------------------------------------------------

                              CCAAT box
TvERV-K1 5' LTR   TATGACGCAGGCCTGCGTCCAATGCTTAGCCACCCCTCTGGGGACGTGGCATTCTGATAA
TvERV-K1 3' LTR   TATGACGCAGGCCTGCGTCCAATGCTTAGCCACCCCTCTGGGGACGTGGCATTCTGATAA
pTvERV-K2         ------------------------------------------------------------

                        TATA box                poly(A) U3  ⟷      R      ⟷ U5
TvERV-K1 5' LTR   CATGCTCATAAAAACCCTTACCAGGAACAATAAAGAGAGACATTTTTTACACCAGCAGCG
TvERV-K1 3' LTR   CATGCTCATAAAAACCCTTACCAGGAACAATAAAGAGAGACATTTTTTACACCAGCAGCG
pTvERV-K2         ------------------------------------------------------------

TvERV-K1 5' LTR   CGTCTAGTCTCCTTCCTCTCCGTGATCCCGTTCCCGAGCCTCTTGGCCCTTTAGAGCCAT
TvERV-K1 3' LTR   CGTCTAGTCTCCTTCCTCTCCGTGATCCCGTTCCCGAGCCTCTTGGCCCTT-AGAGCCAT
pTvERV-K2 U5      ------GTCTCCTTCCTCTCCGTGATCTCGTTCC-GAGCCTCTTGGCCCTT-AGAGCCAT

TvERV-K1 5' LTR   CAGGTCCTCCTCGAGCCTAGCTCCCCCGGCCCTCTCCAGCTTCGCTGACCCACGATTGTG
TvERV-K1 3' LTR   TAGGTTCCCCTCGAGCCTAGCTCCCCCGGACCTCTCCAGCTTCGCTAACCCACGATTGTG
pTvERV-K2 U5      TAGGTTCCCCTCGAGCCTAGCTCCCC-GGACCTCTCCAGCTTCGCTAACCCACGATTGTG

                              U5 IR
TvERV-K1 5' LTR   TGACCTGCGGGCCAGGTCACT…
TvERV-K1 3' LTR   TGACCTGCGGGCCAGGTCA---
pTvERV-K2 U5      TGACCTGCGGGCCAGGTCAAT…
```

FIGURE 5.2 LTR sequences of the TvERV-K1 contig and pTvERV-K2. The TvERV-K1 5' and 3' LTRs correspond to nt 1–378 and nt 9440–9818 of the TvERV-K1 contig (Appendix 4). The pTvERV-K2 U5 sequence corresponds to nt 1–132 of pTvERV-K2 (Appendix 5). The CCAAT box, TATA box, and poly(A) signal are shaded. IR indicates the inverted repeats (underlined) at the ends of the LTRs. ⟷ indicates U3/R and R/U5 boundaries. See text for further details.

dinucleotide at this position is moderately conserved at the site of polyadenylation in other retroviruses (Rabson and Graves, 1997), and is an appropriate distance (20 nt) from the polyadenylation signal (Guntaka, 1993, Rabson and Graves, 1997; and see below).

The TG and CA dinucleotides at the 5' and 3' ends of the LTRs, respectively, represent the internal repeats (IR) which serve as the recognition sequences for IN during integration.

The sequence CATAAAAA, the T of which is 33 bp upstream of the putative transcription start site (U3/R boundary), probably represents the TATA box sequence (Breathnach and Chambon, 1981; Bucher, 1990). This is supported by the presence of a potential CCAAT box (Breathnach and Chambon, 1981; Bucher, 1990) 49 bp upstream from the T of the TATA box.

A potential polyadenylation signal (AATAAA), which is identical to the consensus polyadenylation signal (Guntaka, 1993; Rabson and Graves), is also evident (FIG. 5.2). In FIG. 5.2, the putative CCAAT and TATA boxes, although present in both the 5' and 3' LTRs, are shaded only in the 5' LTR where they control transcription initiation. Likewise, the polyadenylation signal (poly(A)), shaded only in the 3' LTR, is expected to function only at the 3' end of the genome.

5.2.3.2 Splice sites

The sequence AG**GT**AAGT, located 19bp upstream from the start of the *gag* gene in both the TvERV-K1 contig (nt 459–466) and pTvERV-K2 (nt 211–218) conforms with the consensus splice donor sequence AG**GU**AAGU, where the **GU** dinucleotide is at the exact 5' end of the intron (Myers, 1997). The lack of an *env* ORF (see 5.2.4.4 Env protein) precludes the identification of an appropriate splice acceptor.

5.2.3.3 Frameshifts

Analysis of the ORFs of both the TvERV-K1 contig and pTvERV-K2 suggested that they contained ORFs for viral proteins (see 5.2.2 Open reading frames [ORFs]). Indeed, closer inspection revealed that the ORFs in the TvERV-K1 contig and pTvERV-K2 coded for the retroviral proteins Gag, Pro, and Pol (see 5.2.4 Gene products). The ORFs were arranged in a similar way to those of the type B retrovirus MMTV, the type D retroviruses, as well as the HTLV and BLV viruses (Petropoulos, 1997; see also FIG. 1.3). All of these retroviruses rely on ribosomal frameshifting within the overlapping regions of the *gag* and *pro* ORFs and the *pro* and *pol* ORFs for synthesis of the Gag-Pro and Gag-Pro-Pol polyproteins. The overlapping regions of the *gag* and *pro* ORFs and the *pro* and *pol* ORFs of the TvERV-K1 contig and pTvERV-K2 are shown in FIG. 5.3. A candidate slippery sequence for frameshifting (GGGAAAC) is apparent in the *pro-pol* overlap of both the TvERV-K1 contig (nt 2481–2487) and pTvERV-K2 (nt 2223–2229) (FIG. 5.3a). No slippery sequence was apparent in the short overlap between the *pro* and *pol* genes (FIG. 5.3b).

(a) *gag-pro* overlap

```
TvERV-K1     ...TATAGAGGGCAAACCCCTTCCCCCGTTGTCGGGAAACGCCTGGCGGGGTCCGCCCCGGA
pTvERV-K2    ...TATAGACGGCAAACCCCTCCCCCCGTTGTCGGGAAACGGCTGGCGGGGTCGGCCCCGGA

TvERV-K1     CCCCTCAAGGACTGGCCCAGCCAGCCACGGGGACCCACCCTGCCCACAGCTCTCCAGAGC
pTvERV-K2    CCCCTCAAGGACTGGCCCAGCCAGCCACGGGGACCCACCCTGCCCACAGCTCTCCAGAGC

TvERV-K1     AACGGCAGGCAGTGCAGGAGTGGACTTGTGTGCCACCGCCGCCGCAATAT...
pTvERV-K2    AACGGCAGGCAGTGCAGGAGTGGACTTGTGTGCCACCGCCGCCGCAATAT...
```

(b) *pro-pol* overlap

```
TvERV-K1     ...CCGCTCCGGCCTTGGTTTCCAGAAACATTTTTCA...
pTvERV-K2    ...CCGCTCCGGCCTTGGTTTCCAGACACATTTTTCA...
```

FIGURE 5.3 Nucleotide sequences at the overlaps between the *gag* and *pro* (a) and *pro* and *pol* (b) genes of the TvERV-K1 contig and pTvERV-K2. The *gag-pro* overlaps are nt 2451–2619 of the TvERV-K1 contig (Appendix 4) and nt 2193–2361 of pTvERV-K2 (Appendix 5). The *pro-pol* overlaps are nt 3356–3389 of the TvERV-K1 contig (Appendix 4) and nt 3098–3131 of pTvERV-K2 (Appendix 5). The putative frameshift slippery sequences in the *gag-pro* overlaps are underlined.

Visual comparison of the overlapping regions of the *gag* and *pro* ORFs of the TvERV-K1 contig and pTvERV-K2 with the *gag-pro* overlaps in the type D retroviruses revealed some similarities (data not shown). Superficially, it appears that the TvERV-K *gag-pro* overlap regions are capable of forming pseudoknots similar to those shown to influence frameshifting in the type D retroviruses (ten Dam *et al.*, 1994) and other retroviruses (see Swanstrom and Wills, 1997) (FIG. 5.4). It remains to be determined whether this structure forms *in vivo*, or whether it exerts any effect on the efficiency of frameshifting. No such structure has been reported for the *pro-pol* frameshifts of the type B and type D retroviruses, and nor could one be identified in the overlapping regions of the TvERV-K *pro* and *pol* ORFs.

5.2.3.4 Primer binding site (PBS)

The PBS sequences from the TvERV-K1 contig and pTvERV-K2 are compared with the 18 nucleotides at the 3' end of $tRNA^{Lys}_{1,2}$ in FIG. 5.5. The PBS of the TvERV-K1 contig is perfectly complementary to the 3' end of mammalian $tRNA^{Lys}_{1,2}$. The PBS sequence of pTvERV-K2 contains a single G insert relative to the TvERV-K1 contig, but otherwise the sequences are

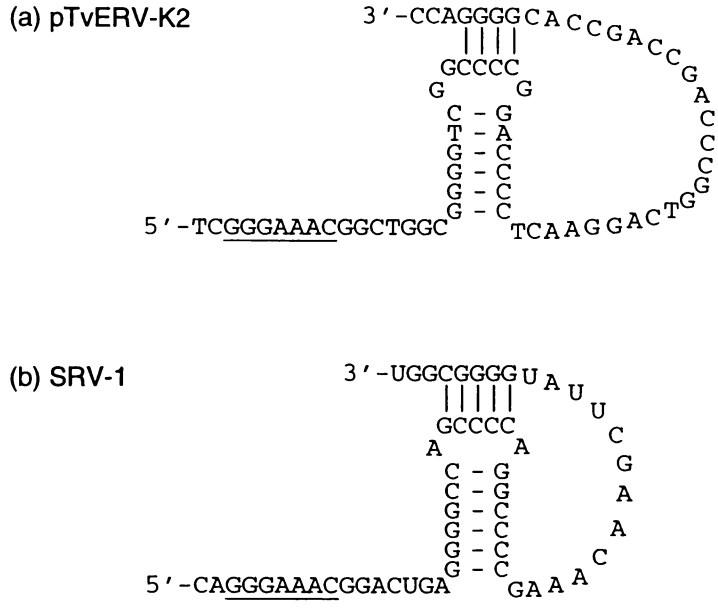


FIGURE 5.4 Pseudoknot structures at *gag-pro* overlaps in pTvERV-K2 (a) and SRV-1 (b). The SRV-1 structure is adapted from ten Dam *et al.* (1994). The putative slippery sequence for pTvERV-K2 and the actual slippery site sequence of SRV-1 are underlined.

identical. Thus it seems almost certain that this retrovirus uses (or used) $tRNA^{Lys}_{1,2}$ to prime minus strand DNA synthesis.

5.2.3.5 Polypurine tract (PPT)

The 5' end of the 3' LTR was not determined experimentally. However, a purine rich region near the 3' end of the TvERV-K1 contig (nt 9425–9441, Appendix 3) may represent the polypurine tract (PPT) of TvERV-K1, and would define nucleotide 9444 as the 5' end of the 3' LTR. The putative PPT of the TvERV-K1 contig is compared with the PPTs of three type D

TvERV-K1 contig	5'	...TGGCGCCCAA-CGTGGGGC... 3'
pTvERV-K2	5'	...TGGCGCCCAAGCGTGGGGC... 3'
$tRNA^{Lys}_{1,2}$	3'	ACCGCGGGUU-GCACCCCG... 5'

FIGURE 5.5 The primer binding sites (PBSs) of the TvERV-K1 contig and pTvERV-K2. The PBS sequences of the TvERV-K1 contig (nt 379–396) and pTvERV-K2 (nt 133–151) are compared with the 3' end of mammalian (rabbit) $tRNA^{Lys}_{1,2}$ (Raba *et al.*, 1979). Gaps introduced for optimal alignment are indicated by dashes (-).

retroviruses in FIG. 5.6. Again, that this sequence actually serves (or served) as the PPT during reverse transcription of the TvERV-K1 contig genome is purely speculative, but its similarity to the PPTs of the type D retroviruses suggests that this is possible.

5.2.4 Gene products

The TvERV-K1 contig and pTvERV-K2 sequences were conceptually translated in all three reading frames as described in Chapter 2 – Materials and Methods. The resulting amino acid sequences were used to perform searches of protein sequence databases and alignments with homologous proteins from related retroviruses. Together, these analyses confirmed that the large ORFs in the possum sequences encoded retroviral proteins.

5.2.4.1 Gag protein

A Gag protein is encoded by ORFs 1 and 2 (nt 486–2619) of the TvERV-K1 contig and by ORF1 (nt 238–2361) of pTvERV-K2 (see FIG. 5.1, Appendix 4, and Appendix 5). The change in reading frame from ORF1 to ORF2 in the TvERV-K1 contig is due to a single nucleotide (G) insertion at position 2210. The *gag* ORF of the TvERV-K1 contig contains two in-frame stop codons (TAG at nt 657–659, TAA at nt 708–710); the corresponding codons in pTvERV-K2 code for amino acids. The TvERV-K1 contig also contains a 9 nucleotide (ie. 3 codon) insert relative to pTvERV-K2 (nt 904–912 in the TvERV-K1 contig). Comparison of the Gag amino acid sequences of the TvERV-K1 contig and pTvERV-K2 revealed that they were 93% identical.

TvERV-K1 contig	...AAAATAAAAATGGGGGACATG...
MPMV	...AAAAAGGGTGACATG...
SMRV-H	...AAAGAAAAAGGGAGGAGATG...
JSRV	...AAAGAATAAAGAGAGGGGAGCTG...

FIGURE 5.6 The polypurine tract (PPT) of the TvERV-K1 contig and the type D retroviruses. The putative PPT sequence of the TvERV-K1 contig (nt 9425-9441) is compared with the PPTs of three type D retroviruses. The TG in each sequence corresponds to the 5' end of the 3' LTR.

Searches of nucleotide sequence databases using the nucleotide sequences of the TvERV-K1 contig and pTvERV-K2 *gag* genes failed to detect any retrovirus-related sequences. However, searches of protein sequence databases using the deduced amino acid sequences of the TvERV-K1 contig Gag and pTvERV-K2 Gag detected significant homology to the Gag proteins of several genera of retroviruses. The possum sequences were most similar to Gag proteins of the type D retroviruses (MPMV, SRV-1, SRV-2, SMRV-H, JSRV), were less similar to Gag proteins of mouse mammary tumour virus (MMTV), HERV-K elements of humans, and rodent intracisternal A-type particles (IAPs), and were distantly related to the Gag proteins of other retroviruses (eg. HIV-1).

The predicted Gag proteins of the TvERV-K1 contig and pTvERV-K2 are shown in FIG. 5.7, and are compared with those of three type D retroviruses. Overall, there are few positions at which amino acid residues are conserved between all of the sequences shown, although some regions of higher homology are apparent (see below).

Several factors suggest that the start codon (ATG) at nucleotides 486–488 of the TvERV-K1 contig and nucleotides 238–240 of pTvERV-K2 is almost certainly the start codon which is (or was) used for synthesis of the Gag protein. First, it is the first ATG codon in the transcripts of these sequences. Second, this codon conforms to the consensus for initiating codons — purine residues are present three bases upstream (A) three bases downstream (G) from the A of the start codon (Kozak, 1984). Third, the Gag protein initiated at this ATG codon bears significant similarity at its amino terminus to those of the type D retroviruses (FIG. 5.7).

It is likely that the Gag proteins of the TvERV-K1 contig and pTvERV-K2 are (or were) myristylated at their amino termini. Myristylation, attachment of a myristyl group to the N terminus of a protein via an amide link, occurs at the N-terminal glycine following removal of the initiating Met residue. The myristylation machinery has an absolute requirement for a Gly residue at the N-terminal position (position 1), and displays a preference for Asn, Gln, Ser,

```
TvERV-K1 Gag     MGSQLSK--EQVFVKDLKQSFRERGIRVKKKDLLKFFIFIDKVCPWFIVNGPDIHPGKWK
pTvERV-K2 Gag    MGSPLSK--EQVFLKDLKQSFKERGIRIKKKDLLKFFIFIDKVCPWFIVNGPDIHPGKWQ
MPMV Gag         MGQELSQ--HERYVEQLKQALKTRGVKVKYADLLKFFDFVKDTCPWFPQEG-TIDIKRWR
SMRV-H Gag       MGQASSHSENDLFISHLKESLKVRRIRVRKKDLVSFFSFIFKTCPWFPQEG-SIDSRVWG
JSRV Gag         MGHTHSR---QLFVHMLSVMLKHRGITVSKTKLINFLSFIEEVCPWFPREG-TVNLETWK
                 **   *          *      *        * * *    ****   *       *

TvERV-K1 Gag     VG#RDLNEKLKQEGPE#VPINAFSLWTLINDIIEGSTGDKGARQLLVQAESCLAPLSRAA
pTvERV-K2 Gag    KVGRDLNERLKQEGPESVPINAFSLWTLINDIIEGSTGDKGARQLLVQAESCLAPLSRAA
MPMV Gag         RVGDCFQDYYNTFGPEKVPVTAFSYWNLIKELIDKKEVNPQVMAAVAQTEEILKSNSQTD
SMRV-H Gag       RVGDCLNDYYRVFGPETIPITTFNYYNLIRDVLTNQSDSPDIQRLCKEGHKILISHSRPP
JSRV Gag         KVGEQIRTHYTLHGPEKVPVETLSFWTLIRDCLDFD--NDE----LKRLGNLLKQ-E---
                           *** *        **                    *

TvERV-K1 Gag     SASSLHAENSQDKEEVKSPPPI-----------------------IMPPTVKSIYPPPP
pTvERV-K2 Gag    SASSLHAENSQDKEEAKSPP-------------------------PTPTAKSIYPPLP
MPMV Gag         LTKTSQNPDLDLISLDSDDE-------------------------GAKSSSLQDKGL
SMRV-H Gag       SRQAPVTITTSEKASSRPPSRAPSTCPSVAIDIGSHDTGQSSLYPNLATLTDPPIQSPHS
JSRV Gag         -EDPLHTPDSVPSYDPPPPP---------------------------------------

TvERV-K1 Gag     NPVEETLTFPVFPTERKVGMNTAPENES--GPLPDKDAANLDAEAAAYNPQE-VFVSHNF
pTvERV-K2 Gag    NPVEETLTFPVFPTEGKVGMNTAPENES--GPLPDENAANLDVEAAAYNPQE-VFLSHAF
MPMV Gag         SSTKKPKRFPVLLTAQTSKDPEDPNPSE--VDWDGLEDEAAKYHNPDWPP-F-LTRPPPY
SMRV-H Gag       RAHTPPQHLPLLANSKTLHNSGSQDDQLNPADQADLEEAAAQYNNPDWPQLTNTPALPPF
JSRV Gag         -----PPSLKMHPSDNDDSLSSTDEA--------ELDEEAAKYHQEDWGF---------L

TvERV-K1 Gag     SPPHYLLPPLVDLSQIQRDLTQRMAELRKTVNMQNQYHQMVKEFSALQASLQKALLPPSA
pTvERV-K2 Gag    SPHPYLLPPLVDLSQIQRDLTQRMAELRKTVNMQNQYHQMVKEFSALQASLQKALLPPSA
MPMV Gag         NKATPSAPTVMAVVNPKEELKEKIAQLEEQIKLEELHQALISKLQKLKTGNETVTHP--D
SMRV-H Gag       RPPSYVSTAVPPVAVAAPVLHAPTSGVPGSPTAPNLPGVALAKPSGPIDETVSLLDGVKT
JSRV Gag         AQEKGALTSKDELVECFKNLTIALQNAGIQLPSNNNTFPSAPPFPPAYTPTVMAGLD---
                                *

TvERV-K1 Gag     PKGTLIAFPQEKIPQPKQKQKSLLFPILRSNTQNRPKDAEDSNTDENPPEISAEKEESGE
pTvERV-K2 Gag    PKSTPIALPQDKIPRPKQKQKSLLFPILRSNTQNRPKDAEDTNSDKHPPENSAGLEESEE
MPMV Gag         TAGG----------------------LSR-TPHWP---G----QHIPKGKCCASREKEE
SMRV-H Gag       LVTKLSDLALLPPAG------VMAFPVTRSQGQVSSNTTG--RASPHPDTHTIPEEEEAD
JSRV Gag         PPPG----------------------FPPPSKHMS-----------PLQKALRQAQRLG
                                                             *

TvERV-K1 Gag     ESQEEEVENDDQESDSEEDSKSCSSKYKAPKFKNIKDLHAAVKKYGPNAPFTLSALEAIG
pTvERV-K2 Gag    ESQEEEVEDDNQESDSEEDNKSCSSKYKAPKFKNIKDLHAAVKKYGPNAPFTLSALEAIG
MPMV Gag         QIPKDIFPVTETVDGQGQAWR----HHNGFDFAVIKELKTAASQYGATAPYTLAIVESVA
SMRV-H Gag       SGESDSEDDEEESSEPTEPTYT--HSYKRLNLKTIEKIKTAVANYGPTAPFTVALVESLS
JSRV Gag         EVVSD-FSLAFPVFENNNQRY-----YESLPFKQLKELKIACSQYGPTAPFTIAMIESLG
                                                      *   **  ** *    *

TvERV-K1 Gag     QGGYLLPGEWVRVAQATLSRGQFLTWKAEFDYHCQTIKKRNLKSKANRDWTFEKLAGRGE
pTvERV-K2 Gag    QGGYLLPGEWVRVARAALSRGQFLTWKAEFEYHCQTIEKRNLKSKANRDWTFEKLAGRGE
MPMV Gag         DN-WLTPTDWNTLVRAVLSGGDHLLWKSEFFENCRDTAKR--NQQAGNGWDFDMLTGSGN
SMRV-H Gag       ER-WLTPSDWFFLSRAALSGGDNILWKSEYEDISKQFAER-TRVRPPPKDGPLKIPGASP
JSRV Gag         TQ-ALPPNDWKQTARACLSGGDYLLWKSEFFEQCARIADV--NRQQGIQTSYEMLIGEGP
                     * *  *     * ** *    ** *                          *
```

FIGURE 5.7 Gag proteins of the TvERV-K1 contig, pTvERV-K2, and three type D retroviruses. The TvERV-K1 contig (Appendix 4) and pTvERV-K2 (Appendix 5) deduced Gag sequences were aligned with those of three type D retroviruses using the CLUSTAL X program as described in Chapter 2 – Materials and Methods. Invariant residues within the major homology region (MHR) of the capsid (CA) are shaded, and conserved hydrophobic residues are in bold text (Craven *et al.*, 1995; Bénit *et al.*, 1997). The Cys and His residues within the Cys-His motifs (CCHC 1, CCHC 2) of NC are shaded, other basic residues within NC are in bold. * indicates identical amino acids in all proteins; # indicates an in-frame stop codon; - indicates gaps introduced for optimal alignment. The type D retroviruses are: MPMV, Mason Pfizer monkey virus; SMRV-H, squirrel monkey retrovirus; JSRV, Jaagsiekte sheep retrovirus .

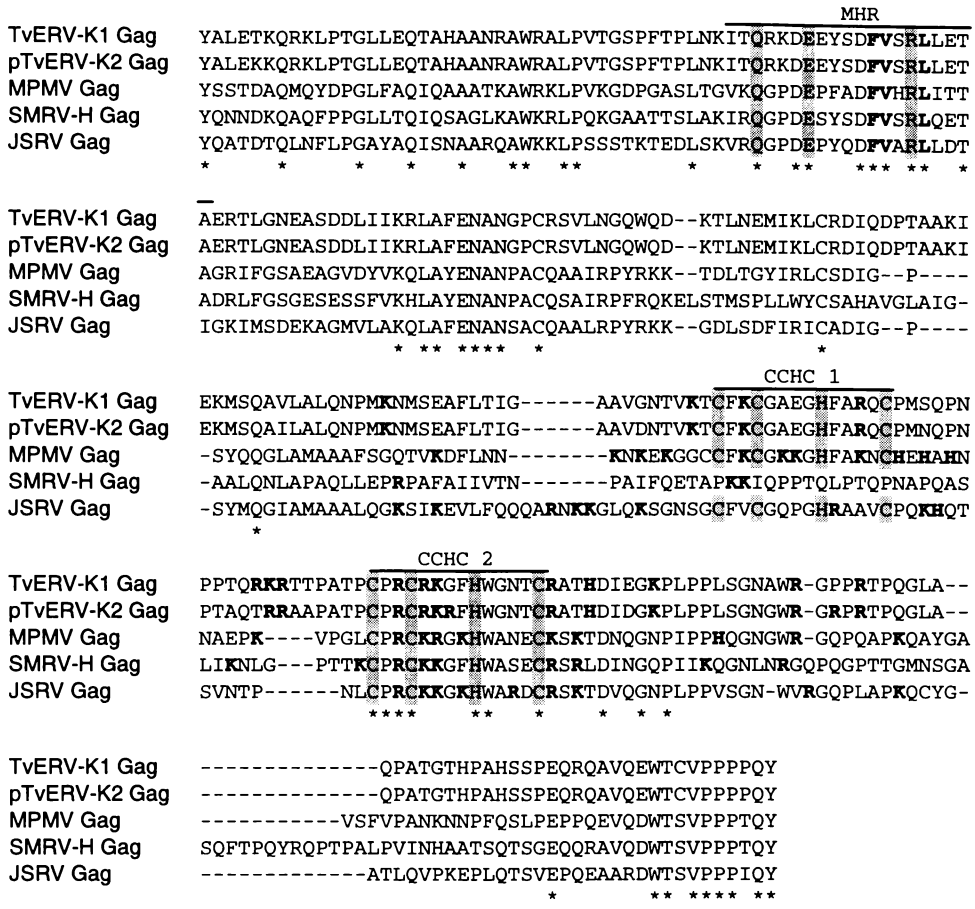


FIGURE 5.7 *Continued.*

Val, Ala, or Leu at position 2, and a strong preference for Ser at position 5 (Grand, 1989; Schmidt, 1989). The Gag protein of the TvERV-K1 contig, with the N-terminal sequence Met-Gly-Ser-Gln-Leu-Ser..., and that of pTvERV-K2, with the N-terminal sequence Met-Gly-Ser-Pro-Leu-Ser..., both conform to the consensus sequences (FIG. 5.7).

Both the TvERV-K1 contig and pTvERV-K2 Gag proteins possess sequence motifs that correspond to the major homology region (MHR), a highly conserved region within the capsid (CA) protein of all retroviruses. The residues conserved in the MHRs of all retroviruses (Craven *et al.*, 1995; Bénit *et al.*, 1997) are also conserved in the MHRs of the TvERV-K1 contig and pTvERV-K2 (FIG. 5.7).

The Gag proteins of the TvERV-K1 contig and pTvERV-K2 also possess two obvious CCHC motifs near their carboxyl termini, which presumably lie within the nucleocapsid (NC) proteins. The second of the CCHC motifs is shared

by all of the type D retrovirus NC proteins, and several other amino acids, in addition to the Cys and His residues, are present in all five proteins shown in FIG. 5.7. Although the first CCHC motif is absent in the SMRV-H NC protein, it is present in the other type D retroviruses, and again, several amino acids within this motif are conserved between the TvERV-K and type D retrovirus sequences (FIG. 5.7). The TvERV-K1 contig and pTvERV-K2 Gag proteins, as well as those of the type D retroviruses, contain numerous basic residues in close proximity to, and within, the Cys-His motifs (FIG. 5.7).

5.2.4.2 Pro protein

The TvERV-K1 contig and pTvERV-K2 both possess uninterrupted *pro* ORFs, corresponding to ORF3 (nt 2451–3389) in the TvERV-K1 contig and ORF2 (nt 2193–3131) in pTvERV-K2 (see FIG. 5.1). The deduced amino acid sequences are 97% identical. Both the TvERV-K1 contig and pTvERV-K2 *pro* ORFs are identical in length at 313 codons. Note, however, that the *pro* reading frame is probably not entered until the C of the slippery sequence GGGAAAC (nt 2481 of the TvERV-K1 contig, nt 2223 of pTvERV-K2) (see 5.2.3.3 Frameshifts), which would eliminate the first 12 amino acids of the Pro proteins.

Searches of protein sequence databases with the amino acid sequences of these ORFs detected highest homology to the Pro proteins of the type D retroviruses, less similarity to the Pro proteins of MMTV, IAPs, and HERV-K, and distant similarity to the Pro proteins of other retroviruses. The TvERV-K Pro sequences are aligned with those of three type D retroviruses in FIG. 5.8.

The 5' ends of the *pro* ORFs of type B and type D retroviruses encode an additional protein, deoxyuridine triphosphatase (DU) (McGeoch, 1990; Elder *et al.*, 1992), not present at this location in other retroviruses (V. M. Vogt, 1997). The five conserved DU motifs identified by McGeoch (1990) are also present in the N-terminal domain of the TvERV-K1 contig and pTvERV-K2 Pro proteins (FIG. 5.8). Numerous amino acids outside the DU motifs are also conserved in the TvERV-K and type D retrovirus DU domains (FIG. 5.8).

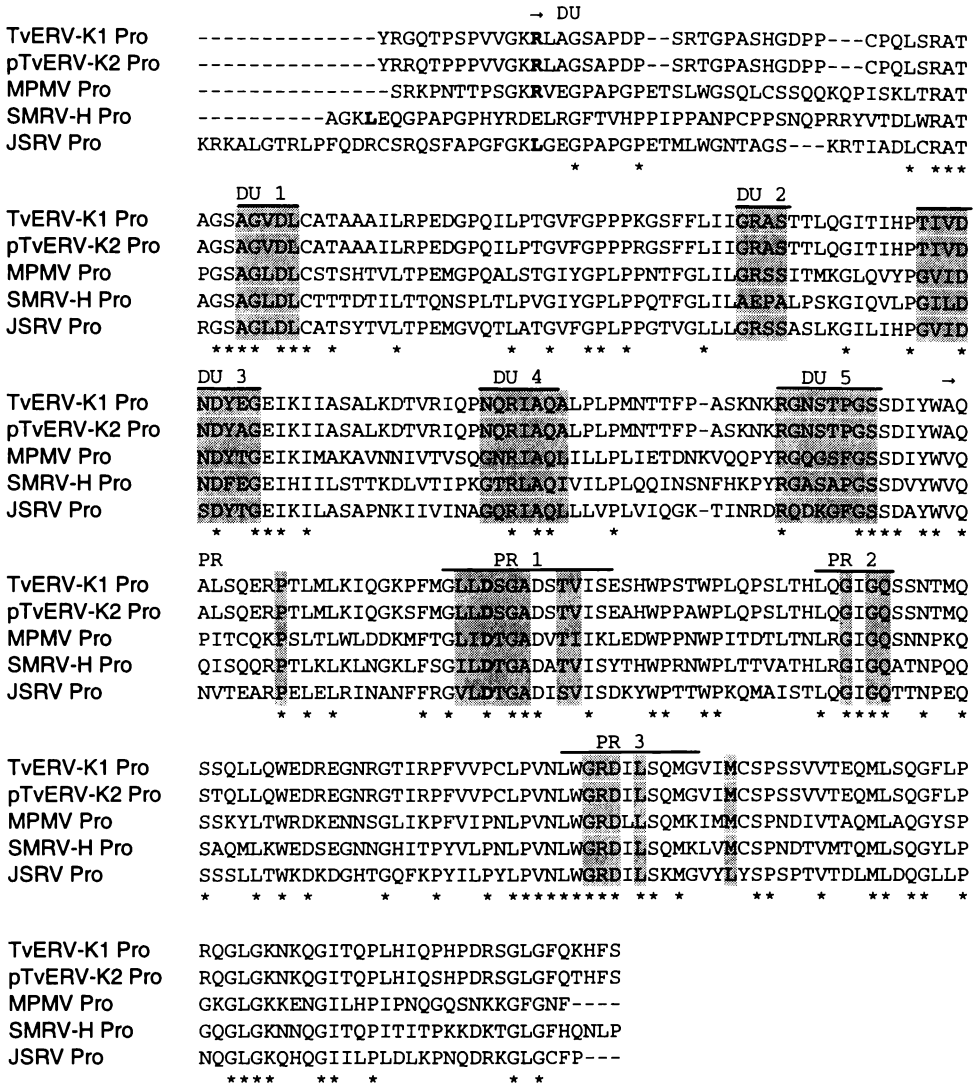


FIGURE 5.8 Pro proteins of the TvERV-K1 contig, pTvERV-K2, and three type D retroviruses. The TvERV-K1 contig (Appendix 4) and pTvERV-K2 (Appendix 5) deduced Pro sequences were aligned with those of three type D retroviruses using the CLUSTAL X program as described in Chapter 2 – Materials and Methods. Note that the sequences are those encoded by the entire ORFs, not the products of translation. The first coded amino acid (ie. the first codon after the *gag-pro* frameshift) in each protein is in bold. The N-terminus of PR (→ PR) is based on that reported by van der Kuyl *et al.* (1997). Conserved motifs within the DU domain (DU 1–DU 3) are as reported by McGeoch (1990). Conserved PR motifs (PR 1–PR 3) and particularly conserved residues (shaded) within PR are as reported by Rao *et al.* (1991) and Vogt (1996). The active site Asp (D) in PR 1 is in bold text. * indicates identical amino acids in all proteins; - indicates a gap introduced for optimal alignment. The type D retroviruses are as listed in FIG. 5.7.

The aspartyl protease (PR) domains of the TvERV-K1 contig and pTvERV-K2 Pro proteins share many conserved amino acids with the same domains from the type D retroviruses (FIG. 5.8). The three PR motifs (Rao *et al.*,

1991; Vogt, 1996) are highly conserved in the TvERV-K and type D retrovirus PR domains, and conserved residues are also apparent outside these motifs (FIG. 5.8). The active site residues in PR 1 (FIG. 5.8) match the consensus active site sequence (Asp-Thr/Ser-Gly) of all retroviral proteases and cellular aspartyl proteases (Toh *et al.*, 1985; Rao *et al.*, 1991; Vogt, 1996).

5.2.4.3 Pol protein

The *pol* ORFs (ORF 4 [nt 3356–5968] in the TvERV-K1 contig and ORF3 [nt 3098–5710] in pTvERV-K2) are uninterrupted, and encode what appear to be full-length Pol polyproteins. The ORFs of the TvERV-K1 contig and pTvERV-K2 are 871 codons long, although again, some of the amino acids upstream from the *pro-pol* frameshift site will be eliminated from the Gag-Pro-Pol polyprotein. The deduced amino acid sequences are 99% identical.

The TvERV-K Pol amino acid sequences were used to search protein sequence databases. Again, the Pol proteins of the TvERV-K1 contig and TvERV-K2 were most similar to those of the type D retroviruses, followed by those of MMTV, HERV-K and the IAPs. The deduced amino acid sequences of the TvERV-K Pol proteins are aligned with those of three type D retroviruses in FIG. 5.9.

The Pol polyprotein comprises the RT and IN proteins. RT can be further subdivided into polymerase and RNase H domains. The TvERV-K1 contig and pTvERV-K2 Pol proteins possess conserved motifs from all of these domains.

The DNA polymerase domains of the RTs of the TvERV-K1 contig and pTvERV-K2 contain eight conserved motifs found in all retroviral RTs (labelled pol 1 to pol 8 in FIG. 5.9) (Jacobo-Molina and Arnold, 1991). Of these, four (pol 3–pol 6) are conserved in all RNA-dependent RNA and DNA polymerases (Poch *et al.*, 1989). Two (pol 3 and pol 5) are also conserved in all other DNA and RNA polymerases (Delarue *et al.*, 1990), and are presumably directly involved in the catalytic functions of these enzymes

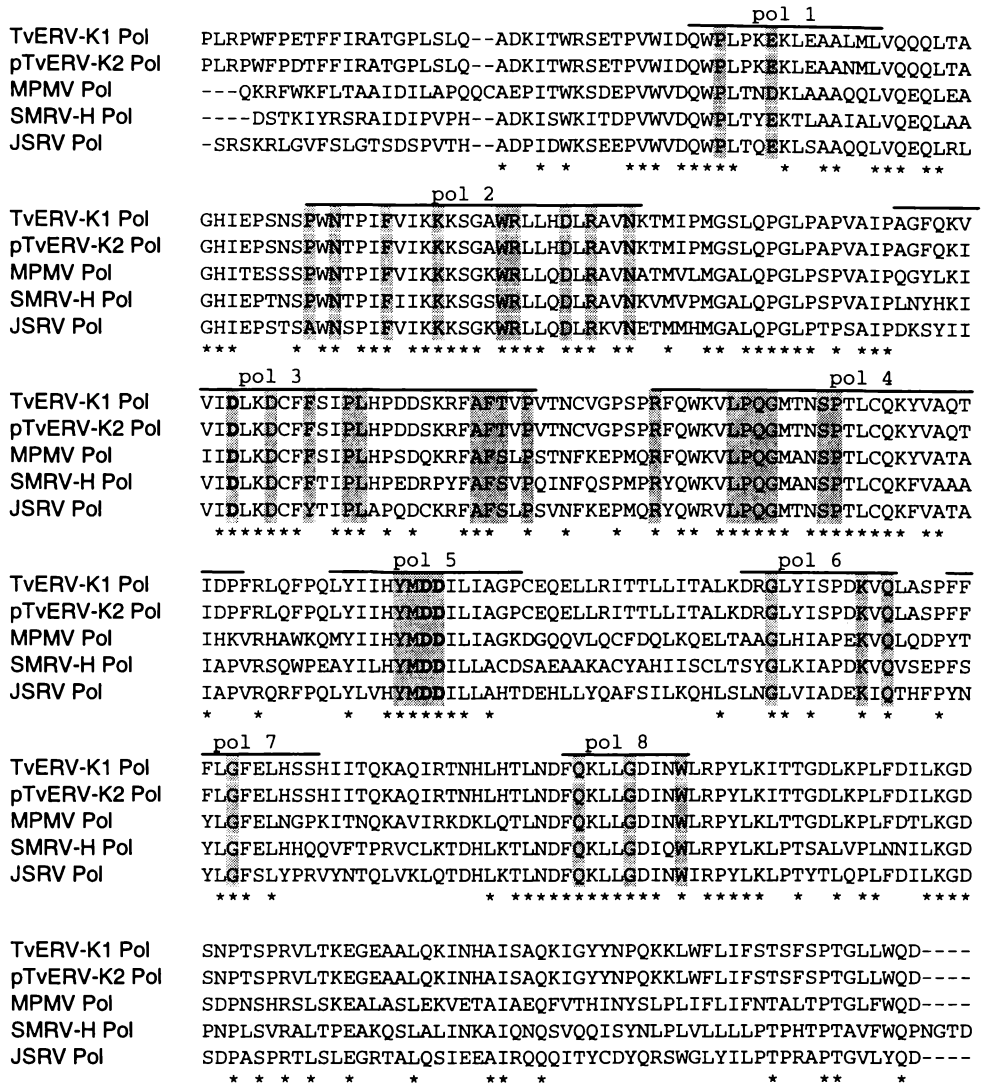


FIGURE 5.9 Pol proteins of the TvERV-K1 contig, pTvERV-K2, and three type D retroviruses. The TvERV-K1 contig (Appendix 4) and pTvERV-K2 (Appendix 5) deduced Pol sequences were aligned with those of three type D retroviruses using the CLUSTAL X program as described in Chapter 2 – Materials and Methods. Note that the sequences are those encoded by the entire ORFs, not the product of translation. Motifs pol 1–pol 8 within the polymerase domain correspond to conserved motifs (with particularly conserved residues shaded, and active site aspartic acid [D] residues in bold) identified in retroviral RTs (Jacobo-Molina and Arnold, 1991). Motifs pol 3–pol 6 within the polymerase domain are conserved among all RNA-dependent polymerases (Poch *et al.*, 1989); motifs pol 3 and pol 5 are conserved in all polymerases (Delarue *et al.*, 1990). Motifs RNase 1–RNase 6 are as described by Jacobo-Molina and Arnold (1991) and McClure (1993); the shaded residues correspond to highly conserved residues, and bold residues were shown to cluster near the active site in HIV-1 RNase H (Davies *et al.*, 1991; McClure, 1993). Conserved residues in the HHCC and D,D(35)E domains of IN are shaded, and are as reported by Katzman (1991), Khan (1991), Katz (1992), and Kulkosky (1992). * indicates identical amino acids in all proteins; - indicates a gap introduced for optimal alignment are indicated by dashes. The type D retroviruses are as listed in FIG. 5.7.

```
TvERV-K1 Pol    -----LPLYWIHGPATPTKVLPTYPLLTASLLQEGGSLSVKFFGKDPDHIVCPYTSAQLS
pTvERV-K2 Pol   -----LPLYWIHGPATPTKVLPTYPLLTASLLQEGRSLSVKFFGKDPDHIVCPYTSSQLS
MPMV Pol        -----NPIMWIHLPASPKKVLLPYYDAIADLIILGRDHSKKYFGIEPSTIIQPYSKSQID
SMRV-H Pol      PTKNGSPLLWLHLPASPSKVLLTYPSLLAMLIIKGRYTGRQLFGRDPHSIIIPYTQDQLT
JSRV Pol        -----KPLRWIYLSATPTKHLLPYYELVAKIIAKGRHEAIQYFGMEPPFICVPYALEQQD
                      *  *    * * * *  *    *      *       **  *  *  **   *

                                                                    RNase 1
TvERV-K1 Pol    WLQQNDDTWAISSVSYQGILDNHYPPDKLIQFLRDNPVAFPRITKRDPLENARLVFTDGS
pTvERV-K2 Pol   WLQQNDDTWAISSVSYQGILDNHYPPDKLIQFLRDNPVAFPRITKRDPLENARLVFTDGS
MPMV Pol        WLMQNTEMWPIACASFVGILDNHYPPNKLIQFCKLHTFVFPQIISKTPLNNALLVFTDGS
SMRV-H Pol      WLLQTSDEWAIALSSFTGDIDNHYPSDPVIQFAKLHQFIFPKITKCAPIPQATLVFTDGS
JSRV Pol        WLFQFSDNWSIAFANYPGQITHHYPSDKLLQFASSHAFIFPKIVRRQPIPEATLIFTDGS
                ** *    * *       *    ***     **       ** *    *   * * *****
                   RNase 2              RNase 3              RNase 4
TvERV-K1 Pol    SSGTAAISIDSNLTSFHTPYKSAQLVELAAILWVFEHVP-DPMNIYTDSAYIAHSIPELE
pTvERV-K2 Pol   SSGTAAISIDSNLTSFHTPYKSAQLVELAAILWVFEHVP-DPMNIYTDSAYIAHSIPELE
MPMV Pol        STGMAAYTLTDTTIKFQTNLNSAQLVELQALIAVLSAFPNQPLNIYTDSAYLAHSIPLLE
SMRV-H Pol      SNGIAAYVIDNQPISIKSPYLSAQLVELYAILQVFTVLAHQPFNLYTDSAYIAQSVPLLE
JSRV Pol        SNGTAALIINHQTYYAQTSFSSAQVVELFAVHQALLTVP-TSFNLFTDSSYVVGALQMIE
                * * **                *** *** *              *  *** *        *
                                                      RNase 5    RNase 6
TvERV-K1 Pol    TVPYIKPST-NAFSMFLQIQKLIRERAQPFFICHIRAHSDLPGPLAKGNDTVDAASKHLV
pTvERV-K2 Pol   TVPYIKPST-NAFSMFLQIQKLIRERAQPFFICHIRAHSDLPGPLAKGNDTVDAASKHLV
MPMV Pol        TVAQIKHIS-ETAKLFLQCQQLIYNRSIPFYIGHVRAHSGLPGPIAQGNQRADLATKIVA
SMRV-H Pol      TVPFIKSST-NATPLFSKLQQLILNRQHPFFIGHLRAHLNLPGPLAEGNALADAATQIFP
JSRV Pol        TVPIIGTTSPEVLNLFTLIQQVLHCRQHPCFFGHIRAHSTLPGALVQGNHTADVLTKQVF
                **  *          *   *     *  *    * ***  ***     **   *
                                          HHCC
TvERV-K1 Pol    FASVLQPSPIETATKAHELHHLNAHSLSQIYKITREQARQIVKQCPGCLVLLPEPHLGVN
pTvERV-K2 Pol   FASVLQPSPIETAIKAHELHHLNAHSLSQIYKITREQARQIVKQCPGCLVLLPEPHLGVN
MPMV Pol        SNIN---TNLESAQNAHTLHHLNAQTLRLMFNIPREQARQIVKQCPICVTYLPVPHLGVN
SMRV-H Pol      IIS----DPIHEATQAHTLHHLNAHTLRLLYKITREQARDIVKACKQCVVATPVPHLGVN
JSRV Pol        FQS-----AIDAARKSHDLHHQNSHSLRLQFKISREAARQIVKSCSTCPQFFVLPQYGVN
                      *   * *** *   *     * ** ** *** *  *      *  ***
                                                                D,D(35)E
TvERV-K1 Pol    PRGLVPGELWQMDVTHFTPFGKLKYIHVSIDTFSGFLIATLQMGEATKHVINHIVASLAM
pTvERV-K2 Pol   PRGLVPGELWQMDVTHFTPFGKLKYIHVSIDTFSGFLIATLQMGEATKHVINHIVASLAM
MPMV Pol        PRGLFPNMIWQMDVTHYSEFGNLKYIHVSIDTFSGFLLATLQTGETTKHVITHLLHCFSI
SMRV-H Pol      PRGLVPNAIWQMDVTHFTPFGKQRFVHVTVDTFSGFILATPQTGEASKNVISHVIHCLAT
JSRV Pol        PRGLRPNHLWQTDVTHIPQFGRLKYVHVSIDTFSNFLMASLHTGESTRHCIQHLLFCFST
                **** *   ** ****   **     **  **** *  *    **     * *

TvERV-K1 Pol    GPKPKILKTDNGPGYVSSSFKRFCAYMNIRNITSIPYNPQGQGIVERAHQTLKRMISKLQ
pTvERV-K2 Pol   GPKPKILKTDNGPGYVSSSFKQFCAYMNIRNITGIPYNPQGQGIVERAHQTLKRMISKLQ
MPMV Pol        IGLPKQIKTDNGPGYTSKNFQEFCSTLQIKHITGIPYNPQGQGIVERAHLSLKTTIEKIK
SMRV-H Pol      IGKPHTIKTDNGPGYTGKNFQDFCQKLQIKHVTGIPYNPQGQGVVERAHQTLKNALNRLA
JSRV Pol        SGIPQTLKTDNGPGYTSRSFQRFCLSFQIHHKTGIPYNPQGQGIVERAHQRIKHQLLKQK
                   *   ********    *  **    *   * ********* *****   *

TvERV-K1 Pol    SGSEKLYIRANNHRTLLNHALFVLNFLTLDAHGKSAADRLWHPSSAHDYAQVKWKDPLTH
pTvERV-K2 Pol   SGSEKLYIRANNHRTLLNHALFVLNFLTLDAHGKSAADRLWHPSSAHDYAQVKWKDPLTH
MPMV Pol        K--GEWYPRKGTPRNILNHALFILNFLNLDDQNKSAADRFWHNNPKKQFAMVKWKDPLDN
SMRV-H Pol      R--SPLGFSMQQPRNLLSHALFQLNFLQLDSQGRSAADRLWHPQTSQQHATVMWRDPLTS
JSRV Pol        K---GNELYSPSPHNALNHALYVLNFLTLDTEGNSAAQRFWGERSSCKKPLVRWKDPLTN
                        * ***  **** **    *** * *          * * ***

TvERV-K1 Pol    QWHGPDPVLIWGKGHACIYDSNAQNARWLPERLLKLVDHAIIKH----------------
pTvERV-K2 Pol   QWHGPDPVLIWGKGHACIYDSNAQNARWLPERLLKLVDHAIIKH----------------
MPMV Pol        TWHGPDPVLIWGRGSVCVYSQTYDAARWLPERLVRQVSNNNQSRE---------------
SMRV-H Pol      VWKGPDPVLIWGRGSACIYDQKEDGPRWLPERLIRHINNQTAPLCDRPSNPNTAPGPKGS
JSRV Pol        LWYGPDPVLIWGRGHVCVFPQDAEAPRWIPERLVRAAEELPDASDATHDPE---------
                 * ********* *  *         ** ****

TvERV-K1 Pol    -
pTvERV-K2 Pol   -
MPMV Pol        -
SMRV-H Pol      P
JSRV Pol        -
```

FIGURE 5.9 *Continued.*

(Delarue *et al.*, 1990). The polymerase domains of the TvERV-K1 contig and pTvERV-K2 also share other regions of significant similarity with the type D retroviruses, for example between the pol 2 and pol 3 motifs, and after the pol 8 motif (FIG. 5.9).

The RNase H domains of the TvERV-K Pol proteins contain six conserved motifs found in all retroviral RNases H (Jacobo-Molina and Arnold, 1991; McClure, 1993) (FIG. 5.9). These motifs include residues that are also conserved in RNases H from other retroelements and in RNases H from bacteria (McClure, 1993; Katz and Skalka, 1994; Telesnitsky and Goff, 1997).

Finally, the TvERV-K1 contig and TvERV-K2 IN proteins contain HHCC and D,D(35)E motifs. The HHCC motif (H-X[3-7]-H-X[23-32]-C-X[2]-C, where X is any amino acid) is located in the N-terminal domain of the IN proteins of retroviruses and other retroelements (Khan *et al.*, 1991; Katz and Skalka, 1994; Brown, 1997). The D,D(35)E motif forms the majority of the central catalytic domain of retroviral INs, and is also present in the INs of other retroelements and the transposases of bacterial insertion sequences (Katzman *et al.*, 1991; Katz *et al.*, 1992; Kulkosky *et al.*, 1992; Polard and Chandler, 1995). The aspartic acid (D) and glutamic acid (E) residues of the D,D(35)E motif are critical components of the active site of INs (for reviews see Katz and Skalka, 1994; Polard and Chandler, 1995; Brown, 1997). The TvERV-K IN proteins share other regions of homology with those of the type D retroviruses, between the HHCC and D,D(35)E motifs and in the carboxy-terminal domain (FIG. 5.9).

5.2.4.4 Env protein

The Env proteins of retroviruses are encoded by an ORF 3' to (although sometimes overlapping) the *pol* ORF. Because both the TvERV-K1 contig and pTvERV-K2 lacked large ORFs downstream of the *pol* ORF (see FIG. 5.1), it was not immediately apparent what, if any, of the 3' sequences encoded Env proteins. BLASTx searches were performed using the nucleotide sequence between the *pol* ORF and the putative PPT of the

TvERV-K1 contig (nt 5969–9424) and the sequence downstream of the *pol* ORF of pTvERV-K2 (nt 5711–9260). These searches revealed a short (247 bp) stretch of sequence (nt 8650–8896 in the TvERV-K1 contig; nt 8518–8764 in pTvERV-K2) with similarity to the transmembrane (TM) domain of the Env proteins of the simian type D retroviruses, baboon endogenous retrovirus (BaEV), avian spleen necrosis virus (SNV) and reticuloendotheliosis virus (REV), and feline endogenous virus RD114. The TM-related sequences in the TvERV-K1 contig and pTvERV-K2 occupied two reading frames. Removal of a single C nucleotide from both sequences (nt 8815 in TvERV-K1 contig; nt 8683 in pTvERV-K2) corrected the frameshift, and allowed alignments with the corresponding regions of other retroviruses (FIG. 5.10). The region of homology between the TvERV-K sequences and those of the other retroviruses included the membrane spanning domain (MSD), which anchors the TM protein in the virus membrane (FIG. 5.10). The TvERV-K sequences and those of the other retroviruses also possessed several conserved residues upstream of the membrane spanning domain (FIG. 5.10).

The distance between the end of the *pol* gene (nt 5968) and the start of the PPT (nt 9425) in the TvERV-K1 contig is 3456 nucleotides (see Appendix 4). The corresponding region in MPMV is 1978 nucleotides (Sonigo *et al.*, 1986), in SRV-1 is 1998 nucleotides (Power *et al.*, 1986), in SRV-2 is 1943 nucleotides (Thayer *et al.*, 1987), and in SMRV-H is 2031 nucleotides (Oda *et al.*, 1988). Thus, the *pol*-PPT distance is considerably (approximately 1500 nucleotides) larger in the TvERV-K1 contig than in the type D retroviruses.

5.2.5 Phylogenetic analysis

Phylogenetic analyses were performed to determine the relationship of the TvERV-K sequences to other retroviruses. Because the *pol* gene products, and in particular the polymerase domain of the RT protein, are the most conserved among retroviruses, analyses were performed using these sequences.

```
                                                        MSD
TvERV-K1 Env      ...SGIVRDKIKILQEVLVQRRKELLNNPLWNGLHGLLPYLLPLLGPLVGLLLLLSLAPGLL
pTvERV-K2 Env     ...SGIVRDKIKKLQKELVQRRKELLNNPLWNGLHGLLPYLLPLLGPLVGLLLLLSLAPGLL
MPMV Env          ...SGIVRDKIKNLQDDLERRRRQLIDNPFWTSFHGFLPYVMPLLGPLLCLLLVLSFGPIIF
SMRV-H Env        ...SGIVRDKIKNLQEDLEKRRKALADNLFLTGLNGLLPYLLPFLGPLFAIILFFSFAPWIL
BaEV Env          ...SGIVRDKIKTLQEELERRRKDLASNPLWTGLQGLLPYLLPFLGPLLTLLLLLTIGPCIF
RD114 Env         ...SGIVRNKIRTLQEELQKRRESLATNPLWTGLQGFLPYLLPLLGPLLTLLLILTIGPCVF
SNV Env           ...SGIVRDKIRKLQEDLIERKRALYDNPLWSGLNGFLPYLLPLLGPLFGLILFLTLGPCIM
REV Env           ...SGIVRDKIRKLQEDLLARKRALYDNPLWNGLNGFLPYLLPLLGPLFGLILFLTLGPCIM
                    ***** **  **  *  *      *       * ***  * ****    *     *

                  ____
TvERV-K1 Env      NKLTTFVKSQVDSAIARKSHIFYSRLLEQDTTEQQEERLQFTKDLPF
pTvERV-K2 Env     NKLTTFVKSQVDSAIARKSHIFYSRLQEEDTTEQQEERLQFTEDLLKEP...
MPMV Env          NKLMTFIKHQIESIQAKPIQVHYHRLEQEDSGGSYLTLT
SMRV-H Env        RRVTALIRDQLNSLLGKPIQIHYHQLATRDLEYGRL
BaEV Env          NRLTAFINDKLNIIHAMVLTQQYQVLRTDEEAQD
RD114 Env         SRLMAFINDRLNVVHAMVLAQQYQALKAEEEAQD
SNV Env           KTLTRIIHDKIQAVKS
REV Env           KTLTRIIHDKIQAVKILALVPQYKPLPTEMDTLGQ
                                  *  *
```

FIGURE 5.10 Alignment of the TM-related sequences of the TvERV-K1 contig and pTvERV-K2 with corresponding sequences from other retroviruses. The TvERV-K1 contig (Appendix 4) and pTvERV-K2 (Appendix 5) deduced TM sequences were aligned with those of other retroviruses using the program CLUSTAL X as described in Chapter 2 – Materials and Methods. The membrane spanning domain (MSD), as described by Einfeld (1996), is indicated. * indicates identical amino acids in all proteins. The other retroviruses are: MPMV, Mason Pfizer monkey virus; SMRV-H, squirrel monkey retrovirus; BaEV, baboon endogenous retrovirus; SNV, spleen necrosis virus; REV, reticuloendotheliosis virus.

The TvERV-K1 contig and pTvERV-K2 Pol proteins were initially compared with representatives of all of the retroviral genera. A 208 amino acid region of the polymerase domains of the RT proteins of the TvERV-K1 contig (nt 3464–4087) and pTvERV-K2 (nt 3206–3829), spanning from the start of the pol 1 domain to within the pol 7 domain in FIG. 5.9, were manually aligned with corresponding sequences from retroviruses representing all of the major genera. The amino acid sequences were aligned according to the alignment published by Xiong and Eickbush (1990), and the resulting alignment is shown in Appendix 6. The aligned sequences were used to construct a phylogenetic tree using the neighbour joining (NJ) method (Saitou and Nei, 1987) within the CLUSTAL X program (Thompson *et al.*, 1997). The resulting tree is shown in FIG. 5.11. Based on the sequences of their polymerase domains, the TvERV-K1 contig and pTvERV-K2 belong to the branch of retroviruses which includes the avian type C, the type B, and the type D genera of retroviruses. Also within this group are the rodent IAP elements and the HERV-K endogenous retroviruses of humans. Within this

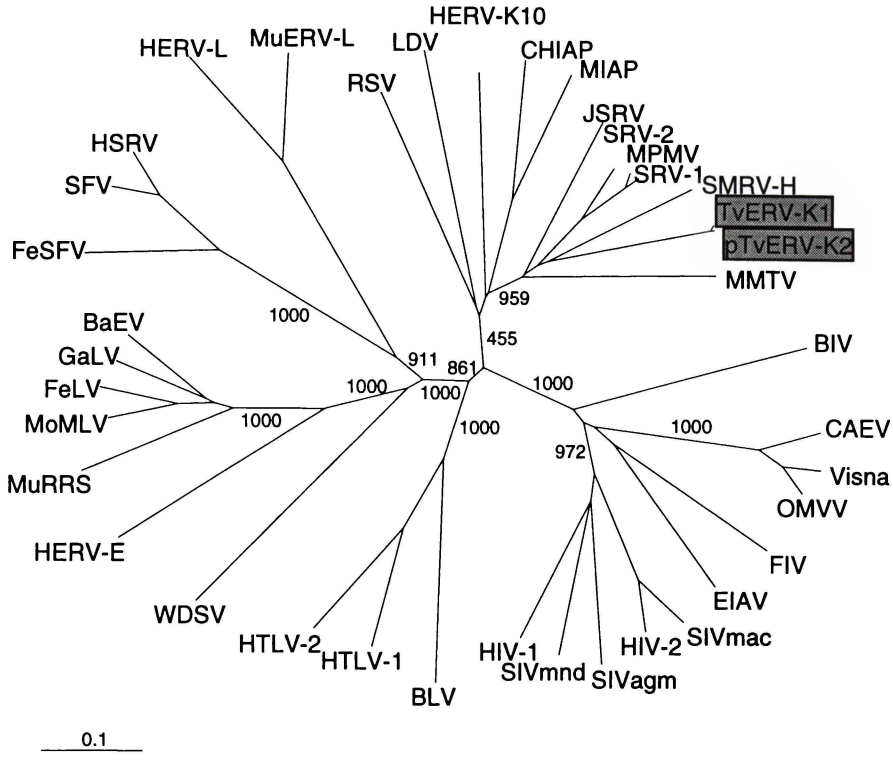


FIGURE 5.11 Phylogenetic relationship of the TvERV-K1 contig and pTvERV-K2 to other retroviruses based on their RT sequences. Amino acid sequences within the polymerase domains of the RT proteins of the TvERV-K1 contig and pTvERV-K2 Pol were aligned with those of other retroviruses (see Appendix 6). This alignment was used to construct a phylogenetic tree using the NJ method (see Chapter 2 – Materials and Methods), which was further analysed using bootstrapping analysis (1000 replicates). Bootstrap scores for some of the major branches are shown. Branch-lengths are proportional to evolutionary distance between sequences, as indicated by the scale. Full names of retroviruses are listed in Appendix 1.

large group, there is strong bootstrap support (bootstrap value = 959) for the placement of the TvERV-K elements within a smaller group of retroviruses. This group includes the type B retroviruses (MMTV), the type B/D retroviruses of sheep and goats (JSRV), the endogenous type D retrovirus of New World monkeys (SMRV-H), and the exogenous type D retroviruses of Old World monkeys (MPMV, SRV-1, SRV-2). Bootstrap scores within the type B and type D group did not allow the precise placement of the TvERV-K1 contig and pTvERV-K2 (data not shown).

In their analysis of the distribution of endogenous retroviruses in vertebrates, Herniou *et al.* (1997) reported two sequences, one from the bower bird and one from the stripe faced dunnart, which placed with the group containing

avian type C, type B, and the type D retroviruses. The relationships between the TvERV-K1 contig and pTvERV-K2, the bower bird and dunnart sequences, and other retroviruses within this large group were analysed in more detail. The analysis was restricted to the region of sequence reported by Herniou *et al.*, which spanned the active site D(S/T)G motif of the Pro protein (PR 1 in FIG. 5.8) at the N-terminus and the YXDD motif of the RT protein (pol 5 in FIG. 5.9) at the C-terminus. Thus, two regions of sequence were used. The first region comprised 30 amino acids from 15 amino acids N-terminal to 12 amino acids C-terminal of the conserved GR(D/N) motif (PR 3, FIG. 5.8) of the Pro protein (nt 3156–3245 of the TvERV-K1 contig and nt 2898–2987 of pTvERV-K2). The second region comprised 166 amino acids from just upstream of the pol 1 motif to the pol 5 motif (FIG. 5.9) of the polymerase domain of the RT protein (nt 3449–3946 of the TvERV-K1 contig and nt 3191–3688 of pTvERV-K2). These regions of the possum sequences were manually aligned, using the alignment published by Herniou *et al.* (1997) as a guide, with the corresponding sequences from the bower bird and dunnart sequences reported by Herniou *et al.* (1997). Also included were the corresponding sequences from the avian type C retroviruses, rodent IAP elements, HERV-K endogenous retroviruses, and the type B and type D retroviruses (see Appendix 7). A phylogenetic tree (FIG. 5.12) was generated from the alignment using the CLUSTAL X program, and using the HTLV/BLV group viruses as an outgroup. As shown in FIG. 5.12, the RV bowerbird sequence reported by Herniou *et al.* (1997) groups with the avian type C retroviruses and the RV stripe faced dunnartI sequence appears to belong to a group comprising the HERV-K elements and the type B and type D retroviruses. However, the bootstrap support for these relationships is very weak, and therefore casts doubt on the branching orders in FIG. 5.12. Nevertheless, there is strong support (bootstrap value = 933) for an association of the TvERV-K sequences with the type B and type D retroviruses, to the exclusion of the other sequences.

The trees in FIG. 5.11 and FIG. 5.12 both strongly suggest that the TvERV-K1 contig and pTvERV-K2 belong to the type B/type D lineage of retroviruses. However, neither of these trees allow conclusive placement of the TvERV-K sequences within this lineage. In an attempt to resolve this issue, a larger

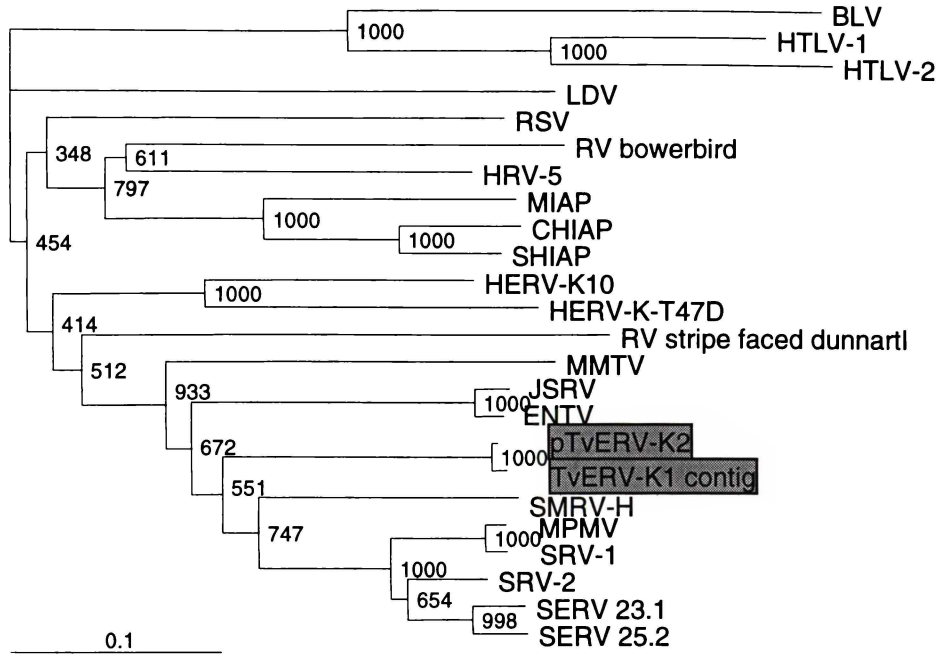


FIGURE 5.12 Phylogenetic relationship of the TvERV-K1 contig and pTvERV-K2 to the avian type C and mammalian type A, type B, and type D retroviruses based on amino acid sequences of their PR and RT proteins. Amino acid sequences from the Pro and Pol proteins of TvERV-K1 contig and pTvERV-K2 were manually aligned with those of other retroviruses according to the alignment of Herniou *et al.* (1997) and as shown in Appendix 7. This alignment was used to construct a phylogenetic tree using the NJ method (see Chapter 2 – Materials and Methods) which was further inspected using bootstrapping analysis. The tree was routed using the relevant sequences from HTLV-1, HTLV-2, and BLV. Branch lengths are proportional to evolutionary distances between sequences, as indicated by the scale. Full names of retroviruses are listed in Appendix 1.

sequence range was used. Again, amino acid sequences from both the Pro and Pol proteins were used. In this case, however, the Pro sequence was from just N-terminal to the DU 1 motif to just C-terminal to the PR 3 motif (FIG. 5.8) of Pro (nt 2577–3254 of the TvERV-K1 contig and nt 2319–2996 of pTvERV-K2), and included both the DU and PR) domains of Pro (the avian type C retroviruses lack DU, and could therefore not be included in this analysis). The Pol sequence was from upstream of the pol 1 motif to downstream of the D,D(35)E motif of IN (FIG. 5.9) (nt 3425–5878 of the TvERV-K1 contig and nt 3167–5620 of pTvERV-K2), and therefore included sequences from the polymerase and RNase H domains of RT, as well as the HHCC, D,D(35E), and carboxy terminal domains of IN. In total, 944 amino acids of sequence from the proteins of each of TvERV-K1 contig and TvERV-

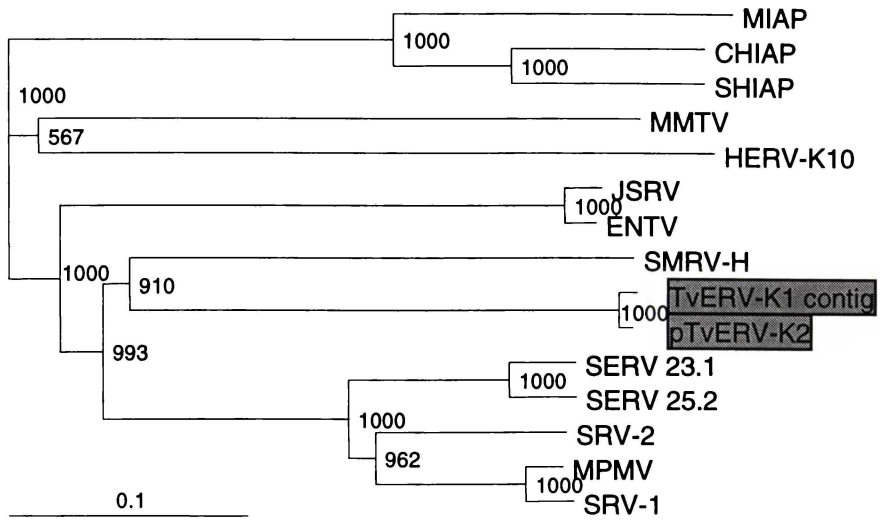
K2 were used. The TvERV-K Pro and Pol sequences were aligned separately with the corresponding sequences from the IAPs, the HERV-K elements, and the type B and type D retroviruses. The aligned Pro and Pol sequences were then concatenated, as shown in Appendix 8, and were used to construct trees using both NJ and maximum parsimony methods (FIG. 5.13). The overall structures of the NJ and maximum parsimony trees are very similar, the only difference being that HERV-K10 and MMTV form a separate group with moderate bootstrap support (bootstrap value = 567) in the NJ tree, whereas they branch off from the type B/typeD lineage sequentially in the maximum parsimony tree. Both trees group the TvERV-K1 contig and pTvERV-K2 with the simian type D retroviruses, to the exclusion of the type B retroviruses (MMTV) and the type B/type D retroviruses of sheep and goats (JSRV, ENTV). In particular, the TvERV-K sequences group with SMRV-H, an endogenous retrovirus in the New World squirrel monkey. The TvERV-K/SMRV-H group is well-supported (bootstrap value = 910) in the NJ tree, although, as the branch lengths in this tree indicate, the relationship is a distant one.

5.3 Discussion

Analysis of the structures and sequences of retroviral genomes can provide information on the likely biological activity of the corresponding retroviruses. Most of the regulatory elements of retroviral genomes contain conserved nucleotide sequences that reflect their importance for the replication of the retrovirus genome and their dependence on proteins, both viral and cellular, for replication. Conserved amino acid sequence motifs are also found in most of the retroviral proteins that, again, reflect the functional constraints on these proteins. Analysis of retroviral sequences also permits deduction of the evolutionary history of a given retrovirus.

The regulatory regions of the TvERV-K elements suggest that they are competent for transcription, polyadenylation, splicing, translation (including frameshifts), reverse transcription and integration.

(a) NJ tree



(b) Maximum parsimony tree

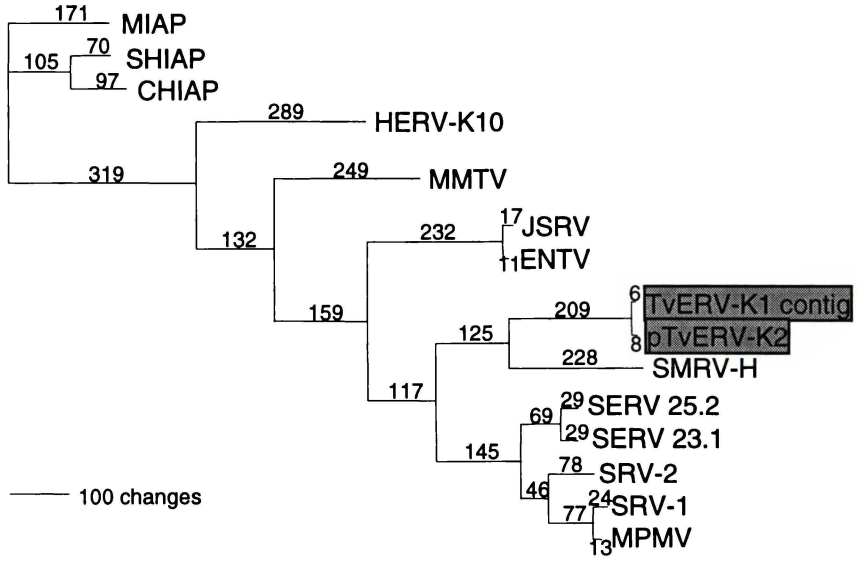


FIGURE 5.13 Phylogenetic relationship of the TvERV-K1 contig and pTvERV-K2 to the mammalian type A, type B, and type D retroviruses based on their Pro and Pol amino acid sequences. Sequences from the Pro and Pol proteins were aligned as described in the text and in Appendix 8. The alignment was used to construct trees using Neighbour-Joining (a) and maximum parsimony (b) methods as described in Chapter 2 – Materials and Methods. For the NJ tree (a), the numbers correspond to the bootstrap values (1000 replicates) and the branch lengths are proportional to the evolutionary distances between sequences as indicated in the scale. The numbers in the maximum parsimony tree (b) correspond to the number of amino acid differences between sequences.

The TATA box and CCAAT box within the TvERV-K elements probably direct transcription initiation. Additional, as yet unidentified, sequences upstream of the CCAAT may further influence transcription initiation. Polyadenylation of the TvERV-K RNA is almost certainly directed by the poly(A) signal.

A potential splice donor site, which matched the consensus eukaryotic splice donor site, was located immediately upstream from the *gag* gene. Retroviral splice donors are normally located in the 5' untranslated region, so that the *gag*, *pro*, and *pol* are removed during splicing, although they are sometimes located within the *gag* gene (Swanstrom and Wills, 1997). Thus, the sequence and position of this sequence make it a likely candidate for splice donor. However, the lack of an obvious *env* ORF prevented the identification of a splice acceptor for the *env* gene.

The overlapping region of the *gag* and *pro* genes contains a potential frameshift slippery sequence for transition from the *gag* to *pro* reading frames. Also, a pseudoknot structure, similar to that shown to be important for frameshifting in the type D retrovirus MPMV (ten Dam *et al.*, 1994), may form just downstream of the frameshift slippery sequence in the TvERV-K RNAs. No perfect slippery sequence was apparent in the region overlapping the *pro* and *pol* genes. Note, however, that the *pro-pol* frameshift site in MMTV (GGAUUUA) does not strictly conform to the consensus, and yet frameshifts still occur at this site (Hatfield and Oroszlan, 1990). It may therefore be possible that frameshifting does occur within the TvERV-K *pro-pol* overlaps despite the absence of a perfect frameshift slippery sequence.

An 18 nucleotide sequence in the TvERV-K1 contig is perfectly complementary to the 3' 18 nucleotides of mammalian $tRNA^{Lys}_{1,2}$. The same sequence is present in pTvERV-K2 except that it contains a single G insert. Clearly, the TvERV-K elements use (or used) $tRNA^{Lys}_{1,2}$ to prime minus-strand synthesis during reverse transcription. The exogenous simian type D retroviruses (MPMV, SRV-1, SRV-2), an endogenous simian type D retrovirus (SMRV-H), exogenous and endogenous type B/type D retroviruses in sheep and goats (JSRV, ENTV), human foamy virus (HFV, a spumavirus),

and Visna Maedi virus (VMV, a lentivirus) also use $tRNA^{Lys}_{1,2}$ to prime minus strand synthesis during reverse transcription (Petropoulos, 1997).

One of the systems for naming human endogenous retroviruses incorporates the tRNA species that is used to prime reverse transcription in the name (Boeke and Stoye, 1997). In accordance with this system, I have named the possum endogenous retroviruses described here TvERV-K, where the K indicates that a lysine tRNA is used to prime reverse transcription.

The putative PPT was assigned because of its position within the TvERV-K1 contig genome, and because of its similarity to the PPTs in other retroviruses. Delineation of the 5' boundary of the LTRs (see above) would permit the definitive identification of the PPT.

The small inverted repeats (IR), TG at the 5' end and CA at the 3' end of the LTR, are likely to act as recognition signals for the retroviral integrase during cleavage of the ends of the virus dsDNA genome and integration of the genome into the DNA of the host cell.

Some regulatory regions remain to be identified in the TvERV-K genomes. A packaging signal would be expected near the 5' end of the RNA genome, upstream (and possibly overlapping) the *gag* gene (Rabson and Graves, 1997). In addition, a constitutive transport element (CTE), required for transport of unspliced RNA out of the nucleus, might also be expected to be located near the 3' end of the RNA genome (Rabson and Graves, 1997). Despite the fact that neither of these sequence elements have yet been identified, the amplification and cloning of TvERV-K fragments from apparently particle-associated RNA in possum blood plasma (see 4.2.1 3'-RACE to amplify the 3' end of the retroviral genome) suggests that they are present.

Analysis of the sequences of both the TvERV-K1 contig and pTvERV-K2 revealed several long open reading frames (ORFs), located in the 5' two thirds of the sequences. Database searches and comparisons of the amino acid sequences of these ORFs with those of retroviral proteins confirmed that

the TvERV-K1 contig and pTvERV-K2 encoded Gag, Pro, and Pol proteins (see below). The arrangement of the ORFs suggested that frameshifts at the Gag-Pro and Pro-Pol overlapping regions would be required for synthesis of Gag-Pro and Gag-Pro-Pol polyproteins. This ORF arrangement is reminiscent of those of the type B, the type D, and the HTLV/BLV retroviruses, all of which require two translational frameshifts for synthesis of the Gag-Pro-Pol polyprotein.

The Gag ORF of pTvERV-K2 is uninterrupted. However, that of the TvERV-K1 contig contains two in-frame stop codons and a single nucleotide insertion that causes a frameshift mutation, and probably does not encode a functional Gag protein.

The Gag protein alone is sufficient for the formation of retroviral particles (see Dickson *et al.*, 1982; Klikova *et al.*, 1995). Thus, pTvERV-K2 encodes a Gag protein that, on the basis of its amino acid sequence, appears to be capable of forming retrovirus particles. The MA protein is most likely myristylated at its N-terminus, and is therefore probably membrane associated both before and after virus particle maturation. The MHR region within the putative CA protein of the pTvERV-K2 Gag may function during particle maturation and infection of host cells (Craven *et al.*, 1995; Craven and Parent, 1996). The intact CCHC motifs within the C-terminal NC protein suggest that the pTvERV-K2 Gag would package retroviral RNA into particles (Berkowitz *et al.*, 1996; Vogt, 1997b; Rein *et al.*, 1998). The high number of basic residues in NC may enable it to promote maturation of the RNA in the virus particle and improve the efficiency of reverse transcription and strand transfer during replication of the retroviral genome (Berkowitz *et al.*, 1996; Vogt, 1997b; Rein *et al.*, 1998).

The *pro* ORFs of the TvERV-K1 contig and pTvERV-K2 are uninterrupted. The products of these ORFs appear to comprise dUTPases (DUs) at their amino termini and proteases (PRs) at their carboxy termini. Rodent IAPs, HERV-K endogenous retroviruses, type B retroviruses, and type D retroviruses also encode DUs at the 5' ends of their *pro* genes (McClure *et al.*, 1987; Elder *et al.*, 1992). Both the DU and PR domains of the TvERV-K

Pro proteins possessed conserved domains present in the DUs and PRs of other retroviruses. Thus, the TvERV-K1 contig and pTvERV-K2 appear to encode functional dUTPases, which would degrade dUTP in the virus particle, as well as functional PRs, which would cleave the Gag, Gag-Pro, and Gag-Pro-Pol polyproteins into their separate components during maturation of the virus particle (Vogt, 1996; Swanstrom and Wills, 1997). However, although the TvERV-K1 contig *pro* ORF is intact, it would probably not be translated due to termination and frameshift mutations within the *gag* gene (see above).

The *pol* ORFs of the TvERV-K1 contig and pTvERV-K2 are also intact. Again, however, the TvERV-K1 contig *pol* ORF would not be translated due to the mutations in the *gag* gene. Retrovirus Pol polyproteins comprise RT and IN proteins, and RT can be further divided into polymerase and RNase H domains. The polymerase and RNase H domains of the TvERV-K RT proteins, possessed motifs conserved in the same domains of all other retroviral RTs. Similarly, the IN proteins from the TvERV-K1 contig and pTvERV-K2 possessed conserved IN motifs. Thus, it appears that the possum elements encode fully functional RT and IN proteins. The RT protein may be responsible for the reverse transcriptase activity detected in the PERT assays performed on possum blood (Chapter 3).

The TvERV-K1 contig and pTvERV-K2 sequences lack long *env* ORFs. However, short sequences near the 3' ends of the TvERV-K elements translate into amino acid sequences that are very similar to the membrane spanning regions of the TM proteins of several retroviruses, including the simian type D retroviruses, endogenous mammalian type C retroviruses in baboons (BaEV) and cats (RD114), and exogenous mammalian type C-related retroviruses in birds (SNV, REV). The nature of the *env* sequences in the TvERV-K elements makes it unlikely that they are expressed as functional proteins.

The estimated length from the 5' end of the R region of the 5' LTR (nt 219) to 3' end of the R region of the 3' LTR (nt 9674) of the TvERV-K1 contig is 9456 nt. The corresponding length of MPMV is approximately 7835 nt

(Petropoulos, 1997), that of SRV-1 is approximately 7848 nt (Petropoulos, 1997), that of SRV-2 is approximately 7785 nt (Thayer *et al.*, 1987), that of SERV clone 23.1 is approximately 7940 nt (van der Kuyl *et al.*, 1997), and that of SMRV-H is 8221 nt (Oda *et al.*, 1988). Thus, the R–R distance of TvERV-K1 contig is approximately 1200–1600 nt longer than those of the simian type D retroviruses. Likewise, the distance between the end of the *pol* gene and the start of the PPT in the TvERV-K1 contig is approximately 1500 nucleotides longer than the corresponding region in the simian type D retroviruses. Therefore, it seems likely that rearrangements in the 3' region of the genomes of the TvERV-K elements have contributed the most towards the differences in lengths between the TvERV-Ks and the simian type D retroviruses.

The evolution of the *env* gene of the simian type D retroviruses is complex. Although the *pol* genes of simian type D retroviruses clearly shared a recent common ancestor with those of MMTV and JSRV, the *env* genes of the simian type D retroviruses are unrelated to those of MMTV and JSRV, which are related to each other (York *et al.*, 1992). As shown in FIG. 5.10, the *env* genes of simian type D retroviruses are, in fact, more closely related to those of the exogenous reticuloendotheliosis virus (REV) family in birds, the baboon endogenous virus (BaEV), and a cat endogenous retrovirus (RD114). It has been proposed that the simian type D retroviruses arose by a recombination event, which resulted in the acquisition of a REV-like envelope gene by a retrovirus of the type B/type D lineage (York *et al.*, 1992). The evolution of the simian type D retroviruses is discussed in more detail in Chapter 7.

In summary, then, the TvERV-K endogenous retroviruses in possums appear to be capable of significant biological activity. They appear to possess regulatory regions that would enable transcription, polyadenylation, splicing, translation (including frameshifts), reverse transcription, and integration of the TvERV-K genomes. In addition, they appear to encode Gag proteins that would package RNA and form retrovirus particles, Pro proteins that could mediate maturation of those particles, and RT and IN proteins that would reverse transcribe the TvERV-K genome and integrate it into the genome of

the host cell, respectively. Neither the TvERV-K1 contig nor pTvERV-K2 encode functional Env proteins, and therefore could only produce infectious particles if an Env protein was provided *in trans*.

Phylogenetic analyses were performed using deduced amino acid sequences from various regions of the TvERV-K1 contig and pTvERV-K2 elements. Initially, an NJ tree generated from an alignment of the polymerase domains of the RT proteins from representatives of all retroviral genera placed the TvERV-K1 contig and pTvERV-K2 within a group of retroviruses that also included the avian type C retroviruses, rodent IAP elements, HERV-K endogenous human retroviruses, type B retroviruses, and type D retroviruses. Within this group, there was strong support for a closer relationship of the TvERV-K1 contig and pTvERV-K2 with the type B and type D viruses to the exclusion of the other members of the group. Further analysis of this group of retroviruses, as well as some related sub-genomic sequences, using sequences from the carboxy-terminus of Pro and the amino-terminus of Pol, again placed the TvERV-K elements with the type B and type D retroviruses. Another marsupial sequence reported by Herniou *et al.* (1998), RV stripe faced dunnartl, appeared to group with HERV-K, the type B and type D retroviruses, and the TvERV-K1 contig and pTvERV-K2, although the levels of bootstrap support cast uncertainty over their precise relationships. Phylogenetic analyses using both NJ and maximum parsimony approaches, and an extended region of amino sequences spanning almost all of the Pro and Pol proteins, placed the TvERV-K1 contig and pTvERV-K2 exclusively with the simian type D retroviruses. Closer homology to the endogenous retrovirus of the New World squirrel monkey to the exclusion of the endogenous and exogenous Old World monkey type D retroviruses (SERVs, MPMV, SRV-1, SRV-2) was suggested.

Thus, all the evidence presented here suggests that the TvERV-K1 contig and pTvERV-K2 should be classified as species of the type D retrovirus genus. The organisation of their ORFs, requiring frameshifts between the *gag* and *pro* and the *pro* and *pol* reading frames for synthesis of the Gag-Pro-Pol polyprotein, is the same as that of the type B and type D retroviruses. They use (or used) a $tRNA^{Lys}_{1,2}$ to prime minus strand synthesis

during reverse transcription, a feature they have in common with most of the type D retroviruses. The TvERV-K1 contig and pTvERV-K2 encode a DU protein immediately upstream from the PR protein, a characteristic of the IAPs, HERV-K, type B retroviruses, and type D retroviruses. Database searches using the TvERV-K Gag, Pro, and Pol amino acid sequences, and a short stretch of sequence corresponding to part of the TM potein, all revealed highest homology to the simian type D retroviruses. Finally, phylogenetic analyses using the amino acid sequences of the Pro and Pol proteins conclusively grouped TvERV-K1 contig and pTvERV-K2 with the simian type D retroviruses.

Chapter 6

Southern Analysis of Possum Genomic DNA Using TvERV-K Probes

6.1 Introduction

Southern analysis is a particularly useful tool for the analysis of the distribution and organisation of specific endogenous retroviruses within the genomes of animals. Southern analysis can be used to estimate the copy number of an endogenous retrovirus, to detect structural rearrangements, and to analyse the distribution of an endogenous retrovirus in the genomes of individuals of one species or of different species.

Two classes of fragments can be detected by Southern analysis of genomic DNA using retroviral probes: internal fragments; and junction fragments (FIG. 6.1) (Boeke and Stoye, 1997). Internal fragments contain only viral sequences, and are used to detect structural rearrangements in proviruses by what is essentially restriction fragment length polymorphism (RFLP)

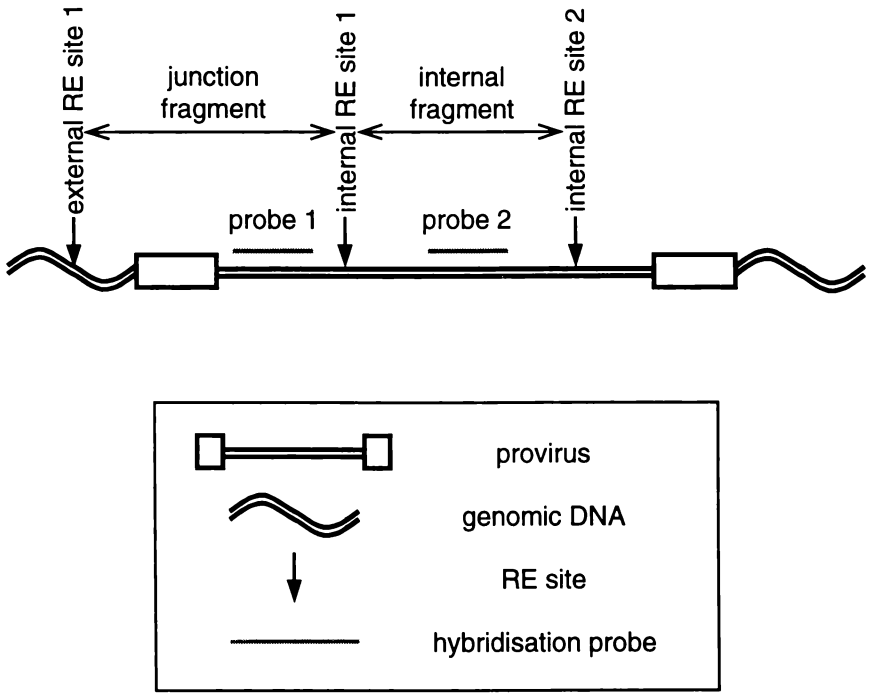


FIGURE 6.1 Analysis of endogenous retroviruses by hybridisation with retroviral probes. Restriction enzyme digestion of genomic DNA will produce junction fragments, which contain proviral DNA and flanking genomic DNA, and internal fragments, which contain only proviral DNA. Junction fragments can be detected by hybridisation using a probe derived from sequence between the internal RE sites. Junction fragments can be detected using a probe that lies between an internal RE site and the end of the provirus. See text for further details.

analysis (FIG. 6.1). Junction fragments contain both viral DNA and flanking cellular DNA. Because each provirus is integrated at a different site within the host cell genomic DNA, the junction fragments of different proviruses will vary in size depending on the distance between the end of the provirus and the nearest restriction enzyme site in the genomic DNA. Thus, separation of the junction fragments by agarose gel electrophoresis of the digested DNA, Southern blotting, and hybridisation with a proviral probe can be used to estimate the copy numbers of endogenous retroviruses and to identify individual proviruses (FIG. 6.1) (Boeke and Stoye, 1997).

In order to gain some preliminary information on TvERV-K provirus copy number and structure, and the distribution of TvERV-K elements in the genomes of possums, a limited number of Southern analyses were carried out.

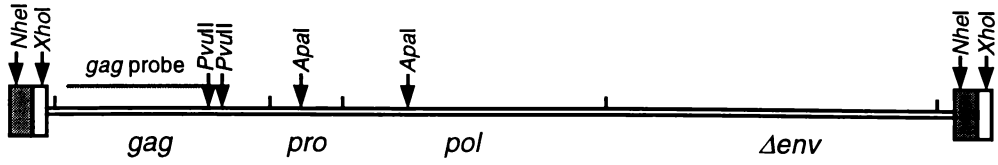
6.2 Results

6.2.1 TvERV-K internal fragments

Internal fragments and junction fragments of TvERV-K were analysed using a probe from the *gag* gene of pTvERV-K2 (FIG. 6.2). Analysis of the sequences of the TvERV-K1 contig and pTvERV-K2 revealed two restriction enzymes (*Nhe*I, *Xho*I) that cut only within the LTRs of the TvERV-K elements. *Nhe*I (recognition site G/CTAGC) cuts within the U3 region (nt 62–67 and nt 9505–9510 of the TvERV-K1 contig) and *Xho*I (recognition site C/TCGAG) cuts within the U5 region (nt 307–312 and nt 9749–9754 of the TvERV-K1 contig, and nt 62–67 of pTvERV-K2) (FIG. 6.2). Thus, based on the sequences of the TvERV-K1 contig, digestion of possum genomic DNA with *Nhe*I or *Xho*I and hybridisation with the *gag* probe would be expected to detect ~9.4kb fragments.

Hybridisation of the *gag* probe to *Xho*I and *Nhe*I digested genomic DNA from three possums is shown in FIG. 6.3. Complete digestion with *Xho*I could not be achieved, possibly due to methylation of the C residue in the central CG dinucleotide in the *Xho*I recognition sites in the possum genome.

(a) TvERV-K1 contig



(b) pTvERV-K2

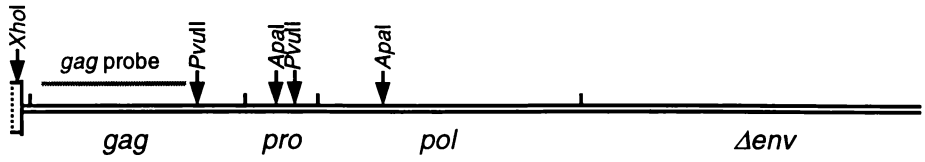


FIGURE 6.2 Restriction enzyme sites in the TvERV-K1 contig (a) and pTvERV-K2 (b), and the position of the *gag* hybridisation probe. The *gag* probe (nt 398–1831 of pTvERV-K2, Appendix 5) was generated by PCR amplification of pTvERV-K2 using the TvP05F and TvP07R primers (see Materials and Methods), and will detect internal fragments following digestion with *Nhe*I or *Xho*I, and junction fragments following digestion with *Apa*I or *Pvu*II.

Nevertheless, with both *Xho*I and *Nhe*I, ~9.4kb restriction fragments could be observed. In addition, both restriction enzymes yielded restriction fragments approximately 6.5–7.5kb in length (FIG. 6.3).

6.2.2 TvERV-K junction fragments

The number of copies of TvERV-K in the genomes of possums and the distribution of TvERV-K in possums were investigated using the *gag* probe and the restriction enzymes *Apa*I and *Pvu*II. *Apa*I (recognition site GGGCC/C) cuts at the same sites within the TvERV-K1 contig (nt 2923–2928 and nt 3959–3964) and pTvERV-K2 (nt 2655–2670 and nt 3701–3706) (see FIG. 6.2). *Pvu*II (recognition site CAG/CTG) has one site in common between the TvERV-K1 contig (nt 2270–2275) and pTvERV-K2 (nt 2012–2017), as well as one site in the TvERV-K1 contig (nt 2017–2022) which is absent in pTvERV-K2, and one site in pTvERV-K2 (nt 2859–2864) which is absent in the TvERV-K1 contig (see FIG. 6.2). The *gag* probe lies almost

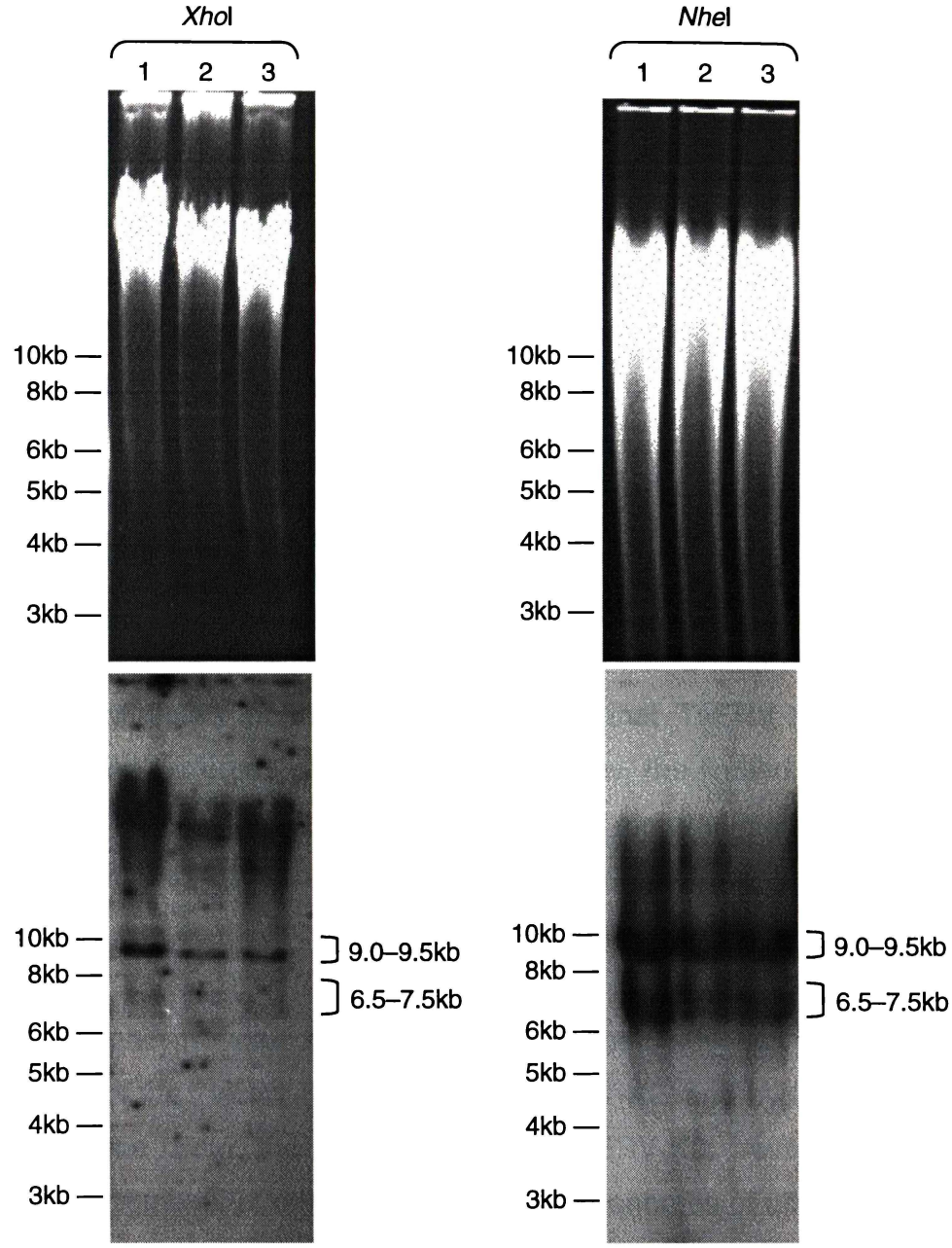


FIGURE 6.3 Detection of TvERV-K variants in the genomes of possums. Genomic DNA (10 µg) from three possums (1–3) was digested with *Xho*I (left) and *Nhe*I (right) and electrophoresed in 0.7% agarose (top). The DNA was blotted to a Hybond N+ membrane, and the membrane was hybridised to the *gag* probe (see FIG. 6.2), washed to moderate stringency, and exposed to XAR-5 film (bottom) as described in Materials and Methods.

entirely 5' to the 5' most *Pvu*II site in the TvERV-K1 contig (the 254bp *Pvu*II fragment is unlikely to be detected by hybridisation with the *gag* probe, with which it overlaps by 69bp), and entirely 5' to the 5' most *Pvu*II site in pTvERV-K2 (see FIG. 6.2).

Hybridisation of the *gag* probe to *Apa*I and *Pvu*II digested genomic DNA from three possums is shown in FIG. 6.4. Both *Apa*I and *Pvu*II generated 15–20 junction fragments for each possum genomic DNA sample. However, the intensity of the bands varied within each possum (FIG. 6.4). In addition, the patterns of restriction fragments generated by both enzymes differed between possums (FIG. 6.4), although four of the fragments in the *Apa*I and *Pvu*II digests appeared to be conserved in all three possums (indicated by arrows in FIG. 6.4). Possible explanations for these observations are discussed below.

6.3 Discussion

As suggested by previous PCR results (see Chapter 4), the DNA hybridisation results presented here suggest that TvERV-K elements are present in the genomes of all possums. Clearly, the limited data presented here needs to be extended by examining many more animals with a number of restriction enzymes and probes. Nevertheless, a number of important conclusions can be made.

Restriction enzyme digestion with enzymes that only cut within the LTR sequences of the TvERV-K1 contig and pTvERV-K2, and hybridisation with a probe derived from the *gag* gene of pTvERV-K2, revealed TvERV-K restriction fragments of different lengths in the genomes of all possums. The ~9.0–9.5kb fragments probably correspond to the elements from which the TvERV-K1 contig and pTvERV-K2 were derived. The shorter (~6.5–7.5kb) fragments might represent TvERV-K elements that have acquired additional restriction enzyme sites near their 3' ends. However, the fact that these shorter fragments were generated using two different restriction enzymes counts against this possibility, and suggests that they were in fact generated by digestion exclusively within the LTRs. This implies that TvERV-K elements substantially (~1.5–3.0kb) shorter than those from which the TvERV-K1 contig and pTvERV-K2 were derived reside within the genomes of all possums. Further support for this possibility comes from the observation that PCR amplification of possum genomic DNA using the TvP03F and

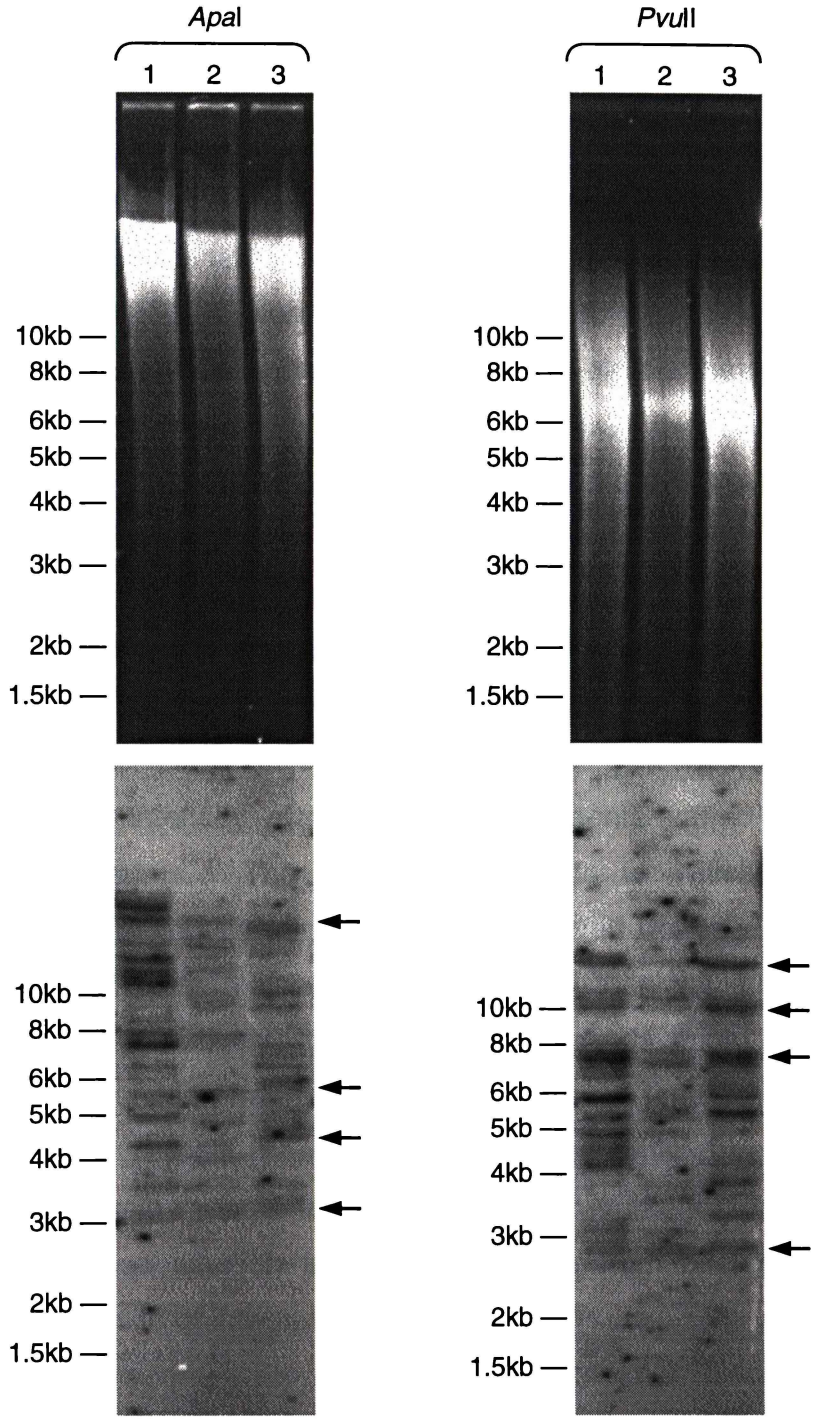


FIGURE 6.4 Detection of TvERV-K junction fragments in the genomes of possums. Genomic DNA (10 µg) from three possums (1–3) was digested with *Apa*I (left) and *Pvu*II (right) and electrophoresed in 0.7% agarose (top). The DNA was blotted to a Hybond N+ membrane, and the membrane was hybridised to the *gag* probe (see FIG. 6.2), washed to moderate stringency, and exposed to XAR-5 film (bottom) as described in Materials and Methods. Arrows indicate junction fragments of similar size in the DNA of all three possums.

TvP27R primers also yielded fragments in the 6.5–7.5kb size range (see FIG. 4.9).

As discussed in Chapter 5, the two TvERV-K sequences analysed so far are approximately 1200–1600 nt longer than the simian type D retroviruses. Most of this difference appears to be due to rearrangements in the *env* genes of the TvERV-K elements relative to those of the type D retroviruses. Thus, the shorter TvERV-K elements that appear to reside within the possum genome may represent more ancestral TvERV-Ks that have fewer mutations within their *env* genes. It is conceivable, given the intact nature of the TvERV-K *gag*, *pro*, and *pol* genes, that one or more of the shorter TvERV-K elements also possesses an intact *env* gene.

TvERV-K junction fragments were generated using two different restriction enzymes and detected by hybridisation with the *gag* probe. Approximately 15–20 copies of TvERV-K were detected in the genomes of three possums. However, the intensity of the junction fragments varied both within and between possums. The junction fragments to which the *gag* probe hybridised more weakly might have been derived from TvERV-K proviruses that are more distantly related to pTvERV-K2 than the strongly hybridising fragments. Alternatively, the more intense bands might comprise two or more distinct junction fragments of similar size. A final possibility is that the more intense bands arise from allelic genomic fragments that are homozygous with respect to the site of provirus integration, as opposed to the less intense junction fragments that are from alleles that differ with respect to the site of integration.

Several of the TvERV-K junction fragments appeared to be conserved in all three possums. These fragments may be derived from unique integration events in each of the possums that, by chance, generate junction fragments of similar sizes. However, they may also represent TvERV-K integration events that occurred in an evolutionary ancestor of the possum and that are therefore present at the same positions in the genomes of all possums. If this is the case, TvERV-K elements would be regarded as ancient endogenous retroviruses of possums.

The patterns of the majority of the junction fragments are highly variable between possums. Three possible phenomena, or a combination of them, could account for the variability in TvERV-K junction fragments between possums: mutation of the restriction sites within the TvERV-K elements and/or the flanking DNA; homologous recombination between TvERV-K LTRs resulting in the loss of a detectable TvERV-K junction fragment; or integration of new TvERV-K copies at different sites of the germline DNA of different possums.

Some of the variability in TvERV-K junction fragments between possums could be attributed to restriction fragment polymorphisms. Thus, possums, when they evolved, might already have possessed a set of TvERV-K elements in their genomes. Mutation of these proviruses, as well as the flanking cellular DNA, could have resulted in the acquisition and loss of different restriction enzyme recognition sites in different possums, and therefore the variation in junction fragment sizes. Support for this scenario is provided by the observation that approximately equal numbers of TvERV-K fragments were detected in the genomes of all three possums tested here, despite the variations in sizes of the fragments. However, a mutation rate far exceeding that normally seen for eukaryotes would be required to generate the junction fragment variability seen in the three possums tested here.

An alternative mechanism for the generation of variable junction fragment patterns might be homologous recombination between LTRs of TvERV-K proviruses. Again, all of the TvERV-K proviruses currently observed might have been present in the possum genome when possums first evolved. Homologous recombination between the LTRs of different proviruses (see 1.3.2 Amplification within the germline) in different possums, and therefore deletion of the junction fragments derived from those proviruses, could then generate the variability that is seen today. However, the TvERV-K variability in possums would require loss by homologous recombination at a rate far greater than the rate of 4–4.5 x 10^{-6} reversions per meiosis observed for ecotropic (Seperack *et al.*, 1988) and non-ecotropic (Frankel *et al.*, 1990) proviral loci in mice.

Finally, unique junction fragments may have been generated by integration of new TvERV-K proviruses at unique sites in the possum germline. Thus, a small set of TvERV-K elements might have been present in the genomes of the first possums. Subsequent retrotranspositional activity of one or more of the original TvERV-K elements within the genomes of possums could have lead to the acquisition of new TvERV-K proviruses at different positions in the genomes of different possums. TvERV-K retrotransposition may be ongoing. Previous studies by Tchenio and Heidmann (1991; 1992) have demonstrated that cloned retroviruses that are defective in *gag*, *pol*, and *env* genes can retrotranspose within host cells if they are complemented for Gag and Pol function. Complementation for Env is not required for retrotransposition, and only non-infectious particles are released from the cells expressing the Gag and Gag-Pol proteins, indicating that the retrotransposition occurs intracellularly (Tchenio and Heidmann, 1991; 1992). Thus, the TvERV-K elements, at least one member of which possesses complete *gag*, *pro*, and *pol* ORFs, may also be able to retrotranspose intracellularly. Alternatively, as mentioned above, one or more TvERV-K elements might encode a functional Env protein, or such a protein could be provided by other endogenous or exogenous retroviruses, enabling amplification of TvERV-K within the possum germline by extracellular reinfection of germline cells.

Chapter 7

General Discussion and Conclusions

7.1 Possums have endogenous retroviruses

Degenerate primers derived from the retroviral *pol* gene were used to PCR amplify possum genomic DNA as described in Chapter 3. Cloning and sequencing of these PCR products, and analysis of their sequences clearly demonstrated that the genome of the common brush-tailed possum (*Trichosurus vulpecula*) contains several families of endogenous retroviruses. The clones could be divided into three groups based on database search results (see FIG. 3.5) and phylogenetic analyses (see FIG. 3.9). All three groups were related, to varying extents, to previously reported marsupial endogenous retrovirus sequences (see FIG. 3.5 and FIG. 3.9).

Although the short clones from possum genomic DNA (TvPol1–TvPol12) were demonstrably related to retroviruses from other mammals (including other marsupials), the amount of sequence data obtained (85–91 nucleotides encoding 28–30 amino acids) prevents any firm conclusions from being drawn as to the origins of the possum endogenous retroviruses and their time of entry into the genome of the possum (or its evolutionary ancestors). It is possible that some of the elements are ancient, and are also present in the genomes of all marsupials, all mammals, or perhaps all vertebrates. It is also possible that some possum endogenous retroviruses arose, or were acquired, more recently, and are found in only a subset of marsupials, mammals, or vertebrates. Isolation of longer regions of sequence of the possum endogenous retroviruses from which the short clones were derived, and comparisons with related endogenous viruses in other marsupials, other mammals, and other vertebrates may clarify some of these issues.

7.2 TvERV-K is a biologically active endogenous retrovirus in possums

Reverse transcriptase activity was detected in the filtered blood plasma of all possums tested (see FIG. 3.6). In addition, products of an appropriate size (~130bp) were generated by RT-PCR amplification, using the *pol*-derived degenerate primers, of RNA isolated from the filtered plasma of all possums tested (see FIG. 3.7). Cloning and sequencing of these products from one

possum generated five clones (TvPol13–TvPol17) that were essentially identical in nucleotide sequence (see FIG. 3.8). These clones were also related to known retroviruses, but were distinct from the short clones generated from possum genomic DNA (see FIG. 3.8 and FIG. 3.9).

A primer (TvP11F) derived from the sequences of clones TvPol13–TvPol17 was used to amplify, using a 3'-RACE approach, the 3' end of the RNA from which the TvPol13–TvPol17 sequences were generated (see 4.2.1 3'-RACE to amplify the 3' end of the retroviral genome). A ~4.5kb product was cloned and sequenced, and was subsequently found to be derived from an endogenous retrovirus in the possum genome (see 4.2.2 Clone pTvRV3' was derived from an endogenous possum retrovirus). All subsequent work focussed on obtaining the full-length sequence of this endogenous retrovirus, which enabled more detailed analysis of its structure and biological activity (see below).

The assembly of a contiguous sequence (the TvERV-K1 contig) representing an entire possum endogenous retrovirus (TvERV-K) and the amplification and cloning of a single, near-full-length copy of this contiguous sequence (pTvERV-K2) are described in Chapter 4. The sequences of the TvERV-K1 contig and pTvERV-K2 are analysed in Chapter 5. The TvERV-K elements are the first full-length marsupial retroviruses to be sequenced.

The TvERV-K elements appear to be capable of significant biological activity. The TvERV-K1 contig and pTvERV-K2 appear to possess all of the regulatory regions required for replication of a retroviral genome (see 5.2.3 Regulatory regions). They also appear to encode functional Gag, Pro, and Pol polyproteins (see 5.2.4 Gene products). Thus, TvERV-Ks may express Gag proteins that form particles and package RNA. The successful PCR-amplification of apparently particle-associated TvERV-K RNA in possum blood plasma supports this possibility (see 3.2.3 Detection of retroviral RNA in possum blood). These putative particles could also contain TvERV-K Gag-Pro and Gag-Pro-Pol proteins, which, following particle maturation, would yield active DU, PR, RT, and IN enzymes. Again, support for this possibility is provided by the observation that particle-associated RT activity

was present in the blood plasma of all possums tested (see 3.2.2 Detection of reverse transcriptase activity in possum blood). Note, however, that the correct functioning of TvERV-K regulatory regions and gene products remain to be verified by experimental investigation. Whether the TvERV-K ORFs do in fact encode functional proteins will require expression of these ORFs in heterologous or *in vitro* systems, and assays for the relevant biological activities. Antibodies raised against recombinant proteins should also enable the question of whether they are expressed *in vivo* to be answered.

7.3 Selective maintenance of some TvERV-K genes

Both the TvERV-K1 contig and pTvERV-K2 possessed uninterrupted (or minimally interrupted) *gag*, *pro*, and *pol* genes, but highly mutated *env* genes (see 5.2.4 Gene products). There are two possible explanations for this observation. First, the TvERV-K elements from which the TvERV-K1 contig and pTvERV-K2 were derived may have entered the possum genome relatively recently, and may have possessed defective *env* genes at the time of entry. The retrovirus particle that infected the germline of the possum (or an evolutionary ancestor) might have comprised the protein products of TvERV-K *gag*, *pro*, and *pol* genes. Functional Env proteins might have been provided by an as yet unidentified TvERV-K element possessing a functional *env* gene, or by another endogenous or exogenous retrovirus. A brief presence within the possum genome would explain why the *gag*, *pro*, and *pol* genes are relatively (or completely) unscarred by mutation. Second, TvERV-Ks may have resided in the genomes of possums (and their ancestors) for a significant period of time. In this case, TvERV-Ks might have possessed complete *gag*, *pro*, *pol*, and *env* genes at the time of entry, but subsequently suffered numerous mutations almost exclusively in the *env* gene. Perhaps there has been selection for possums with Gag-, Pro-, and/or Pol-encoding TvERV-K elements in their genomes, because these proteins provided some advantage for their hosts. It is difficult to imagine what those advantages might be. Alternatively, it is possible that the original TvERV-K *env* gene product exerted a deleterious effect on its host, and that TvERV-K elements possessing functional *env* ORFs have been selected against.

Maintenance of extensive ORFs despite a long presence in the germline of a host is not uncommon for endogenous retroviruses. The HERV-K group of human endogenous retroviruses is one of the most studied groups of endogenous retroviruses (for reviews see Löwer *et al.*, 1996; Tönjes *et al.*, 1996). HERV-K elements entered the germline of an Old World primate ancestor sometime after the divergence of the Old World and New World primates approximately 30 MYBP (Steinhuber *et al.*, 1995). They are consequently found in the genomes of all Old World primates, great apes, and hominids, but not in the genomes of New World monkeys (Steinhuber *et al.*, 1995; Mayer *et al.*, 1998). Despite this long presence in the primate/human genome, HERV-K elements still display extensive biological activity. HERV-K transcripts have been observed in numerous human tissues (see Löwer *et al.*, 1996; Tönjes *et al.*, 1996). Cloning and heterologous expression of various regions of the HERV-K genome have demonstrated that at least some HERV-K elements encode functional Gag (Mueller-Lantzsch *et al.*, 1993; Sauter *et al.*, 1995), dUTPase (Harris *et al.*, 1997), protease (Mueller-Lantzsch *et al.*, 1993; Schommer *et al.*, 1996), reverse transcriptase (Berkhout *et al.*, 1999), integrase (Kitamura *et al.*, 1996), and Env (Tönjes *et al.*, 1997) proteins. There is strong evidence that HERV-K elements express retroviral particles (Boller *et al.*, 1993; Löwer *et al.*, 1993) which possess reverse transcriptase activity (Patience *et al.*, 1996) and package HERV-K RNA genomes (Seifarth *et al.*, 1995; Simpson *et al.*, 1996). To date, however, no infectious particles of human endogenous retrovirus origin have been detected (see Wilkinson *et al.*, 1994; Löwer *et al.*, 1996; Patience *et al.*, 1997), although HERV-K elements have undergone recent retrotranspositional expansion within the human genome (Medstrand and Mager, 1998).

7.4 Some members of the TvERV-K family may encode Env proteins

Both the TvERV-K1 contig and pTvERV-K2 were defective in the *env* gene (see 5.2.4.4 Env protein). Thus, particles expressed from the TvERV-K1 contig or pTvERV-K2 would not be infectious. However, both PCR and Southern blotting and hybridisation analyses suggested that shorter TvERV-K-related elements are present in the possum genome (see 4.2.8

Amplification and cloning of a near-full-length endogenous virus, and 6.2.1 TvERV-K internal fragments), and it is possible that one or more of these encodes a functional Env protein. Alternatively, it is conceivable that one of the other endogenous retroviruses in the possum genome, or possibly an exogenous retrovirus, could complement the TvERV-K elements for Env function, and enable the production of infectious pseudotypes.

7.5 TvERV-K may express, or be closely related to, exogenous retroviruses in possums

Although TvERV-K exists as an endogenous retrovirus in possums, it may also be capable of exogenous replication in possums, or it may be closely related to an exogenous possum retrovirus. A number of endogenous retroviruses that can replicate within the cells of their hosts have been identified. The best studied are the avian type C retroviruses (ASLVs), the mammalian type C retroviruses (MLVs) in mice, and the type B retrovirus (MMTV) in mice (Coffin, 1982; Boeke and Stoye, 1997). Some endogenous retroviruses are closely related to, but distinct from, exogenous retroviruses infecting their host species. The type B/type D endogenous sheep retrovirus (ESRV) provides an example of such a retrovirus. Approximately 15–20 copies of ESRVs are present in the genomic DNAs of various breeds of sheep and goat (York *et al.*, 1992; Hecht *et al.*, 1996). The ESRVs are closely related to exogenous retroviruses, JSRV and ENTV, which infect and cause diseases (ovine pulmonary carcinomas and enzootic nasal tumours, respectively) in sheep (Cousens *et al.*, 1999; Palmarini *et al.*, 1999). However, the endogenous ESRVs and the exogenous JSRV and ENTV can be distinguished genetically (Bai *et al.*, 1996; Palmarini *et al.*, 1996; Bai *et al.*, 1999). Whether the TvERV-K elements are capable of replicating in possum cells or have closely related exogenous cousins in possums remain to be determined.

7.6 TvERV-K may be retrotranspositionally active

TvERV-K elements may have been retrotranspositionally active within the possum genome subsequent to speciation. Southern hybridisation analysis

of TvERV-K junction fragments using a probe from the *gag* gene revealed unique junction fragments in the genomes of each of three possums (see 6.2.2 TvERV-K junction fragments). Although mutation of restriction enzyme recognition sites and loss of TvERV-K proviruses by homologous recombination between their LTRs are both potentially responsible for the observed variability, they would both have been required at frequencies orders of magnitude higher than has previously been observed. The alternative explanation, as described in the discussion in Chapter 6, is that TvERV-K has undergone recent retrotransposition within the genomes of possums. This retrotransposition might have occurred intracellularly using proteins expressed exclusively from TvERV-K elements. Alternatively, retrotransposition might have occurred by expression of extracellular particles and reinfection of germline cells. The requisite Env proteins might have been provided by as yet unidentified TvERV-K elements (see 7.4 Some members of the TvERV-K family may encode Env proteins), or by other endogenous or exogenous possum (or possum-ancestor) retroviruses. That some members of the TvERV-K family still possess intact *gag*, *pro*, and *pol* genes suggests that amplification of TvERV-K within the possum genome may be ongoing.

7.7 Possums may have inherited TvERV-K from their evolutionary ancestors

Several lines of evidence suggest that possums inherited TvERV-K elements from their evolutionary ancestors, and that TvERV-K should therefore be thought of as an ancient endogenous retrovirus. First, TvERV-K elements were present in the genomes of all possums tested. Although this could have resulted from endemic infection and germline entry in possums, it could also be a result of common inheritance. Second, although the patterns of junction fragments in the genomic DNAs of three possums revealed many bands that were unique to each possum, several of the bands were present in all of the possums (see 6.2.2 TvERV-K junction fragments). This could indicate that at least some of the TvERV-K integrants were present in the possum genome prior to the establishment of possums as a species, although it might also be the result of unique TvERV-K integration events that

generate junction fragments of the same size. Third, both PCR and Southern blotting and hybridisation investigations suggested elements identical in length to the TvERV-K1 contig and pTvERV-K2 are present in the genomes of all possums (see FIG. 4.9 and FIG. 6.3). Both the TvERV-K1 contig and pTvERV-K2 were defective in the *env* gene (see 5.2.4.4 Env protein). The elements of similar length in various possums may have undergone independent rearrangements, but it seems most likely that they all originated from a single element that contained the rearrangements. Although elements defective in the *env* gene can be complemented by retroviruses expressing functional Env, transfer from the germline of one possum to the germline of another would probably require several rounds of replication, each of which would require Env complementation. This scenario seems unlikely. Further analysis of the TvERV-K elements and their distribution in the possum genome, as well as the occurrence of TvERV-K-related elements in the genomes of closely and distantly related marsupials should shed more light on the time of entry into the possum germline.

7.8 Distribution of TvERV-K in possums and other marsupials

It should be emphasised that all possums used in this study were caught in New Zealand. The presence of TvERV-K in the genomic DNA of brush-tailed possums of Australian origin has not been investigated. Given that TvERV-K was present in the DNA of all New Zealand possums tested, it seems likely that TvERV-K is also present in all Australian brush-tailed possums. Nevertheless, it is possible that TvERV-K is present in only a subset of Australian possums, and that its fixation in the germline of New Zealand possums was a chance event. The small number of possums that were introduced into New Zealand from Australia increases the likelihood of such a founder effect. Alternatively, TvERV-K might have entered the genomes of New Zealand possums subsequent to their arrival in New Zealand. If this is the case, it is difficult to imagine what the immediate source of TvERV-K could be, given that TvERV-K belongs to a genus of retroviruses that, at this stage, appears to be of non-human primate origin (see 7.9 Origin of the TvERV-K elements). Analysis of the genomes of Australian possums will be required to resolve this issue.

The distribution of TvERV-K integrants appears to be highly variable between possums (see 7.6 TvERV-K may be retrotranspositionally active). It is possible that this variation in New Zealand possums reflects variation in the genomes of the possums that were originally introduced from Australia. Another possibility is that the retrotransposition events that gave rise to the observed variation have occurred in possums subsequent to their arrival in New Zealand. Again, resolution of this issue will require analysis of the genomic DNA of Australian possums and a greater number of New Zealand possums.

If TvERV-K is present in the genomes of Australian possums, analysis of the genomes of other marsupials might help to pinpoint the time of entry into the genomes of the evolutionary ancestors of possums. In this regard, the genomes of marsupials both closely and distantly related to possums should be investigated. It should be emphasised that the mere presence of TvERV-K in the genomes of two marsupials is not definitive proof that they inherited it from a common ancestor — they may both have obtained it by cross-species transmission followed by germline infection. Rather, the presence of TvERV-K integrants at identical positions within the genomes of two marsupials would be more reliable evidence of common inheritance.

7.9 Origin of the TvERV-K elements

The sequences of the TvERV-K Gag, Pro, and Pol polyproteins, as well as a short sequence corresponding to the TM protein, all have highest homology to those of the simian type D retroviruses (see 5.2.4 Gene products). Phylogenetic analysis using the Pro and Pol sequences also placed the TvERV-K elements exclusively with the simian type D retroviruses (see 5.2.5 Phylogenetic analysis).

The simian type D retroviruses have both exogenous and endogenous members. The three exogenous members from macaques (Old World monkeys) that have been cloned and sequenced are Mason-Pfizer monkey virus (MPMV) (Sonigo *et al.*, 1986), simian retrovirus type 1 (SRV-1) (Power *et al.*, 1986), and SRV-2 (Thayer *et al.*, 1987). These viruses are pathogenic

in their hosts, and induce fatal immunodeficiency syndromes (for example see Marx *et al.*, 1984). Endogenous type D retroviruses, known as simian endogenous retroviruses (SERVs), are widespread in Old World monkeys, and have recently been extensively characterised by van der Kuyl *et al.* (1997). They are discussed further below. No exogenous type D retroviruses have been reported in New World monkeys. However, an endogenous retrovirus, squirrel monkey retrovirus (SMRV), is present in the genome of the squirrel monkey (*Saimiri sciureus*). SMRV was first isolated from squirrel monkey lung tissue by cocultivation with canine cells (Heberling *et al.*, 1977). A full-length SMRV sequence, designated SMRV-H, was cloned from a human cell-line apparently contaminated by SMRV (Oda *et al.*, 1988). Regions of this clone were almost identical to previously reported SMRV sequences (Chiu *et al.*, 1984; Chiu and Skuntz, 1986). SMRV has not been detected in the genomes of any New World monkeys apart from the squirrel monkey, or in the genomic DNAs of Old World monkeys, apes, or non-primates (Fine and Schochetman, 1978, Teich, 1982).

Recent analysis of endogenous type D retroviruses (SERVs) in Old World monkeys has shed further light on the evolution of the simian type D retroviruses (van der Kuyl *et al.*, 1997). SERVs are present in the genomes of all species of Old World monkeys tested, but not in the genomes of humans and apes, indicating that they integrated into the genomes of Old World monkeys sometime after the divergence of the monkey and hominoid branches approximately 36 MYBP (van der Kuyl *et al.*, 1997). The phylogenies of one subgroup of SERVs (the SERV 23.1-related subgroup) reflects the phylogenies of the host families, suggesting integration into the Old World monkey genome before the divergence of the separate families, and subsequent evolution with the host genomes (van der Kuyl *et al.*, 1997). The SERV Gag, Pro, Pol, and P20 (TM) proteins are very similar in sequence to those of the exogenous simian type D retroviruses (MPMV, SRV-1, and SRV-2). However, the GP70 (SU) proteins of the SERVs and the exogenous type D retroviruses share no obvious homology, leading to the proposal that the exogenous type D retroviruses arose by acquisition of an SU-encoding sequence, of unknown origin, by a SERV (van der Kuyl *et al.*, 1997). Similarly, baboon endogenous retrovirus (BaEV), an endogenous

retrovirus in some Old World monkeys, is suggested to be the product of recombination between the *env* gene of a SERV (van der Kuyl *et al.*, 1997) and the genome of a primate type C retrovirus (PcEV) (Mang *et al.*, 1999). SMRV, the endogenous type D retrovirus from the New World squirrel monkey, appears to be a separate lineage of simian type D retroviruses. van der Kuyl *et al.* (1997) suggest that the type D retroviruses have been associated with primates since the type D retroviruses first evolved, and that they have entered the germlines of different primate hosts at different times: SMRV entered the germline of the squirrel monkey following its separation from the other New World monkeys (Fine and Schochetman, 1978); and the SERVs became endogenous in the Old World monkeys, after they and the hominoid primates diverged (van der Kuyl *et al.*, 1997).

If the type D retroviruses did evolve within a primate context then cross-species transmission from a primate to possums (or an evolutionary ancestor of possums) must be assumed. However, possums inhabit a different continent (Australia) than do the Old World monkeys (Africa and Asia) and the squirrel monkey (central and South America) (Nowak, 1991). Thus it seems unlikely that a retrovirus was transferred directly between possums and either the Old World monkeys or the squirrel monkey. Interestingly, in their study of the distribution of mammalian type C retroviruses, Martin *et al.*, (1999) found an endogenous retrovirus in koalas (an Australian marsupial) that was closely related to the exogenous GaLV in gibbon apes (from southeast Asia). The authors invoked an intermediate, possibly a rodent, to explain the intercontinental transmission (Martin *et al.*, 1999). Similarly, a vector would need to be involved in the transfer of retroviruses from primates in Africa, Asia, or America to possums in Australia. The recent report of an endogenous retrovirus, with homology to the simian type D retroviruses, in several strains of mice suggests that rodents might also be responsible for the transfer of the simian type D retroviruses (Ristevski *et al.*, 1999). An alternative explanation might be that the simian type D retroviruses are not actually of primate origin, and that they were transferred to both primates and possums (or their evolutionary ancestors) from a common, as yet unidentified, source. Further characterisation of the mouse type D virus — at this stage only a short region of sequence, corresponding to parts of the *gag*

and *pro* genes, is known — may shed light on the phylogeny of the simian type D retroviruses.

7.10 Implications for possums

The presence of TvERV-K, and endogenous retroviruses in general, has a number of potential implications for possums (see 1.3.6 Implications of presence of retroviruses in the genome/germline).

It is not known to what extent retroviruses have been involved in shaping the genomes of possums (and other marsupials). TvERV-K elements, which are only present in 15–30 full-length copies in the possum genome, might not be expected to have influenced the possum genome significantly. However, full-length endogenous retroviruses are often accompanied by a vast excess of solitary LTRs (see 1.3.2 Amplification within the germline). High numbers of solitary LTRs provide more potential foci for chromosomal rearrangements (Boeke and Stoye, 1997). The number of solitary TvERV-K LTRs in the possum genome has not been investigated.

Expansion of chromosomal centromeres and genome-wide undermethylation was observed in a hybrid produced from the mating of two kangaroo species, *Wallabia bicolor* (swamp wallaby) and *Macropus eugenii* (tammar wallaby) (O'Neill *et al.*, 1998). Amplification of a retroviral element, termed kangaroo endogenous retroviral element-1 (KERV-1), was associated with the centromeric expansion. However, elements of the same structure as KERV-1 could not be detected in the genomes of either of the parent species of the hybrid. The authors suggested that KERV-1 arose by xenologous recombination between two partial viruses present in the genomes of one or both parents, and subsequently invaded the centromeres of the swamp wallaby-derived chromosomes. Database searches using a two kilobase fragment from KERV-1 suggested it was related to the type B and type D retroviruses, the human endogenous HERV-K elements, and the murine IAPs (O'Neill *et al.*, 1998). The sequence of KERV-1 contained numerous frameshift mutations and rearrangements, and detailed comparison with the TvERV-K sequences was not possible. That the possum genome contains

endogenous retroviruses, including the TvERV-K elements, which are related to the type B and type D retroviruses, HERV-K, and the IAPs, suggests that similar rearrangements have the potential to occur in possums.

The presence of the TvERV-K elements in the possum genome may have consequences for the infection of possums by other retroviruses. It is conceivable that some members of the TvERV-K group encode Env proteins. Expression of these Env proteins in possum tissues could block receptor binding by retroviruses with related Env proteins in a manner similar to that seen with the *Fv4* locus in mice (see 1.3.6 Implications of presence of retroviruses in the genome/germline). Similarly, infection by related retroviruses could be blocked at a post-entry stage analogous to the mechanism of restriction by the *Fv1* locus in mice (see 1.3.6 Implications of presence of retroviruses in the genome/germline). In this respect, a marsupial kidney epithelial cell line derived from a rat kangaroo has been shown to be resistant to infection by MoMLV pseudotypes carrying Env proteins from some members of the SNV/RD114/type D retrovirus receptor interference group (Koo *et al.*, 1994). Specifically, viral pseudotypes carrying SNV and RD114 Env proteins were unable to infect the marsupial cells, which could be infected by pseudotypes carrying BaEV Env proteins. The authors suggested that the marsupial cells might express a slightly modified receptor protein that allows entry of BaEV but not SNV or RD114 (Koo *et al.*, 1994). An alternative explanation might be that the presence, within the kangaroo genome, of endogenous retroviruses expressing Env proteins of the SNV/RD114/type D retrovirus group resulted in selection for an altered receptor, which still performs its function in the cell but which has altered Env-binding characteristics. As stated above (see 7.8 Distribution of TvERV-K in possum and other marsupials) the distribution of TvERV-K elements, which appear to have once encoded (or may still encode) an Env protein belonging to the SNV/RD114/type D retrovirus interference group, has not been investigated in other marsupials. The receptor for the SNV/RD114/type D retrovirus group of viruses has recently been cloned from the human genome, and is a neutral amino acid transporter known as RDR (for RD114/type D retrovirus receptor) (Rasko *et al.*, 1999). The amino acid transport activity of RDR was reduced by exposure of cells to replicating

RD114, MPMV, and BaEV viruses, and was consequently suggested to be a potential mechanism of pathogenesis (Rasko *et al.*, 1999). It would be interesting to compare the properties of the possum/marsupial RDR (assuming it exists) with those of the human RDR.

The TvERV-K elements are potentially available for recombination with other endogenous and exogenous retroviruses. It appears likely that TvERV-K elements are transcribed and packaged into retrovirus particles. Co-packaging of a TvERV-K genome with that of another retrovirus might lead to recombination and the generation of a new virus with altered biological properties. As described above (see 7.9 Origin of the TvERV-K elements), the SERVs, also members of the simian type D retrovirus genus, appear to have contributed to the generation of at least two retroviruses — the exogenous type D retroviruses of macaques (MPMV, SRV-1, SRV-2) and the baboon endogenous retrovirus (BaEV) — by recombination. The exogenous type D retroviruses are clearly pathogenic to their hosts (see above). Thus, recombination between TvERV-Ks and other retroviruses could potentially give rise to retroviruses that are pathogenic in possums.

7.11 Implications for other animals (including humans)

TvERV-K elements with intact *env* genes may be present in the possum genome, and may be able to express infectious retrovirus particles. Although replication of these retroviruses may be blocked in possum cells, it may be supported in cells of other animals (ie. TvERV-K may be a xenotropic possum retrovirus). SMRV is a xenotropic retrovirus which is unable to replicate in cells of its natural host (the squirrel monkey), but which can infect cultured cells from a variety of mammals, including Old World monkeys, apes (including humans), minks, dogs, cats, and rabbits (for review see Fine and Schochetman, 1978). Indeed, SMRV was first isolated by coculturing squirrel monkey lung cells with a fetal canine thymus culture (Heberling *et al.*, 1977). The first clones of SMRV to be sequenced were generated from such a cocultures (Chiu *et al.*, 1983), and the only full-length SMRV sequence to be reported was generated from an SMRV replicating in a human lymphoblastoid cell line (Oda *et al.*, 1988). MPMV, an exogenous type D

retrovirus from macaques, can also infect human cells in culture (Fine *et al.*, 1979). That the closest known relatives of TvERV-K are able to replicate in cells of several other animals, including humans, suggests that the potential for cross-species transmission of TvERV-K from possums to other animals, including humans, exists. Inoculation of cultures of cells from other animals with TvERV-K particles, or coculturing of these cells with possum cells, might reveal whether TvERV-K particles can infect other animals.

7.12 Implications for the detection of exogenous retroviruses in possums

Assuming that the reverse transcriptase activity and TvERV-K RNA detected in possum blood were derived from endogenous TvERV-Ks, these findings have obvious implications for future searches for exogenous retroviruses in possums. In the absence of additional information, future searches for retroviruses are likely to employ electron microscopy, reverse transcriptase assay, and PCR amplification approaches. TvERV-Ks probably express virus particles, which, although they have not yet been observed, would be expected to have similar morphology to those of the simian type D retroviruses. Such particles may complicate searches for virions of other retroviruses by electron microscopy. Similarly, the reverse transcriptase activity that has been detected in the blood of all possums tested may eliminate RT assays as a tool for searching for exogenous retroviruses in possums. Finally, RT-PCR using degenerate *pol*-derived primers will also detect TvERV-K RNAs in blood plasma. Presumably, cloning of the PCR products and sequencing of sufficient numbers of clones will allow the detection of other retroviruses, if they are present, although this may depend on the nature and relative abundance of the retroviruses.

7.13 Potential use of TvERV-K as a possum-specific biocontrol agent

At this stage, no TvERV-K element encoding a functional Env protein has been identified. Even if such an element exists, it seems unlikely that it could be used as a possum biocontrol agent.

Replication of TvERV-K in possum cells might be blocked. If Env proteins are expressed from endogenous TvERV-K elements, these Env proteins may block extracellular infection by receptor interference (see 1.3.6 Implications of the presence of retroviruses in the genome/germline). Alternatively, expression of TvERV-K Gag proteins, or Gag proteins from one of the other endogenous retroviruses in the possum genome, might block infection at a post-entry stage, analogous to the *Fv1* locus in mice (see 1.3.6 Implications of the presence of retroviruses in the genome/germline). The Gag protein encoded by the *Fv1* gene has highest homology to those of the ERV-L retroviruses, and yet it blocks infection by MLVs (Best *et al.*, 1996). Interestingly, the possum genome also harbours ERV-L-related endogenous retroviruses (see 3.2.4 Phylogenetic analysis of possum retroviral sequences), which may also be capable of blocking infection of possum cells by exogenous retroviruses.

Even if some TvERV-K elements do express particles that are able to infect possum cells, such viruses may not be pathogenic. That they are tolerated within the possum genome suggests that this is the case.

The TvERV-K elements are most closely related to SMRV and the exogenous and endogenous type D retroviruses of macaques. That some of these viruses are able to infect cells other than those of their hosts suggests that TvERV-K elements, if they do possess intact *env* genes (see 7.4 Some members of the TvERV-K family may encode Env proteins), might also be capable of cross-species transmission. Clearly, this would not be a favourable characteristic of a vector intended to infect only possums.

Although the endogenous nature of TvERV-K might prevent its use as a biocontrol vector, it might also, under some circumstances, be beneficial in terms of biological control of possums. It is not known if a TvERV-K-related exogenous retrovirus exists in possums. However, expression of TvERV-K proteins during the establishment of immune tolerance in possums might reduce their immune response to such a virus, if it does exist. In sheep, infection by the JSRV and ENTV retroviruses does not induce an immune response, and it has been hypothesised that this is due to expression of

endogenous retroviruses closely related to JSRV and ENTV during ontogeny (York *et al.*, 1992; Cousens *et al.*, 1999).

7.14 Future work

There are a number of questions regarding the TvERV-K elements that remain unanswered. In what tissues are TvERV-K elements expressed? Is TvERV-K expression spatially or temporally regulated? Are the TvERV-K proteins functional? Are TvERV-K elements responsible for the reverse transcriptase activity in possum blood? Does TvERV-K express particles? Can these particles replicate in possum cells? Can they infect other animals? Does a TvERV-K-related exogenous retrovirus exist? Are TvERV-Ks present in all New Zealand and Australian possums? Are TvERV-Ks present in the genomes of other marsupials? When did TvERV-Ks enter the germlines of possums (or marsupials)? Finally, where did the TvERV-K originate? Clearly, much more experimental work would be required to address these questions.

More thorough analysis of the TvERV-K elements, as well as the other endogenous retroviruses in possums, may shed light on the evolution of these retroviruses and their marsupial hosts.

Appendices

Appendix 1 – Retrovirus names, accession numbers, and source references

Abbreviated name	Full name	Accession number	Reference
BaEV	Baboon endogenous virus	D10032	Kato *et al.* (1987)
BIV	Bovine immunodeficiency virus	M32690	Garvey *et al.* (1990)
BLV	Bovine leukemia virus	AF033818	Petropoulos (1997)
CAEV	Caprine arthritis-encaphalitis virus	M33677	Saltarelli *et al.* (1990)
CHIAP	Chinese hamster IAP	M73970	Dorner *et al.* (1991)
EIAV	Equine infectious anemia virus	M16575	Stephens *et al.* (1986)
ENTV	Enzootic nasal tumor virus	Y16627	Cousens *et al.* (1999)
ESRV	Endogenous sheep retrovirus	-	Bai *et al.* (1996)
FeLV	Feline leukemia virus	M18247	Donahue *et al.* (1988)
FeSFV	Feline syncytium-forming virus	U78765	-
FIV	Feline immunodeficiency virus	M25381	Olmsted *et al.* (1989)
GaLV	Gibbon ape leukemia virus	M26927	Delassus *et al.* (1989)
HERV-E	Human endogenous retrovirus, glutamate tRNA PBS	M10976	Repaske *et al.* (1985)
HERV-L	Human endogenous retrovirus, leucine tRNA PBS	X89211	Cordonnier *et al.* (1995)
HERV-K10	Human endogenous retrovirus, lysine tRNA PBS	M14123	Ono *et al.* (1986)
HERV-K-T47D	Human endogenous retrovirus, lysine tRNA PBS, from T47D cell-line	AF020092	Seifarth *et al.* (1998)
HIV-1	Human immunodeficiency virus, type 1	AF033819	Petropoulos (1997)
HIV-2	Human immunodeficiency virus, type 2	M15390	Guyader *et al.* (1987)
HRV-5	Human retrovirus 5	U46939	Griffiths *et al.* (1997)
HSRV	Human spumaretrovirus	AF033816	Petropoulos (1997)
HTLV-1	Human T-cell leukemia virus, type 1	AF033817	Petropoulos (1997)
HTLV-2	Human T-cell leukemia virus, type 2	M10060	Shimotohno *et al.* (1985)
IAP	Intracisternal A-particle	-	-
JSRV	Jaagsiekte sheep retrovirus	M80216	York *et al.* (1992)
LDV	Lymphoproliferative disease virus	U09568	Sarid *et al.* (1994)
MIAP	Mouse IAP	M17551	Mietz *et al.* (1987)
MMTV	Mouse mammary tumor virus	M15122	Moore *et al.* (1987)
MoMLV	Moloney murine leukemia virus	AF033811	Petropoulos (1997)
MPMV	Mason-Pfizer monkey virus	AF033815	Petropoulos (1997)
MuERV-L	Murine endogenous retrovirus, leucine tRNA PBS	Y12713	Benit *et al.* (1997)
MuRRS	Murine retrovirus-related DNA sequences	X02487	Schmidt *et al.* (1985)
OMVV	Ovine maedi/visna virus	M31646	Querat *et al.* (1990)
RSV	Rous sarcoma virus	AF033808	Petropoulos (1997)
SERV 23.1	Simian endogenous retrovirus, clone 23.1	U85505	van der Kuyl *et al.* (1997)

Abbreviated name	Full name	Accession number	Reference
SERV 25.2	Simian endogenous retrovirus, clone 23.1	U85506	van der Kuyl *et al.* (1997)
SFV	Simian foamy virus	X54482	Kupiec *et al.* (1991)
SHIAP	Syrian hamster IAP	M10134.1	Ono *et al.* (1985)
SIVagm	Simian immunodeficiency virus, African green monkey	X07805	Fukasawa *et al.* (1988)
SIVmac	Simian immunodeficiency virus, rhesus macaque	Y00277	Chakrabarti *et al.* (1987)
SIVmnd	Simian immunodeficiency virus, mandrill	M27470	Tsujimoto *et al.* (1989)
SMRV-H	Squirrel monkey retrovirus, from human lymphoblastoid cell-line	M23385	Oda *et al.* (1988)
SRV-1	Simian retrovirus, type 1	M11841	Power *et al.* (1986)
SRV-2	Simian retrovirus, type 2	M16605	Thayer et. (1987)
Visna	Visna virus	M10608	Sonigo *et al.* (1985)
WDSF	Walleye dermal sarcoma virus	AF033822	Petropoulos (1997)

Appendix 2 – Primer sequences and T_ms

Primer sequences and T_ms are listed below. All primers were ordered from Life Technologies, were resuspended and diluted in either TE (pH 8.0) or deionised water, and were stored at –20°C.

Primer name	Primer sequence	T_m[a]
Degenerate *pol* primers[b]:		
polF	5' ctc gaa ttc GTN YTN CCN CAR GG 3'	–
polR	5' ctc aag ctt RTC RTC CAT RTA 3'	–
PERT assay primers:		
RT-1	5' CAT AGG TCA AAG GTG GTA CCA ATG 3'	61.7°C
RT-2	5' TCC TGC TCA ACT TCC TGT CGA G 3'	64.2°C
3'-RACE primers[c, d]:		
oligoTextn	5' CTC AAG CTT ACG GAC AGT CAG TGG TTT TTT TTT TTT TTT MN 3'	–
extn	5' CTC AAG CTT ACG GAC AGT CAG TGG 3'	64.5°C
TvERV-K-specific primers:		
TvP01F	5' GGG GAC GTG GCA TTC TGA TAA CA 3'	67.3°C
TvP02F	5' CAG GAA CAA TAA AGA GAG ACA 3'	52.2°C
TvP03F	5' GTC TCC TTC CTC TCC GTG AT 3'	58.3°C
TvP04F	5' CGG ACT CCG AAA ACC CCA TTC 3'	65.2°C
TvP05F	5' CGG GGA AAT GGC AGA AAG TAG G 3'	65.9°C
TvP06R	5' CTC CTC CTC TTG GCT TTC CT 3'	60.9°C
TvP07R	5' TGC ATT CTC AAA AGC CAG TCT 3'	60.0°C
TvP08R	5' GCA CGG CCA ATA ATG AGG AAA 3'	64.2°C
TvP09R	5' ATG TAA AGG TTG AGT GAT GC 3'	52.6°C
TvP10R	5' CAG TGG CCC TTA TGA AAA ATG TT 3'	61.4°C
TvP11F[e]	5' ctc gaa ttC CAG AAG TAT GTT GC 3'	60.1°C
TvP12R	5' TGT GGA AAT TGT AAG CGA AAG G 3'	61.3°C
TvP13R	5' GCA GGG CCC AGC AAT AAG AAT G 3'	67.0°C
TvP14F	5' TTC TAC ATC ATT CTC CCC TAC GG 3'	61.5°C
TvP15F	5' CAA TGC CTT CTC TAT GTT CCT TCA A 3'	62.8°C
TvP16F	5' CCA AGC ACG TTA TCA ATC ACA T 3'	59.9°C
TvP17F	5' TTA TGC ATC CCT CTC AAA CTG T 3'	58.7°C
TvP18R	5' GGG ACA AAC CAC AGG CTC AA 3'	63.4°C
TvP19R	5' CTT TCT TGC CTT TCC TTT CTG GT 3'	62.2°C
TvP20R	5' ACG GAT TGG AGG GGG ATA CAG AA 3'	66.8°C
TvP21R	5' CAG GGG TTG TCA GGT TTG GAA GA 3'	66.6°C
TvP22F	5' CTG CAA CTA TGA GGA GAG A 3'	51.1°C
TvP23F	5' ACC GGA GAT AGT GTT AGA C 3'	48.8°C
TvP24R	5' AAA GCT CTT TAC GCC TCT GGA C 3'	54.9°C
TvP25F	5' AGC AAT TAT TAA TCT GAA TTC TGT T 3'	61.2°C
TvP26F	5' CAG CAG GAA GAG CGA CTA CAA T 3'	60.9°C
TvP27R	5' GCA ACT TGG GTC TGA TAA TGA G 3'	58.7°C
TvP28F	5' GAC ATG CCG GAG ACT GGA AAA TC 3'	66.2°C

[a] T_m determined by nearest neighbour calculation (Breslauer *et al.*, 1986) at www.williamstone.com; – indicates T_m was not calculated because primer contained degeneracies

[b] N = any nucleotide, R = purine, Y = pyrimidine
Lower case letters represent extensions (ie. not derived from RT sequence)
Restriction enzyme sites are shaded

[c] M = A, C, or G, N = any nucleotide
Restriction enzyme sites (*Hin*dIII) are shaded

[d] The MN sequence at the 3' end serves to "lock-dock" the oligoTextn primer at the junction of the gene-specific DNA and the poly(A) tail (Borson *et al.*, 1992)

[e] Lower case letters represent extension (ie. not derived from TvERV-K sequence)
Restriction enzyme site (*Eco*RI) is shaded

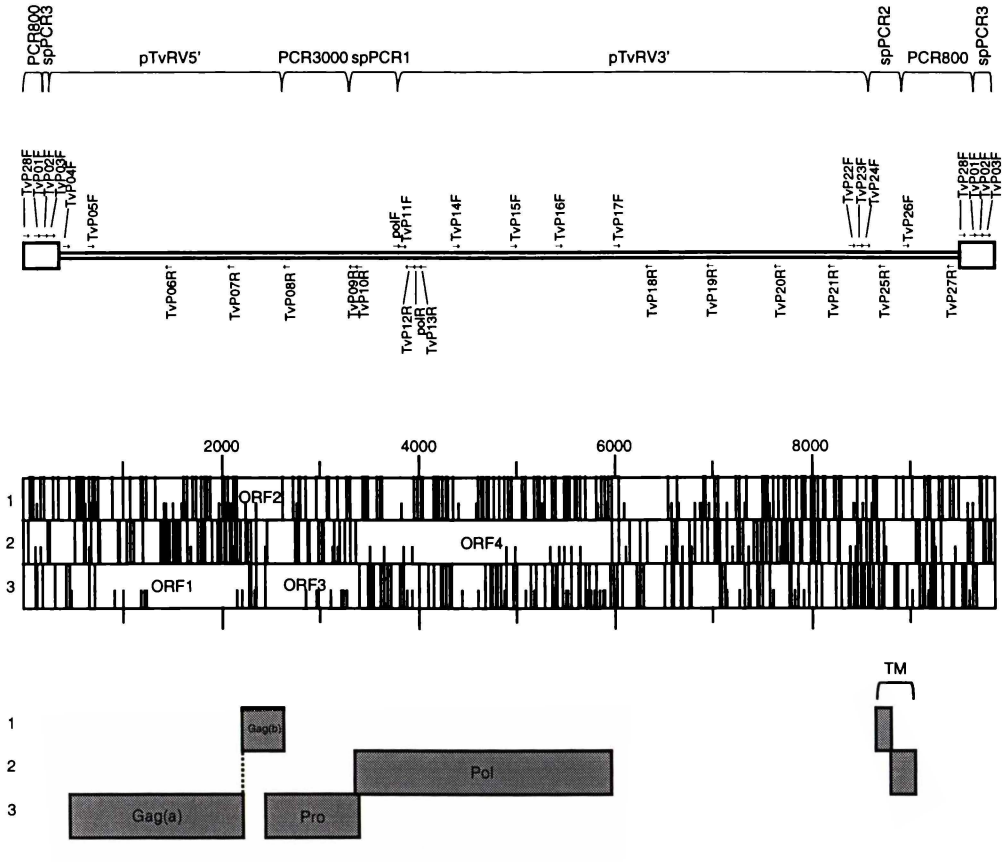
Appendix 3 – Short RT alignment

The amino acid sequences below were aligned using CLUSTAL X and the BLOSUM weight matrix as described in Chapter 2 – Materials and Methods. The alignment was used to construct a neighbour-joining tree, which is shown in FIG. 3.9. TvPol1–TvPol12 are the clones from possum genomic DNA (see FIG. 3.4). TvPol13 represents the clones from plasma RNA, which are identical over the sequence range used here (see FIG. 3.8). RV rock-wallaby was reported by Martin *et al.* (1997). RV common possum, RV sfdI, and RV sfdII are sequences from the common brushtail possum and the stripe faced dunnart reported by Herniou *et al.* (1998). RV koala and RV opossum were reported by Martin *et al.* (1999). Details of all other retroviruses are listed in Appendix 1.

TvPol7	YLHSPTICHRTVAQHLDELELPG--IQLTH
TvPol8	YLHSPTICHRTVAQHLDELELPG--IQLTH
TvPol9	YIHSPTICHRIVAEHLDELKLPG--VQLTH
TvPol10	YIHSPTICHRIVAEHLDELKLPG--VQLTH
RV common possum	DLHSPTICHRIVAEHLDKLKLAG--IQLTH
TvPol11	YLYNPTICHRIVAEYLDKLKSPG--VQLTH
MuERV-L	YINSPALCHNLVRRDLDRLDLPQS-ITLVH
HERV-L	YINSPALCHNLIQSDLDHFSLPQD-ITLVH
TvPol3	MKNSPAMCQQYVAQALQPIXEKYP-AYLIH
TvPol4	MKNSPTMCQQYVAQALQPIXEKYP-AYLIH
TvPol5	MKNSPTMCQQYVAQVLQPIREKYPKTYLIH
TvPol6	MKNSPTMCQQYAAQTLQAIREK-PKAYLTH
TvPol1	MANSPTVCQMYVAAIIQPVREQFPGAMIIH
LDV	MKNSPYICQQVVAEVIRPIRERFRDAVIIH
TvPol2	MKNSPTLCQWYVQQVIRPVHERYPTVLLIH
CHIAP	MANSPTICQLYVQKALEPVRKQFTSMIMIH
RSV	MTCSPTICQLVVGQVLEPLRLKHPSLCMLH
RVsfdI	MANSPILCQ--VDKVLAPVRNLYPNVYMLH
MPMV	MANSPTLCQKYVATAIHKVRHAWKQMYIIH
SMRV-H	MANSPTLCQKFVAAAIAPVRSQWPEAYILH
TvPol13	MTNSPTLCQKYVAQTIDPFRLQFPQLYIIH
JSRV	MTNSPTLCQKFVATAIAPVRQRFPQLYLVH
MMTV	MKNSPTLCQKFVDKAILTVRDKYQDSYIVH
HERV-K10	MLNSPTICQTFVGRALQPVREKFSDCYIIH
GaLV	FKNSPTLFDEALHRDLAPFRALNPQVVLLQ
RV koala	FKNSPTLFDEALHRDLASFRALNPQVVMLQ
MoMLV	FKNSPTLFDEALHRDLADFRIQHPDLILLQ
RV opossum	LKLSPTLFSDALARDLEEFRSSHPDVTLLQ
HTLV-1	FKNSPTLFEMQLAHILQPIRQAFPQCTILQ
TvPol12	FRDSPHLFGQALGKDLRDLKLTD--SSLIQ
RV sfdII	FRDSPHLFGQALAKDLRDLELKG--SCLLQ
ERV9	FRDSPHLFGQALAKDLGHFSSPG--TLVLQ
HERV-H	FRDNPHYFSQALSHDLLSFHSSA--SHLIQ
HSRV	FLNSPALFTADVVDLLKEIPNVQ--V----
FeSFV	FLNSPGLFTGDVVDLLQGIPNVE--V----
WDSV	FIHSPTLFSQALYQSLHKIKFKISSEICI-
HIV-1	WKGSPAIFQSSMTKILEPFRKQNPDIVIYQ
Visna	WKLSPAVYQFTMQKILRGWIEEHPMIQFGI
BLV	FINSPALFERALQEPLRQVSAAFSQSLLVS
RV rock-wallaby	YTESPNLFGQVLEGILEGFDPMAR-TRVLQ

Appendix 4 – The TvERV-K1 contig

The TvERV-K1 contig was assembled from the sequences of a number of different RT-PCR, PCR, and spPCR products, as described in Chapter 4 and indicated below. Also indicated below are the positions of the primers used to amplify and sequence the various regions of the TvERV-K1 contig, the arrangement of the TvERV-K1 contig ORFs, the proteins encoded by the TvERV-K1 contig. The table on p. 172 indicates the expected sizes (based on the TvERV-K1 contig sequence) of amplification products using various primer pairs. The sequence of the TvERV-K1 contig is listed on pp. 173–175, and relevant regulatory regions and protein products are shown.



Expected PCR product sizes based on the TvERV-K1 contig sequence

Primer name	TvP06R	TvP07R	TvP08R	TvP09R	TvP10R	TvP12R	TvP13R	TvP18R	TvP19R	TvP20R	TvP21R	TvP25R	TvP27R
TvP01F	1340	1931	2551	3182	3245	3760	3811	6185	6845	7481	8084	8560	9236
TvP02F	1298	1889	2509	3140	3203	3718	3769	6143	6803	7439	8042	8518	9194
TvP03F	1253	1844	2464	3095	3158	3673	3724	6098	6758	7394	7997	8473	9149
TvP04F	1059	1650	2270	2901	2964	3479	3530	5904	6564	7200	7803	8279	8955
TvP05F	851	1442	2062	2693	2756	3271	3322	5696	6356	6992	7595	8071	8747
TvP11F	-	-	-	-	-	49	100	2474	3134	3770	4373	4849	5525
TvP14F	-	-	-	-	-	-	-	1964	2624	3260	3863	4339	5015
TvP15F	-	-	-	-	-	-	-	1376	2036	2672	3275	3751	4427
TvP16F	-	-	-	-	-	-	-	909	1569	2205	2808	3284	3960
TvP17F	-	-	-	-	-	-	-	347	1007	1643	2246	2722	3398
TvP22F	-	-	-	-	-	-	-	-	-	-	-	341	1017
TvP23F	-	-	-	-	-	-	-	-	-	-	-	182	858
TvP24F	-	-	-	-	-	-	-	-	-	-	-	161	837
TvP26F	-	-	-	-	-	-	-	-	-	-	-	-	466
TvP28F	-	-	-	-	-	-	-	-	-	-	-	-	-

IR U3

1 TGCCGGAGACTGGAAAATCGCCTTACCCCTCCCCTCTGCCTGACCCGCGCCCCCTCCCTTCGCTAGCTTTCATTTCTCGCTTAGCTCTCATTTCCTGGCT 100

CCAAT box TATA box

101 AACTCTCGTTTCTGGGATATGACGCAGGCCTGCGTCCAATGCTTAGCCACCCCTCTGGGGACGTGGCATTCTGATAACATGCTCATAAAAACCCTTACCA 200

Poly(A) U3 ←. R .← U5

201 GGAACAATAAAGAGAGACATTTTTTACACCAGCAGCGCGTCTAGTCTCCTTCCTCTCCGTGATCCCGTTCCCGAGCCTCTTGGCCCTTTAGAGCCATCAG 300

U5 IR PBS

301 GTCCTCCTCGAGCCTAGCTCCCCCGGCCCTCTCCAGCTTCGCTGACCCACGATTGTGTGACCTGCGGGCCAGGTCACTTGGCGCCCAACGTGGGGCTCGA 400

SD

401 ACGCTGGACAGGGACCCGATAGACGCCCCGACTCGTCCGGACTCCGAAAACCCCATTAAGGTAAGTGTTCGCGTCTATCGACATCATGGGATCACAACTA 500

Gag> M G S Q L

501 TCCAAAGAACAGGTTTTTGTTAAAGATTTAAAACAAAGCTTCAGAGAGAGAGGAATAAGGGTAAAGAAAAAGGACCTTTTAAAATTTTTTATTTTTATAG 600

S K E Q V F V K D L K Q S F R E R G I R V K K K D L L K F F I F I

601 ATAAGGTATGTCCCTGGTTCATTGTTAATGGTCCAGATATACATCCGGGGAAATGGTAGAAAGTAGGGAGAGATCTTAATGAGAAATTGAAACAAGAAGG 700

D K V C P W F I V N G P D I H P G K W * K V G R D L N E K L K Q E G

701 TCCCGAGTAAGTTCCCATAAATGCCTTCTCCCTCTGGACCTTAATTAATGACATTATTGAGGGATCAACAGGAGACAAAGGTGCCCGCCAGTTACTCGTT 800

P E * V P I N A F S L W T L I N D I I E G S T G D K G A R Q L L V

801 CAGGCGGAATCCTGTCTCGCGCCACTTTCTCGCGCTGCTTCTGCATCCTCCCTTCACGCAGAAAACTCACAAGACAAAGAGGAAGTAAAATCTCCTCCCC 900

Q A E S C L A P L S R A A S A S S L H A E N S Q D K E E V K S P P

901 CTATCATCATGCCCCCTACCGTCAAGTCCATTTACCCGCCACCCCCTAACCCTGTTGAAGAAACACTCACTTTCCCGGTATTCCCTACCGAGAGGAAAGT 1000

P I I M P P T V K S I Y P P P P N P V E E T L T F P V F P T E R K V

1001 CGGCATGAATACTGCCCCCGAAAACGAAAGCGGCCCGCTCCCTGACAAGGACGCAGCCAATTTAGACGCTGAGGCTGCTGCATACAACCCTCAAGAAGTG 1100

G M N T A P E N E S G P L P D K D A A N L D A E A A A Y N P Q E V

1101 TTCGTATCCCATAATTTCTCCCCACCCCACTATTTACTACCCCCCCTTGTGGACCTTTCCCAAATCCAAAGAGATTTGACCCAGCGCATGGCTGAATTAA 1200

F V S H N F S P P H Y L L P P L V D L S Q I Q R D L T Q R M A E L

1201 GAAAGACAGTCAATATGCAGAATCAATATCACCAGATGGTAAAGGAGTTCTCGGCGCTACAAGCGTCCCTGCAGAAGGCACTCCTTCCTCCCTCTGCCCC 1300

R K T V N M Q N Q Y H Q M V K E F S A L Q A S L Q K A L L P P S A P

1301 AAAAGGTACTCTTATAGCATTCCCTCAAGAAAAAATTCCCCAACCAAAACAGAAACAGAAATCCCTTCTCTTTCCTATTCTTAGAAGCAATACACAGAAT 1400

K G T L I A F P Q E K I P Q P K Q K Q K S L L F P I L R S N T Q N

1401 AGACCTAAGGATGCAGAGGATTCAAACACTGATGAGAACCCTCCCGAAATTAGTGCAGAGAAGGAGGAGTCCGGTGAGGAAAGCCAAGAGGAGGAGGTGG 1500

R P K D A E D S N T D E N P P E I S A E K E E S G E E S Q E E E V

1501 AGAATGATGATCAGGAATCTGATTCAGAAGAGGACAGTAAATCATGCTCCTCTAAATACAAAGCCCCAAAGTTTAAGAATATCAAGGATCTCCATGCAGC 1600

E N D D Q E S D S E E D S K S C S S K Y K A P K F K N I K D L H A A

1601 CGTAAAAAAGTACGGTCCAAATGCCCCTTTCACTCTCTCTGCCCTTGAGGCTATAGGACAAGGAGGTTATCTGCTGCCCGGGGAATGGGTGCGAGTAGCT 1700

V K K Y G P N A P F T L S A L E A I G Q G G Y L L P G E W V R V A

1701 CAGGCCACCCTATCAAGAGGACAATTCCTAACCTGGAAAGCAGAATTTGATTATCACTGTCAAACCATCAAAAAAAGAAACTTGAAATCTAAGGCTAATA 1800

Q A T L S R G Q F L T W K A E F D Y H C Q T I K K R N L K S K A N

1801 GGGATTGGACTTTTGAAAAATTAGCAGGACGAGGTGAATATGCGTTAGAGACAAAACAAAGAAAGTTACCTACAGGCTTGCTAGAACAGACTGCTCACGC 1900

R D W T F E K L A G R G E Y A L E T K Q R K L P T G L L E Q T A H A

1901 GGCTAATCGCGCTTGGAGAGCCCTTCCTGTTACAGGCTCCCCCTTTACACCCCTTAATAAGATCACTCAAAGGAAAGATGAAGAGTACTCTGACTTTGTC 2000

A N R A W R A L P V T G S P F T P L N K I T Q R K D E E Y S D F V

2001 AGTCGACTGTTAGAGACAGCTGAGAGGACATTAGGAAATGAAGCATCAGATGATTTAATTATCAAAAGACTGGCTTTTGAGAATGCAAATGGCCCATGTA 2100

S R L L E T A E R T L G N E A S D D L I I K R L A F E N A N G P C

2101 GGAGTGTACTGAATGGTCAATGGCAGGATAAGACATTGAATGAGATGATTAAACTATGCCGCGATATCCAAGATCCTACAGCAGCAAAAATTGAAAAAAT 2200

R S V L N G Q W Q D K T L N E M I K L C R D I Q D P T A A K I E K M

2201 GTCTCAGGCGGGTATTGGCCTTACAAAACCCTATGAAAAACATGTCAGAGGCATTTCTAACTATAGGAGCAGCTGTTGGTAATACTGTGAAAACCTGTTT 2300

S Q A /V L A L Q N P M K N M S E A F L T I G A A V G N T V K T C F

2301 TAAATGTGGAGCCGAGGGTCATTTTGCCAGACAATGCCCCATGAGCCAACCTAACCCCCCCACTCAACGAAAGCGCACAACACCCGCTACTCCGTGTCCC 2400

K C G A E G H F A R Q C P M S Q P N P P T Q R K R T T P A T P C P

Frameshift

2401 AGGTGCAGAAAAGGGTTTCACTGGGGGAATACATGCAGAGCAACACATGATATAGAGGGCAAACCCCTTCCCCCGTTGTCGGGAAACGCCTGGCGGGGTC 2500

R C R K G F H W G N T C R A T H D I E G K P L P P L S G N A W R G

Pro> Y R G Q T P S P V V G K R L A G S

2501 CGCCCCGGACCCCTCAAGGACTGGCCCAGCCAGCCACGGGGACCCACCCTGCCCACAGCTCTCCAGAGCAACGGCAGGCAGTGCAGGAGTGGACTTGTGT 2600

P P R T P Q G L A Q P A T G T H P A H S S P E Q R Q A V Q E W T C V

A P D P S R T G P A S H G D P P C P Q L S R A T A G S A G V D L C

2601 GCCACCGCCGCCGCAATATTGAGACCGGAGGACGGCCCACAAATCCTCCCTACGGGAGTCTTCGGCCCTCCGCCCAAAGGTTCCTTTTTCCTCATTATTG 2700

P P P P Q Y <Gag

A T A A A I L R P E D G P Q I L P T G V F G P P P K G S F F L I I

2701 GCCGTGCTTCTACCACCCTTCAAGGGATAACAATACACCCAACTATTGTCGATAATGACTATGAAGGAGAGATTAAGATAATTGCTTCTGCTCTGAAAGA 2800

G R A S T T L Q G I T I H P T I V D N D Y E G E I K I I A S A L K D

2801 TACTGTGAGAATCCAGCCCAATCAGCGCATTGCCCAAGCCCTCCCTCTTCCTATGAATACCACCTTTCCTGCTTCTAAAAATAAGAGGGGCAATTCAACA 2900

T V R I Q P N Q R I A Q A L P L P M N T T F P A S K N K R G N S T

2901 CCTGGGTCCTCAGATATTTACTGGGCCCAGGCTCTGTCACAAGAGAGGCCTACACTAATGCTCAAAATACAAGGTAAACCCTTCATGGGCTTACTCGATT 3000

P G S S D I Y W A Q A L S Q E R P T L M L K I Q G K P F M G L L D

3001 CTGGGGCTGATTCAACCGTCATTTCTGAATCACACTGGCCGTCTACCTGGCCGTTACAGCCCTCTCTAACCCACCTACAAGGCATAGGGCAAAGTTCCAA 3100

S G A D S T V I S E S H W P S T W P L Q P S L T H L Q G I G Q S S N

3101 CACCATGCAATCCTCACAACTGTTACAATGGGAAGATCGAGAGGGTAATAGAGGAACAATCCGTCCCTTTGTAGTCCCATGCCTCCCCGTTAATCTATGG 3200

T M Q S S Q L L Q W E D R E G N R G T I R P F V V P C L P V N L W

3201 GGAAGAGATATCCTCTCACAAATGGGAGTTATCATGTGCAGCCCTAGCTCAGTTGTTACAGAACAAATGCTAAGTCAAGGTTTTTTACCCCGCCAAGGAC 3300

G R D I L S Q M G V I M C S P S S V V T E Q M L S Q G F L P R Q G

3301 TAGGAAAAAATAAACAGGGCATCACTCAACCTTTACATATACAACCTCACCCTGACCGCTCCGGCCTTGGTTTCCAGAAACATTTTTCATAAGGGCCACT 3400

L G K N K Q G I T Q P L H I Q P H P D R S G L G F Q K H F S <Pro

Pol> P L R P W F P E T F F I R A T

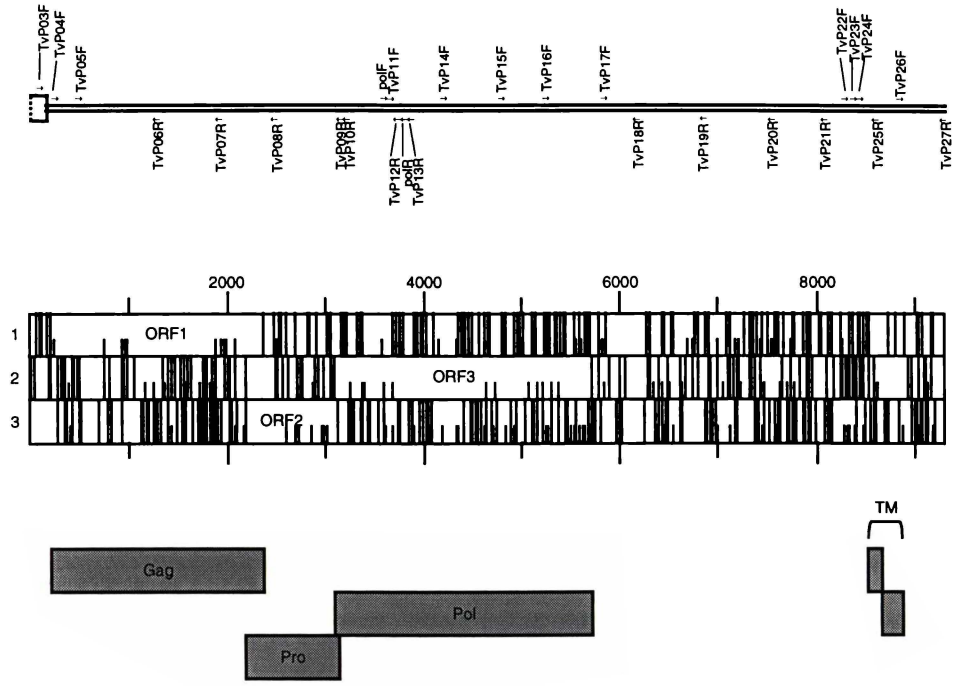
```
3401 GGTCCCCTCAGCCTACAGGCGGATAAAATCACATGGAGATCTGAGACTCCCGTCTGGATTGACCAGTGGCCCCTTCCTAAAGAAAAACTAGAGGCTGCAT 3500
       G  P  L  S  L  Q  A  D  K  I  T  W  R  S  E  T  P  V  W  I  D  Q  W  P  L  P  K  E  K  L  E  A  A

3501 TAATGTTAGTTCAACAACAATTGACTGCGGGTCACATAGAACCCTCCAACTCTCCTTGGAATACTCCAATTTTTGTTATTAAAAAGAAGTCAGGTGCCTG 3600
       L  M  L  V  Q  Q  Q  L  T  A  G  H  I  E  P  S  N  S  P  W  N  T  P  I  F  V  I  K  K  K  S  G  A  W

3601 GAGATTGCTCCATGATTTAAGGGCGGTCAATAAGACCATGATCCCCATGGGATCGCTACAACCGGGTCTTCCTGCACCCGTAGCAATCCCGGCAGGTTTC 3700
         R  L  L  H  D  L  R  A  V  N  K  T  M  I  P  M  G  S  L  Q  P  G  L  P  A  P  V  A  I  P  A  G  F

3701 CAAAAAGTTGTAATAGACCTCAAGGACTGCTTCTTTTCCATTCCTCTCCACCCCGACGACTCCAAAAGATTCGCCTTCACTGTCCCAGTTACCAATTGCG 3800
       Q  K  V  V  I  D  L  K  D  C  F  F  S  I  P  L  H  P  D  D  S  K  R  F  A  F  T  V  P  V  T  N  C

3801 TAGGACCCTCTCCTCGCTTTCAATGGAAAGTTCTCCCCCAAGGCATGACCAACAGCCCTACCCTTTGCCAGAAGTATGTTGCCCAGACAATCGACCCCTT 3900
       V  G  P  S  P  R  F  Q  W  K  V  L  P  Q  G  M  T  N  S  P  T  L  C  Q  K  Y  V  A  Q  T  I  D  P  F

3901 TCGCTTACAATTTCCACAACTTTATATCATTCACTATATGGATGACATTCTTATTGCTGGGCCCTGCGAGCAGGAATTACTTAGGATCACCACATTGCTA 4000
         R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D  I  L  I  A  G  P  C  E  Q  E  L  L  R  I  T  T  L  L

4001 ATAACTGCCTTAAAAGATAGGGGATTATACATATCCCCGGATAAAGTCCAGCTTGCCTCCCCTTTCTTTTTTCTTGGATTCGAATTGCATTCCTCTCATA 4100
       I  T  A  L  K  D  R  G  L  Y  I  S  P  D  K  V  Q  L  A  S  P  F  F  F  L  G  F  E  L  H  S  S  H

4101 TTATAACACAGAAAGCTCAGATACGGACAAATCATTTGCACACGCTGAATGACTTCCAGAAGTTATTGGGTGATATTAATTGGCTACGCCCTTACCTTAA 4200
       I  I  T  Q  K  A  Q  I  R  T  N  H  L  H  T  L  N  D  F  Q  K  L  L  G  D  I  N  W  L  R  P  Y  L  K

4201 AATCACCACAGGAGATTTGAAACCCCTCTTTGACATTCTGAAAGGTGACTCCAACCCAACCTCGCCCCGTGTTCTGACTAAAGAAGGTGAGGCTGCCCTA 4300
         I  T  T  G  D  L  K  P  L  F  D  I  L  K  G  D  S  N  P  T  S  P  R  V  L  T  K  E  G  E  A  A  L

4301 CAAAAAATAAATCATGCTATTTCGGCCCAAAAAATAGGATATTATAACCCCCAGAAAAAACTCTGGTTTCTTATCTTTTCTACATCATTCTCCCCTACGG 4400
       Q  K  I  N  H  A  I  S  A  Q  K  I  G  Y  Y  N  P  Q  K  K  L  W  F  L  I  F  S  T  S  F  S  P  T

4401 GCCTTTTATGGCAGGACCTGCCCCTATATTGGATCCATGGTCCTGCAACACCCACCAAAGTGCTACCCACCTACCCGCTTCTTACTGCCTCCCTCCTTCA 4500
       G  L  L  W  Q  D  L  P  L  Y  W  I  H  G  P  A  T  P  T  K  V  L  P  T  Y  P  L  L  T  A  S  L  L  Q

4501 GGAAGGCGGGTCTCTTTCAGTCAAATTTTTTGGAAAAGACCCCGACCACATTGTCTGCCCTTACACCTCTGCACAACTGTCATGGCTTCAACAGAATGAT 4600
         E  G  G  S  L  S  V  K  F  F  G  K  D  P  D  H  I  V  C  P  Y  T  S  A  Q  L  S  W  L  Q  Q  N  D

4601 GATACTTGGGCCATCTCCTCTGTATCTTATCAGGGAATCCTTGATAATCACTACCCTCCTGACAAATTGATCCAATTTCTTAGGGACAATCCAGTGGCCT 4700
       D  T  W  A  I  S  S  V  S  Y  Q  G  I  L  D  N  H  Y  P  P  D  K  L  I  Q  F  L  R  D  N  P  V  A

4701 TCCCCAGAATCACTAAAAGAGATCCTCTAGAAAATGCTCGTCTAGTCTTCACTGACGGGTCCTCCTCAGGAACTGCTGCCATATCTATAGACTCAAATCT 4800
       F  P  R  I  T  K  R  D  P  L  E  N  A  R  L  V  F  T  D  G  S  S  S  G  T  A  A  I  S  I  D  S  N  L

4801 AACTTCCTTTCATACACCTTATAAATCAGCCCAACTAGTGGAGCTTGCCGCTATACTTTGGGTCTTTGAACATGTACCGGACCCAATGAATATATACACA 4900
         T  S  F  H  T  P  Y  K  S  A  Q  L  V  E  L  A  A  I  L  W  V  F  E  H  V  P  D  P  M  N  I  Y  T

4901 GATAGTGCTTATATTGCTCATTCAATCCCCGAGCTTGAAACTGTTCCTTATATAAAACCTTCCACCAATGCCTTCTCTATGTTCCTTCAAATACAAAAAC 5000
       D  S  A  Y  I  A  H  S  I  P  E  L  E  T  V  P  Y  I  K  P  S  T  N  A  F  S  M  F  L  Q  I  Q  K

5001 TGATCAGAGAAAGAGCACAGCCTTTCTTCATTTGCCATATCAGAGCGCACTCTGACCTCCCTGGACCATTGGCAAAAGGGAACGATACTGTTGATGCGGC 5100
       L  I  R  E  R  A  Q  P  F  F  I  C  H  I  R  A  H  S  D  L  P  G  P  L  A  K  G  N  D  T  V  D  A  A

5101 TTCCAAGCACTTGGTTTTCGCCTCTGTATTGCAACCCTCCCCCATAGAGACTGCCACCAAGGCACACGAGTTGCACCATCTTAATGCCCACAGCTTAAGT 5200
         S  K  H  L  V  F  A  S  V  L  Q  P  S  P  I  E  T  A  T  K  A  H  E  L  H  H  L  N  A  H  S  L  S

5201 CAAATTTATAAAATCACCAGAGAACAGGCTAGACAGATTGTCAAACAATGTCCAGGATGTCTGGTGCTCCTGCCTGAACCTCACCTTGGAGTGAACCCCC 5300
       Q  I  Y  K  I  T  R  E  Q  A  R  Q  I  V  K  Q  C  P  G  C  L  V  L  L  P  E  P  H  L  G  V  N  P

5301 GAGGCCTGGTTCCCGGGGAACTGTGGCAGATGGACGTCACTCATTTCACACCTTTTGGTAAACTGAAATACATCCATGTCTCCATAGATACCTTTAGTGG 5400
       R  G  L  V  P  G  E  L  W  Q  M  D  V  T  H  F  T  P  F  G  K  L  K  Y  I  H  V  S  I  D  T  F  S  G

5401 TTTCCTCATAGCGACCCTACAAATGGGAGAAGCCACCAAGCACGTTATCAATCACATTGTGGCCTCATTAGCCATGGGGCCTAAACCCAAAATTCTCAAA 5500
         F  L  I  A  T  L  Q  M  G  E  A  T  K  H  V  I  N  H  I  V  A  S  L  A  M  G  P  K  P  K  I  L  K

5501 ACTGATAATGGACCGGGATATGTTAGTTCATCCTTTAAACGGTTTTGTGCATATATGAACATCAGAAATATCACAAGTATTCCATATAATCCCCAAGGAC 5600
       T  D  N  G  P  G  Y  V  S  S  S  F  K  R  F  C  A  Y  M  N  I  R  N  I  T  S  I  P  Y  N  P  Q  G

5601 AAGGTATTGTTGAGAGAGCCCATCAGACTTTGAAAAGAATGATATCTAAATTGCAATCGGGGTCAGAAAAACTATACATACGGGCCAATAACCATAGGAC 5700
       Q  G  I  V  E  R  A  H  Q  T  L  K  R  M  I  S  K  L  Q  S  G  S  E  K  L  Y  I  R  A  N  N  H  R  T

5701 CCTCCTAAATCATGCACTCTTCGTCCTAAATTTTCTTACATTGGACGCTCATGGCAAATCCGCCGCAGATCGCCTTTGGCACCCATCGTCAGCCCATGAT 5800
         L  L  N  H  A  L  F  V  L  N  F  L  T  L  D  A  H  G  K  S  A  A  D  R  L  W  H  P  S  S  A  H  D

5801 TATGCCCAAGTGAAATGGAAGGATCCCCTCACCCATCAGTGGCATGGCCCTGATCCAGTCCTCATATGGGGCAAGGGACATGCCTGTATATATGATTCCA 5900
       Y  A  Q  V  K  W  K  D  P  L  T  H  Q  W  H  G  P  D  P  V  L  I  W  G  K  G  H  A  C  I  Y  D  S

5901 ATGCACAAAACGCCAGATGGCTACCTGAACGGTTGCTTAAGTTAGTAGATCATGCCATCATTAAACACTAACAGCATCTTGTAATAAAACAGCTTTATGC 6000
       N  A  Q  N  A  R  W  L  P  E  R  L  L  K  L  V  D  H  A  I  I  K  H <Pol

6001 ATCCCTCTCAAACTGTTTAGTGCTCAAATCTCATCAGATGGCTGAGTAAGCATTACACTTTGGGACAGTAAATCTGTTTCCAATATGGTCTCCATTCACA 6100

6101 GATATGTACTTTGGTATTCATTCCAAGACACTTCCAGCTATTTCTTAGAATTGTTTGTACAAGGTTATATATCCCCTTGCTTGCCTTCTCCCCCTTGCCT 6200

6201 CTCTGCACTACCATCCCACCCTGGTAGTGGGAGATGCCCAATTCTCGCGAGACTTGATCAAATAAAACTTCTGTTTTGTTTCTACCTTGAGAAAACCGAG 6300

6301 ACACCTCTGTTTTTGTTTATTTTGAGCCTGTGGTTTGTCCCACACAGTTGGGGGCTCGGCCGGGACCCCCTCGATTGGGGGGGGTTCTCCTCACCAGGGA 6400

6401 GCAGGATCCAGCGCGCACCACAAAGATTTATTGGGGGTTCCTGGTCCTGGTGTGGGGCTCCACCCAGTTGGGGGAAAATCCCGAAGAGAGGGACTCGGAG 6500

6501 GAACAGAAGCTGTGCTAGGTTCCGGAGCGCATATGTGGGGTACCTTTAGGCAATCGATTGCAAAATTGGAATGAACAACAAAAGTAGAAAGGAAAATGGC 6600

6601 CATAGACAAGAGCTCTTAAAATATTTGACCTAGGTGGGTGGAATGGATTTTTCCTGGTGTTTTGAGAACAGCTCTCCCAGGGAAAATGAAAAACAAAAAG 6700

6701 CTTTTGGATTTCTGTATCTTTGTGTCTGTGTCTGTCTGTATCTGTGCCCCTGTGTGTGTAATCTGACACCCTGTAATGTCCACTATCTGGCTTTCTCTCC 6800

6801 CTCCCTGACTCCTCCTGGCATGCAGAGGAGCGGGAGGGCGGGGGAGAGAGAAAGACAGAGAAGAATGTTTTCTTGTTTAAGTTATGTGATGGAAAAGGCA 6900

6901 AGGTAAAGAGGAATGTCAGGTCAGAATTAGAAAGAAATAAAGTATGTGCAGTTTTACATGGGCTTTCATCAAAGTCCCACCAGAAAGGAAAGGCAAGAAA 7000
```

```
7001  GTTTAAGCATGTAGGAGAGAGGAGGAGGTATAGAAAGTTTAAAAAAGGCATAATCTATTTGCACAAATAGCGTGAATAGAGGTTTTGAGAGTGCAAGGTG  7100

7101  GAAACTTTTGGGAAGTTGTTTTAAAAAAATGATATATATATATATATGCTCTTTCTGTCTCCCTCCATCCCAGATCTGGCCAATTTGGAATGTGGAGAGA  7200

7201  CAAGAAAGGAGGCGGGGACCTTTTCCCTGAAATTCAAATGTGTTACTCCTGATTCGATGCTTAAGATAAAGTCAGAATTTCTCATACCATTTGGCACTAA  7300

7301  GGCAAATTTGGAAAATGCATAAGGAAGAGACACTCTGAGTCTCCCAGTATGAGGAGAAAGGGACATGGTACCACTGGATCTTTTCCGGAGTCCCACTCAC  7400

7401  CTTAGACTGGCAAATTCAAAGCTCTCACTTCCAAAAAAAGTTTTTAAGGAGTGAACACAGAAGTAAAATTTACTTGAGAAAATTAGAACCATTAAGATTT  7500

7501  TGATTTAGCAGATTCTGTTAACCATAGATAATCTGGTGAGGTATATAAAACCTTTTTCAAAATAATAATGCACGTATGAATGTGAAGAAAAATCAGTTTA  7600

7601  TATGACGAAAGATATTCTGTATCCCCCTCCAATCCGTAGACTCCTTGGGGAGTTAAGAATTCCAGCCCTATGTCAGAAGTATATACAAAGATGTTTTAAT  7700

7701  TCCTGTGTGGTGAATGAGAATCACCATGAGACTAGTATAGATAAAAAAGTTGGAAAATATTTAATGACATAACGGTATTAAATCCATAGAGTTAATAGTA  7800

7801  GTTTTCAATATACAACTACAGAGATCCTTATGTGCAATTTAGTGACCCTTGTTCCTATTTGAATTTAAAAATGGCTTAAAGCTTCCAAAATAGAGAAATA  7900

7901  TGACCAGAGAATGAGAGCCAATCTGAACATATTTTACCCAAAAGACTGTTGCTTATTACATTCCTGAGAAACAGAAATACTTCCTAAAATAGAAATTAGC  8000

8001  TCAATAGATATAGATATTTCTAACTTGCTGTGATTTTCCTAAGTTTGTGACTTGAAAAGTTTAAAAAGTGATTTGGCTGCGAATATACAATAGAGGGTTG  8100

8101  GCTCCTTTTCAAATATAATTACCCCACTGACAACAGCCCTGGACCCCGGGCACCCTGAAGCCCACACTTCTAGGCAATTGCCCCCCTCCCACTGGAACCT  8200

8201  GGAATTTTTGCTCCTAGTCTTCCAAACCTGACAACCCCTGTTAGAAAGTAAAGATCTCTCAGAGACCCTCCTAAATGACTCCTTAGACCTTTAATTCAGG  8300

8301  TGCTGGACTCTGCAGCTTCCCAGCACCTTCCCCCCCTTCGCGCCCCCCCCCCTTTTTTCAGTCCACAGCTAGGCTCTGCAACTATGAGGAGAGAGGAAAA  8400

8401  GCCTCTATGCTTGGTAAGCCTAAAATATGTAATGATTTGGGAATGCAATTAATGTAATCTGAAAATCTTTCTATTTGTACTATTAGTGTGCTTCTAGATT  8500

8501  GTAGTAAAATGTCCTATATTGCAAAACTTAAGTAACCGGAGATAGTGTTAGACAAAGCAATTATTAATCTGAATTCTGTTGAAGAGAAGAAAAAAAAAAT  8600

8601  GATATGAAAACAATTGTGTAAAACTGGTTAAGTTACTTCCTCAGCCTAATCTGGCATAGTAAGGGATAAAATTAAGATATTGCAAGAAGTATTGGTCCAG  8700
                                                    TM> S  G  I  V  R  D  K  I  K  I  L  Q  E  V  L  V  Q

8701  AGGCGTAAAGAGCTTTTAAATAACCCCCTCTGGAATGGTTTACATGGGCTTCTCCCATACCTCCTCCCCCTTCTTGGGCCTCTGGTGGGTCTTCTCTTGC  8800
       R  R  K  E  L  L  N  N  P  L  W  N  G  L  H  G  L  L  P  Y  L  L  P  L  L  G  P  L  V  G  L  L  L

8801  TCCTTTCCTTTGGCCCCTGGGCTTTTAAACAAATTAACCACTTTTGTTAAGTCTCAAGTAGATTCAGCCATTGCCAGGAAGAGCCACATATTTTACAGCA  8900
      L  L  S  F  G  /P  G  L  L  N  K  L  T  T  F  V  K  S  Q  V  D  S  A  I  A  R  K  S  H  I  F  Y  S

8901  GACTACTGGAACAAGACACCACTGAACAGCAGGAAGAGCGACTACAATTCACGAAAGACCTTCTGAAGGAGCCCTGGCATCAGCGTTGGAAGACAAACCT  9000
      R  L  L  E  Q  D <TM

9001  TTATGCCCTGTGCTGTAAAGCCAATCATCCCCTTCCTTTCCCCCACTCGCCTTAGCTCTGTCACCAAATTCTTATCTTTTGACCTTTGAATTCACTCCTC  9100

9101  CCACCCTTCTTTCTCTATTTAAACAGTCCTGCTGACCCTCTGAAGGACGGCTGTGGCTGAGCCCATGACGGGATTATTGCAACCCTGAATGAAGGATGGT  9200

9201  TGAGGTCTGGCTGAAACTGGCCGTAAGACCCTCTGATCACCATGACGGGCAACCTCCAGTGGCGCAAGTCGCAGGGCTGTAGTACCCTACTGTAAACACC  9300

9301  CCGCCCGGATGACCTAAGACAGTGCTCAAGGGCGGGCTCTCAAAATCCATTGTACACCAGACGCTCTCACCTCATTATCAGACCCAAGTTGCATAGCCTA  9400

                    PPT          IR U3
9401  AGACAGGCCAGTCACCCTCTCCTTAAAATAAAAATGGGGGACATGCCGGAGACTGGAAAATCGCCTTACCCCTCCCCTCTGCCTGACCCGCGCCCCCTCC  9500

                                                                    CCAAT box
9501  CTTCGCTAGCTTTCATTTCTCGCTTAGCTCTCATTTCCTGGCTAACTCTCGTTTCTGGGATATGACGCAGGCCTGCGTCCAATGCTTAGCCACCCCCTCTG  9600

                          TATA box                Poly(A)  U3.←—    R  .  ←— U5
9601  GGGACGTGGCATTCTGATAACATGCTCATAAAAACCCTTACCAGGAACAATAAAGAGAGACATTTTTTACACCAGCAGCGCGTCTAGTCTCCTTCCTCTC  9700

9701  CGTGATCCCGTTCCCGAGCCTCTTGGCCCTTAGAGCCATTAGGTTCCCCTCGAGCCTAGCTCCCCCGGACCTCTCCAGCTTCGCTAACCCACGATTGTGT  9800

          U5 IR
9801  GACCTGCGGGCCAGGTCA                                                                                  9818
```

Nucleotides 1–230: PCR800/TvP26F
Nucleotides 231–263: spPCR3/TvP01F
Nucleotides 264–2686: clone pTvRV5'
Nucleotides 2687–3294: PCR3000/TvP09R
Nucleotides 3295–3881: spPCR1/TvP12R
Nucleotides 3882–8579: clone pTvRV3'
Nucleotides 8580–8957: spPCR2/TvP23F
Nucleotides 8958–9673: PCR800
Nucleotides 9674–9818: spPCR3/TvP01F

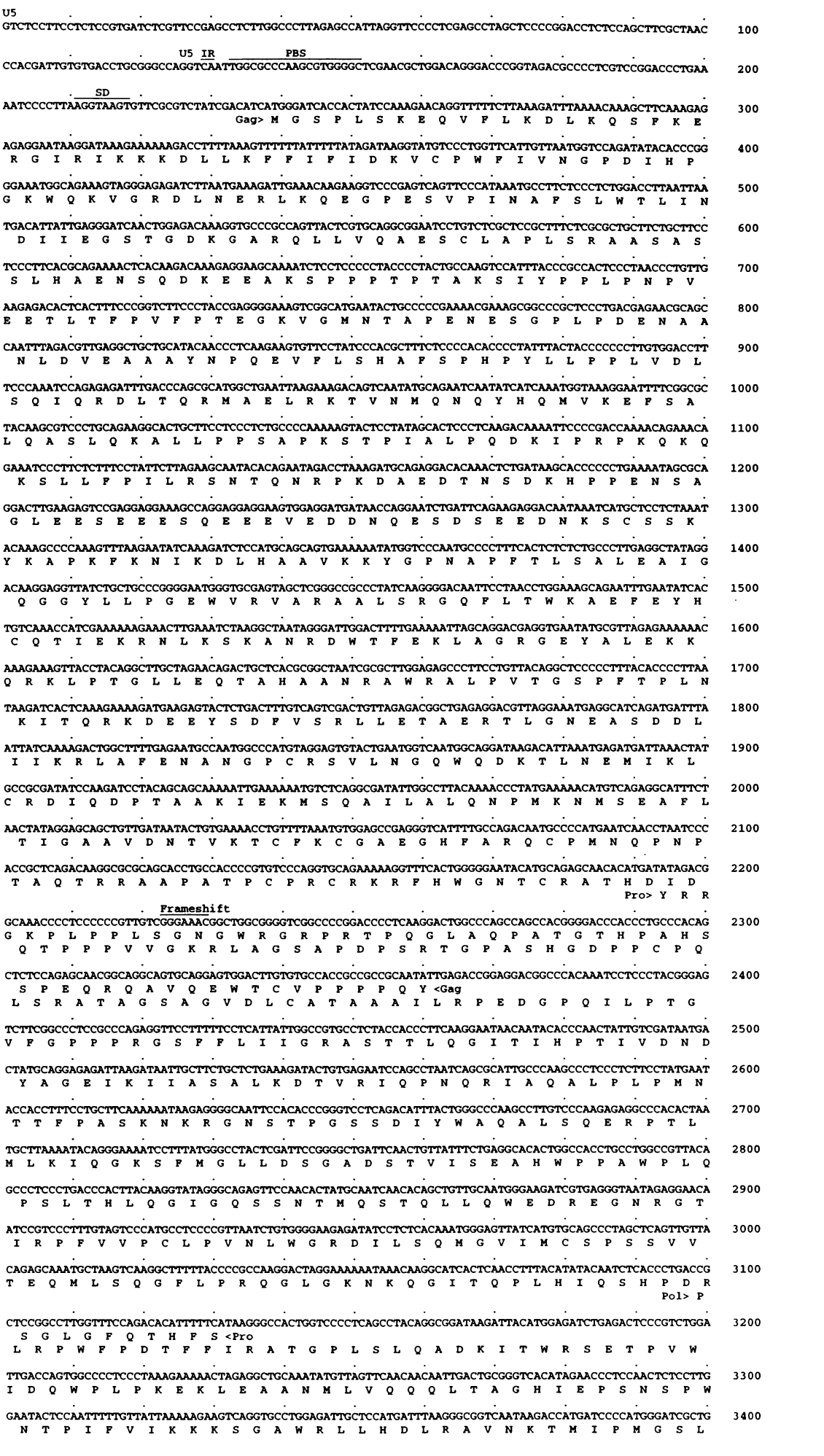
Appendix 5 – pTvERV-K2

The pTvERV-K2 insert was cloned from PCR products amplified from possum genomic DNA using the TvP03F and TvP27R primers (see 4.2.7 Amplification and cloning of a near-full-length endogenous retrovirus). The positions of the primers used to amplify and sequence the various regions of the TvERV-K1 contig and to sequence pTvERV-K2 are indicated below. Also shown below are the ORF map and corresponding proteins encoded by pTvERV-K2. The table at the bottom of this page indicate the expected sizes (based on the pTvERV-K2 sequence) of amplification products using various primer pairs. The sequence of TvERV-K2 is listed on pp. 177–179, and relevant regulatory regions and protein products are shown.



Expected PCR product sizes based on the sequence of pTvERV-K2

Primer name	TvP06R	TvP07R	TvP08R	TvP09R	TvP10R	TvP12R	TvP13R	TvP18R	TvP19R	TvP20R	TvP21R	TvP25R	TvP27R
TvP03F	1236	1831	2449	3080	3143	3658	3709	6083	6859	7501	8104	8584	9260
TvP04F	1047	1642	2260	2891	2954	3469	3520	5894	6670	7312	7915	8395	9071
TvP05F	839	1434	2052	2683	2746	3261	3312	5686	6462	7104	7707	8187	8863
TvP11F	-	-	-	-	-	57	108	2482	3258	3900	4503	4983	5659
TvP14F	-	-	-	-	-	-	-	1964	2740	3382	3985	4465	5141
TvP15F	-	-	-	-	-	-	-	1376	2152	2794	3397	3877	4553
TvP16F	-	-	-	-	-	-	-	906	1682	2324	2927	3407	4083
TvP17F	-	-	-	-	-	-	-	347	1123	1765	2368	2848	3524
TvP22F	-	-	-	-	-	-	-	-	-	-	-	339	1015
TvP23F	-	-	-	-	-	-	-	-	-	-	-	180	856
TvP24F	-	-	-	-	-	-	-	-	-	-	-	159	835
TvP26F	-	-	-	-	-	-	-	-	-	-	-	-	466


U5
U5 IR
PBS
SD
Gag> M G S P L S K E Q V F L K D L K Q S F K E
R G I R I K K K D L L K F F I F I D K V C P W F I V N G P D I H P
G K W Q K V G R D L N E R L K Q E G P E S V P I N A P S L W T L I N
D I I E G S T G D K G A R Q L L V Q A E S C L A P L S R A A S A S
S L H A E N S Q D K E E A K S P P P T P T A K S I Y P P L P N P V
E E T L T F P V F P T E G K V G M N T A P E N E S G P L P D E N A A
N L D V E A A A Y N P Q E V F L S H A F S P H P Y L L P P L V D L
S Q I Q R D L T Q R M A E L R K T V N M Q N Q Y H Q M V K E F S A
L Q A S L Q K A L L P P S A P K S T P I A L P Q D K I P R P K Q K Q
K S L L F P I L R S N T Q N R P K D A E D T N S D K H P P E N S A
G L E E S E E E S Q E E E V E D D N Q E S D S E E D N K S C S S K
Y K A P K F K N I K D L H A A V K K Y G P N A P F T L S A L E A I G
Q G G Y L L P G E W V R V A R A A L S R G Q F L T W K A E F E Y H
C Q T I E K R N L K S K A N R D W T F E K L A G R G E Y A L E K K
Q R K L P T G L L E Q T A H A A N R A W R A L P V T G S P F T P L N
K I T Q R K D E E Y S D F V S R L L E T A E R T L G N E A S D D L
I I K R L A F E N A N G P C R S V L N G Q W Q D K T L N E M I K L
C R D I Q D P T A A K I E K M S Q A I L A L Q N P M K N M S E A F L
T I G A A V D N T V K T C F K C G A E G H F A R Q C P M N Q P N P
T A Q T R R A A P A T P C P R C R K R F H W G N T C R A T H D I D
Pro> Y R R
Frameshift
G K P L P P L S G N G W R G R P R T P Q G L A Q P A T G T H P A H S
Q T P P P V V G K R L A G S A P D P S R T G P A S H G D P P C P Q
S P E Q R Q A V Q E W T C V P P P P Q Y <Gag
L S R A T A G S A G V D L C A T A A A I L R P E D G P Q I L P T G
V F G P P P R G S F F L I I G R A S T T L Q G I T I H P T I V D N D
Y A G E I K I I A S A L K D T V R I Q P N Q R I A Q A L P L P M N
T T F P A S K N K R G N S T P G S S D I Y W A Q Q A L S Q E R P T L
M L K I Q G K S F M G L L D S G A D S T V I S E A H W P P A W P L Q
P S L T H L Q G I G Q S S N T M Q S T Q L L Q W E D R E G N R G T
I R P F V V P C L P V N L W G R D I L S Q M G V I M C S P S S V V
T E Q M L S Q G F L P R Q G L G K N K Q G I T Q P L H I Q S H P D R
Pol> P
S G L G F Q T H F S <Pro
L R P W F P D T F F I R A T G P L S L Q A D K I T W R S E T P V W
I D Q W P L P K E K L E A A N M L V Q Q Q L T A G H I E P S N S P W
N T P I F V I K K K S G A W R L L H D L R A V N K T M I P M G S L

```
3401  CAACCGGGTCTTCCTGCACCCGTAGCAATCCCGGCAGGCTTCCAAAAAATTGTAATAGACCTCAAGGACTGCTTCTTTTCCATTCCTCTCCACCCCGACG  3500
        Q  P  G  L  P  A  P  V  A  I  P  A  G  F  Q  K  I  V  I  D  L  K  D  C  F  F  S  I  P  L  H  P  D

3501  ACTCCAAAAGATTCGCCTTCACTGTCCCAGTTACCAATTGCGTAGGACCCTCTCCTCGCTTTCAATGGAAAGTTCTCCCCCAAGGCATGACCAACAGCCC  3600
      D  S  K  R  F  A  F  T  V  P  V  T  N  C  V  G  P  S  P  R  F  Q  W  K  V  L  P  Q  G  M  T  N  S  P

3601  TACCCTCTGCCAGAAGTATGTTGCCCAGACAATCGACCCCTTTCGCTTACAATTTCCACAACTTTATATCATTCACTATATGGATGACATTCTTATTGCT  3700
        T  L  C  Q  K  Y  V  A  Q  T  I  D  P  F  R  L  Q  F  P  Q  L  Y  I  I  H  Y  M  D  D  I  L  I  A

3701  GGGCCCTGCGAGCAGGAATTACTTAGGATCACCACATTGCTAATAACTGCCTTGAAAGATAGGGGATTATACATATCCCCGGATAAAGTCCAGCTTGCCT  3800
        G  P  C  E  Q  E  L  L  R  I  T  T  L  L  I  T  A  L  K  D  R  G  L  Y  I  S  P  D  K  V  Q  L  A

3801  CCCCCTTCTTTTTTCTTGGATTCGAATTGCATTCCTCTCATATTATAACACAGAAAGCTCAGATACGGACAAATCATTTGCACACGCTGAATGACTTCCA  3900
      S  P  F  F  F  L  G  F  E  L  H  S  S  H  I  I  T  Q  K  A  Q  I  R  T  N  H  L  H  T  L  N  D  F  Q

3901  AAAGTTATTGGGTGATATTAATTGGCTACGCCCTTACCTTAAAATCACCACAGGAGATTTGAAACCCCTCTTTGACATTCTGAAAGGTGACTCCAACCCA  4000
        K  L  L  G  D  I  N  W  L  R  P  Y  L  K  I  T  T  G  D  L  K  P  L  F  D  I  L  K  G  D  S  N  P

4001  ACCTCGCCCCGTGTTCTGACTAAAGAAGGTGAGGCTGCCCTACAAAAAATAAATCATGCTATTTCGGCCCAAAAAATAGGATATTATAACCCCCAGAAAA  4100
      T  S  P  R  V  L  T  K  E  G  E  A  A  L  Q  K  I  N  H  A  I  S  A  Q  K  I  G  Y  Y  N  P  Q  K

4101  AACTCTGGTTTCTTATCTTTTCTACATCATTCTCCCCTACGGGCCTTTTATGGCAGGACCTGCCCCTATATTGGATCCATGGTCCTGCAACACCCACCAA  4200
      K  L  W  F  L  I  F  S  T  S  F  S  P  T  G  L  L  W  Q  D  L  P  L  Y  W  I  H  G  P  A  T  P  T  K

4201  AGTGCTACCCACCTACCCGCTTCTTACTGCCTCCCTCCTTCAGGAAGGCAGGTCTCTTTCAGTCAAATTTTTTGGAAAAGACCCCGACCACATTGTCTGC  4300
        V  L  P  T  Y  P  L  L  T  A  S  L  L  Q  E  G  R  S  L  S  V  K  F  F  G  K  D  P  D  H  I  V  C

4301  CCTTACACCTCTTCACAACTGTCATGGCTTCAACAGAATGATGATACTTGGGCCATCTCCTCTGTATCTTATCAGGGAATCCTTGATAATCACTACCCTC  4400
        P  Y  T  S  S  Q  L  S  W  L  Q  Q  N  D  D  T  W  A  I  S  S  V  S  Y  Q  G  I  L  D  N  H  Y  P

4401  CTGACAAATTGATCCAATTTCTTAGGGACAATCCAGTGGCCTTCCCCAGAATCACTAAAAGAGATCCTCTAGAAAATGCTCGTCTAGTCTTCACTGACGG  4500
      P  D  K  L  I  Q  F  L  R  D  N  P  V  A  F  P  R  I  T  K  R  D  P  L  E  N  A  R  L  V  F  T  D  G

4501  GTCCTCCTCAGGAACTGCTGCCATATCTATAGACTCAAATCTAACTTCCTTTCATACACCTTATAAATCAGCCCAACTAGTGGAGCTTGCCGCTATACTT  4600
        S  S  S  G  T  A  A  I  S  I  D  S  N  L  T  S  F  H  T  P  Y  K  S  A  Q  L  V  E  L  A  A  I  L

4601  TGGGTCTTTGAACATGTACCGGACCCAATGAATATATACACAGATAGTGCTTATATTGCTCATTCAATCCCCGAGCTTGAAACTGTTCCTTATATAAAAC  4700
        W  V  F  E  H  V  P  D  P  M  N  I  Y  T  D  S  A  Y  I  A  H  S  I  P  E  L  E  T  V  P  Y  I  K

4701  CTTCCACCAATGCCTTCTCTATGTTCCTTCAAATACAAAAACTGATCAGAGAAAGAGCACAGCCTTTCTTCATTTGCCATATCAGAGCGCACTCTGACCT  4800
      P  S  T  N  A  F  S  M  F  L  Q  I  Q  K  L  I  R  E  R  A  Q  P  F  F  I  C  H  I  R  A  H  S  D  L

4801  CCCTGGACCATTGGCAAAAGGGAACGATACTGTTGATGCGGCTTCCAAGCACTTGGTTTTCGCCTCTGTATTGCAACCCTCCCCCATAGAGACTGCCATC  4900
        P  G  P  L  A  K  G  N  D  T  V  D  A  A  S  K  H  L  V  F  A  S  V  L  Q  P  S  P  I  E  T  A  I

4901  AAGGCACACGAGTTGCACCATCTTAATGCCCACAGCTTAAGTCAAATTTATAAAATCACCAGAGAACAGGCTAGACAGATTGTCAAACAATGTCCAGGAT  5000
        K  A  H  E  L  H  H  L  N  A  H  S  L  S  Q  I  Y  K  I  T  R  E  Q  A  R  Q  I  V  K  Q  C  P  G

5001  GTCTGGTGCTCCTGCCTGAACCTCACCTTGGAGTGAACCCCCGAGGCCTGGTTCCCGGGGAACTGTGGCAGATGGACGTCACTCATTTCACACCTTTTGG  5100
      C  L  V  L  L  P  E  P  H  L  G  V  N  P  R  G  L  V  P  G  E  L  W  Q  M  D  V  T  H  F  T  P  F  G

5101  TAAACTGAAATACATCCATGTCTCCATAGATACCTTTAGTGGTTTCCTCATAGCGACCCTACAAATGGGAGAAGCCACCAAGCACGTTATCAATCACATT  5200
        K  L  K  Y  I  H  V  S  I  D  T  F  S  G  F  L  I  A  T  L  Q  M  G  E  A  T  K  H  V  I  N  H  I

5201  GTGGCCTCATTAGCCATGGGGCCTAAACCCAAAATTCTCAAAACTGATAATGGACCGGGATATGTTAGTTCATCCTTTAAACAGTTTTGTGCATATATGA  5300
        V  A  S  L  A  M  G  P  K  P  K  I  L  K  T  D  N  G  P  G  Y  V  S  S  S  F  K  Q  F  C  A  Y  M

5301  ACATCAGAAATATCACAGGTATTCCATATAATCCCCAAGGACAAGGTATTGTTGAGAGAGCCCATCAGACTTTGAAAAGAATGATATCTAAATTGCAATC  5400
      N  I  R  N  I  T  G  I  P  Y  N  P  Q  G  Q  G  I  V  E  R  A  H  Q  T  L  K  R  M  I  S  K  L  Q  S

5401  GGGGTCAGAAAAACTATACATACGGGCCAATAACCATAGAACCCTCCTAAATCATGCACTCTTCGTCCTAAATTTTCTTACATTGGACGCTCATGGCAAA  5500
        G  S  E  K  L  Y  I  R  A  N  N  H  R  T  L  L  N  H  A  L  F  V  L  N  F  L  T  L  D  A  H  G  K

5501  TCCGCCGCAGATCGCCTTTGGCACCCATCGTCAGCCCATGATTATGCCCAAGTGAAATGGAAGGATCCCCTCACCCATCAGTGGCATGGCCCTGATCCAG  5600
        S  A  A  D  R  L  W  H  P  S  S  A  H  D  Y  A  Q  V  K  W  K  D  P  L  T  H  Q  W  H  G  P  D  P

5601  TCCTCATATGGGGCAAGGGACATGCCTGTATATATGATTCCAATGCACAAAACGCCAGATGGCTACCTGAACGGTTGCTTAAGTTAGTAGATCATGCCAT  5700
      V  L  I  W  G  K  G  H  A  C  I  Y  D  S  N  A  Q  N  A  R  W  L  P  E  R  L  L  K  L  V  D  H  A  I

5701  CATTAAACACTAACAGCATCTTGTAATAAAACAGCTTTATGCATCCCTCTCAAACTGTTTAGTGCTCAAATCTCATCAGATGGCTGAGTAAGCATTACAC  5800
        I  K  H  <Pol

5801  TTTGGGACAGTAAATCTGTTTCCAATATGGTCTCCATTCACAGATATGTACTTTGGTATTCATTCTAAGACACTTCCAGCTATTTCTTAGAATTGTTTGT  5900

5901  ACAAGGTTATATATCCCCTTGCTTACCTTCTCCCCCTTGCCTCTCTGCACTACCATCCCACCCTGGTAGTGGGAGATGCCCAATTCTCGCGAGACTTGAT  6000

6001  CAAATAAAACTTCTGTTTTGTTTCTACCTTGAGAAAACCGAGACACCTCTGTTTTTGTTTATTTTGAGCCTGTGGTTTGTCCCACACAGTTGGGGGCTCG  6100

6101  GCCGGGACCCCCTCGATTGGGGGGGGTTCTCCTCACCAGGGAGCAGGATCCAGCGCGCACCACAAAGATTTATTGGGGGTTCCTGGTCCTGGTGTGGGGC  6200

6201  TCCACCCAGTTGGGGGAAAATCCCGAAGAGAGGGACTCGGAGGAACAGAAGCTGTGCTAGGTTCCGACATTCCAAAAAGTAAGAGATAAGAGAAAAGAGT  6300

6301  CTGTTGACTCTCTGACTCGGGGAGTGCGGAGACGGCTCCCGACCAAATTATGGGGAAAATCGATCTAAGTCTGTGCAGTCGGAGCACATATGTGAGGTAC  6400

6401  CTTTAGGCAATCGATTGCAAAATTGGAATGAACAACAAAAGTAGAAAGGAAAATGGCCATAGACAAGAGCTCTTAAAATATTTGACCTAGGTGGGTGGAA  6500

6501  TGGATTTTTCCTGGTGTTTTGAGAAAAGCTCTCCCAGGGAAAATGAAAAACAAAAAGCTTTTGGATTTCTGTATCTTTGTGTCTGTGTCTGTCTGTATCT  6600

6601  GTGCTTCTGTGTGTGTGTAATCTGACACCCTGTAATGTCCACTATCTGGCTTTCTCTCCCTCCCTGACTCCTCCTGGCATGCAGAGGAGCGGGAGGGCGG  6700

6701  GGGAGAGAGAAAGACAGAGAAGAATATTTTCTTGTTTAAGTTATGTGATGGAAAAGGCAAGGTAAAGAGGAGTGTCAGGTCAGAATTAGAAAGAAATAAG  6800

6801  GTATGTGCAGTTTTACAGGGGCTTTCATCAAAGTCCACCAGAAAGGAAAGGCAAGAAAGTTTAACCATGTAGGGAAGAGACGTGTGACGGAACTTTAAAA  6900

6901  AAGGCATAATCTATTTGCACAAATAGTGTGAATAGAGGTTTTGAGAGTGCAAGGTGGAAACTTTTGGGAAGTTGTTTTAAAAAAATGATATATATATATA  7000
```

```
7001  TATGCTCTTTCTGTCTCCCTCCATCCCAGATCTGGCCAATTTGGAATGTGGAGAGACAAGAAAGGAGGCGGGGACCTTTTCCCTGAAATTCAAATGTGTT  7100

7101  ACTCCTGATTTGATGCTTAAGATAAAGTCAGAATTTCTCATACCATTTGGCACTAAGGCAAATTCGGAAAATGCATAAATCTCTACTAAAGAGACACTTT  7200

7201  GAGCTCCCCAGTATGAGGAGAAAGGGACATGGTACCACTGGATCTTTTCCGGAGTCCCACTCACCTTAGACTGGCAAATTCAAAGCTCTCACTTCCAAAA  7300

7301  AAAGTTTTTAAGGAGTGAACACAGAAGTAAAATTTACTTGAGAAAATTAGAACCATTAAGATTTTGATTTAGCAGATTCTGTTAACCATAGATAATCTGG  7400

7401  TGAGGTATATAAAACCTTTTTCAAAATAATAATGCACGTATGAATGTGAACAAAAATCAGTTTATATGACAAAAGATATTCTGTATCCCCCTCCAATCCG  7500

7501  TAGACTCCTTGGGGGGTTAAGAATTCCAGCCCTATGTCAGAAGTATATACAAAGATGTTTTAATTCCTGTGTGGTGAATGAAAATCACCATGAGACTAGT  7600

7601  ATAGATAAAAAAGTTGGAAAATATTTAATGACATAACGGTATTAAATCCATAGAGTTAATAGTAGTTTTCAATATACAACTACAGAGATCCTTATGTGCA  7700

7701  ATTTAGTGACCCTTGTTCCTATTTGAATTTAAAAATGGCTTAAAGCTTCCAAAATAGAGAAATATGACCAGAGAATGAGAGCCAATCTGAACGTATTTTA  7800

7801  CCCAAAAGACTGTTGCTTATTACATTCCTGAGAAACAGAAATACTTCCTAAAATAGAAATTAGCTCAATAGATATAGATATTTCTAACTTGCTGTGATTT  7900

7901  TCCTAAGTTTGTGACTTGAAAAGTTTAAAAAGTGATTTGGCTGCGAATATACAATAGAGGGTTGGCTCCTTTTCAAATATAATTACCCCACTGACAACAG  8000

8001  CCCTGGACCCCGGGCACCCTGAAGCCCACACTTCTAGGCAATTGCCCCCCTCCCACTGGAACCTGGAATTTTTGCTCCTAGTCTTCCAAACCTGACAACC  8100

8101  CCTGTTAGAAAGTAAAGATCTCTCAGAGACCCTCCTAAATGACTCCTTAGACCTTTAATTCAGGTGCTGGACTCTGCAGCTTCCCAGCACCTTCCCCCCC  8200

8201  CCCCCCGTTCGCCCCCTCCTCCCGCCTTCAGTCCATAGCTAGGCTCTGCAACTATGAGGAGAGAAGAACAGCCTCTATGCTTGGTAAGCCTAAAATATGT  8300

8301  AATAATTTGGGAATGCAATTAATGTCATCTGAAAATCTTTCTATTTGTACTATTAGTGTGCTTCTAGATTGTAGTAAAATGTCCTATATTGCAAAACTTA  8400

8401  AGTAACCGGAGATAGTGTTAGACAAAGCAATTATTAATCTGAATTCTGTTGAAAAGAAAAAAAAAAATGATATGAAAACAATTGTGTAAAACTGGCTAAG  8500

8501  TTACTTCCTCAGCCTAATCTGGCATAGTAAGGGATAAAATTAAGAAATTGCAAAAAGAATTGGTCCAGAGGCGTAAAGAGCTTTTAAATAACCCCCTCTG  8600
                TM> S  G  I  V  R  D  K  I  K  K  L  Q  K  E  L  V  Q  R  R  K  E  L  L  N  N  P  L  W

8601  GAACGGTTTACATGGGCTTCTCCCATACCTCCTCCCCCTTCTTGGTCCTCTGGTGGGTCTTCTCTTGCTCCTTTCCTTTGGCCCCTGGGCTTTTAAACAA  8700
        N  G  L  H  G  L  L  P  Y  L  L  P  L  L  G  P  L  V  G  L  L  L  L  L  S  F  G  ^P  G  L  L  N  K

8701  ATTAACCACTTTTGTTAAGTCTCAAGTAGATTCAGCCATTGCCAGGAAGAGCCACATATTTTACAGCAGACTACAGGAAGAAGACACCACTGAACAGCAG  8800
        L  T  T  F  V  K  S  Q  V  D  S  A  I  A  R  K  S  H  I  F  Y  S  R  L  Q  E <TM

8801  GAAGAGCGACTACAATTCACGGAAGACCTTCTGAAGGAGCCCTGGCATCAGCGTTGGAAGACAAACCATTATGCCCTGTGCTGTAAAGCCAATCATCCCC  8900

8901  TTCCTTTCCCCCACTCGCCTTAGCTCTGTCACCAAATTCTTATCTTTTGACCTTTGAATTCACTCCTCCCACCCTTCTTTCTCTATTTAAACAGTCCTGC  9000

9001  TGACCCTCTGAAGGACGGCTGTGGCTGAGCCCATGACGGGATTATTGCAACCCTGAATGAAGGATGGTTGAGGTCTGGCTGAAACTGGCCGTAAGACCCT  9100

9101  CTGATCACCATGACGGGCAACCTCCAGTGGCGCAAGTCGCAGGGCTGTAGTACCCTACTGTAAACACCCCGCCCGGATGACCTAAGACAGTGCTCAAGGG  9200

9201  CGGGCTCTCAAAATCCATTGTACACCAGACGCTCTCACCTCATTATCAGACCCAAGTTGC                                          9260
```

Appendix 6 – RT alignment

The amino acid sequences below were manually aligned using the alignment in Xiong and Eickbush (1990) as a guide. The overlined sequences were used to construct a phylogenetic tree using CLUSTAL X and the BLOSUM weight matrix, as described in Chapter 2 – Materials and Methods. The resulting tree is shown in FIG. 5.11. Retrovirus names, accession numbers, and references are listed in Appendix 1.

```
MPMV      QWPLTNDKLAAAQQLVQEQ-LEAGHITE--SSSPWNTPIFVIKKKSGK------WRLLQDLRAVNATMVLMGALQPGLPSPVA---IPQGYLKIIIDLKD
SRV-1     QWPLTSEKLAAAQQLVQEQ-LEAGHITE--SNSPWNTPIFVIKKKSGK------WRLLQDLRAVNATMVLMGALQPGLPSPVA---IPQGYLKIIIDLKD
SRV-2     QWPLTQEKLAAAQQLVQEQ-LQAGHIIE--SNSPWNTPIFVIKKKSGK------WRLLQDLRAVNATMVLMGALQPGLPSPVA---IPQGYFKIVIDLKD
TvERV-K1  QWPLPKEKLEAALMLVQQQ-LTAGHIEP--SNSPWNTPIFVIKKKSGA------WRLLHDLRAVNKTMIPMGSLQPGLPAPVA---IPAGFQKVVIDLKD
pTvERV-K2 QWPLPKEKLEAANMLVQQQ-LTAGHIEP--SNSPWNTPIFVIKKKSGA------WRLLHDLRAVNKTMIPMGSLQPGLPAPVA---IPAGFQKIVIDLKD
SMRV-H    QWPLTYEKTLAAIALVQEQ-LAAGHIEP--TNSPWNTPIFIIKKKSGS------WRLLQDLRAVNKVMVPMGALQPGLPSPVA---IPLNYHKIVIDLKD
JSRV      QWPLTQEKLSAAQQLVQEQ-LRLGHIEP--STSAWNSPIFVIKKKSGK------WRLLQDLRKVNETMMHMGALQPGLPTPSA---IPDKSYIIVIDLKD
MMTV      QWPLKQEKLQALQQLVTEQ-LQLGHLEE--SNSPWNTPVFVIKKKSGK------WRLLQDLRAVNATMHDMGALQPGLPSPVA---VPKGWEIIIIDLQD
CHIAP     QWPLSSEKLEVVTRLIQEQEEQLGHLES--STSPWNSPIFIIKKKSGK------WRLLHDLRAINNQMRPLGPVQRGLPLLSA---LPQNWKLIIIDIKD
MIAP      QWHLSSEKLEAVIQLVEEQ-LKLGHIDP--STSPWNTPIFVIKKKSGK------WRLLHDLRPINEQMNLFGPVQRGLPVLSA---LPRGWNLIIIDIKD
HERV-K10  QWPLPKQKLEALHLLANEQ-LEKGHIEP--SFSPWNSPVFVIQKKSGK------WHTLTDLRAVNAVIQPMGPLQPGLPSPAM---IPKDWPLIIIDLKD
RSV       QWPLPEGKLVALTQLVEKE-LQLGHIEP--SLSCWNTPVFVIRKASGS------YRLLHDLRAVNAKLVPFGAVQQGAPVLSA---LPRGWPLMVLDLKD
LDV       QWPLTAQKLDAVQNIIQDL-LKDGRIIP--SRSQWNSPIFVIQKKDKS-----KFRMLHDLRAVNALIKDWGALQPGTPWPGA---IPSEWPVIAMDISD
Visna     QWPLTQEKLEGLKEIVDRL-EKEGKVGRAPPHWTCNTPIFCIKKKSGK------WRMLIDFRELNKQTEDLAEAQLGLPHPGG---LQRKKHVTILDIGD
OMVV      QWPLTQEKLEGLKEIVDKL-EKEGKVGRAPPHWTCNTPIFCIKKKSGK------WRMLIDFRELNKQTEDLAEAQLGLPHPGG---LQKKKHVTILDIGD
CAEV      QWPLTEEKLKGLTEIIDKL-VEEGKLGKAPPHWTCNTPIFCIKKKSGK------WRMLIDFRELNKQTEDLTEAQLGLPHPGG---LQKKKHVTILDIGD
FIV       QWPLTNEKIEALTEIVERL-EREGKVKRADPNNPWNTPVFAIKKKSGK------WRMLIDFRELNKLTEKGAEVQLGLPHPAG---LQIKKQVTVLDIGD
EIAV      QWPLTKEKLEGAKEIVQRL-LSEGKISEASDNNPYNSPIFVIKKRSGK------WRLLQDLRELNKTVQVGTEISRGLPHPGG---LIKCKHMTVLDIGD
SIVmac    QWPLSKEKIVALREICEKM-EKDGQLEEAPPTNPYNTPTFAIKKKDKN-----KWRMLIDFRELNRVTQDFTEVQLGIPHPAG---LAKRKRITVLDIGD
HIV-2     QWPLTKEKIEALKEICEKM-EKEGQLEEAPPTNPYNTPTFAIKKKDKN-----KWRMLIDFRELNKVTQDFTEIQLGIPHPAG---LAKKRRITVLDVGD
SIVmnd    QWPLSKEKIVGLQKICDRL-EEEGKISRVDPGNNYNTPIFAIKKKDKN-----EWRKLIDFRELNKLTQDFHELQLGIPHPAG---IKKCKRITVLDIGD
HIV-1     QWPLTEEKIKALVEICTEM-EKEGKISKIGPENPYNTPVFAIKKKDST-----KWRKLVDFRELNKRTQDFWEVQLGIPHPAG---LKKKKSVTVLDVGD
SIVagm    QWPLSKEKIEALQEICSQL-EQEGKISRVGGENAYNTPIFCIKKKDKS-----QWRMLVDFRELNKATQDFFEVQLGIPHPAG---LRKMRQITVLDVGD
BIV       QWPLTKEKYQALKEIVKDL-LAEGKISEAAWDNPYNTPVFVIKKKGTG-----RWRMLMDFRELNKITVKGQEFSTGLPYPPG---IKECEHLTAIDIKD
HTLV-1    QPPLNPERLQALQHLVRKA-LEAGHIEP--YTGPGNNPVFPVKKANGT------WRFIHDLRATNSLTIDLSSSSPGPPDLSS--LPTTLAHLQTIDLKD
HTLV-2    PISFKPERLQALNDLVSKA-LEAGHIEP--YSGPGNNPVFPVKKPNGK------WRFIHDLRATNAITTTLTSPSPGPPDLTS--LPTALPHLQTIDLTD
BLV       QPPLNLERLQALQDLVHRS-LEAGYISP--WDGPGNNPVFPVRKPNGT------WRFVHDLRATNALTKPIPALSPGPPDLTA--IPTHLPHIICLDLKD
MoMLV     QYPMSQEARLGIKPHIQRL-LDQGILVP--CQSPWNTPLLPVKKPGTN-----DYRPVQDLREVNKRVEDIHPTVPNPYNLLS-GLPPSHQWYTVLDLKD
FeLV      QYPMPHEAYQGIKPHIRRM-LDQGILKP--CQSPWNTPLLPVKKPGTK-----DYRPVQDLREVNKRVEDIHPTVPNPYNLLS-TLPPSHPWYTVLDLKD
GaLV      QYPMSKEAREGIRPHIQKF-LDLGVLVP--CRSPWNTPLLPVKKPGTN-----DYRPVQDLREINKRVQDIHPTVPNPYNLLS-SLPPSYTWYSVLDLKD
BaEV      QYPMSLEAHMGIRQHIIKF-LELGVLRP--CRSPWNTPLLPVKKPGTQ-----DYRPVQDLREINKRTVDIHPTVPNPYNLLS-TLKPDYSWYTVLDLKD
MuRRS     FT-EGPRVTWESPTSVVLA-LQLGILVP--CQSPWNTPLLSVKKPGTS-----DY-PVQDLREVNKRVQDIHPTVPNPYNLLS-SLPPEQKWYTVLDLKD
HERV-E    QYPVLREALEGIQVHLKCL-RTFRIIVP--CQSPWNTPLLPVPKPGTK-----DYRPVQDLRLVNQATVTLHPTVPNLYTLLG-LLPAEDSWFTCLDLKD
WDSV      QYPLPKDKTEGLRPLISSL-ENQGILIK--CHSPCNTPIFPIKKAGRD-----EYRMIHDLRAINNIVAPLTAVVASPTTVLS-NLAPSLHWFTVIDLSN
SFV       QYPINPKAKPSIQIVIDDL-LKQGVLIQ--QNSTMNTPVYPVPKPDGK------WRMVLDYREVNKTIPLIAAQNQHSAGILS--SIYRGKYKTTLDLTN
HSRV      QYPINPKAKASIQIVIDDL-LKQGVLTP--QNSTMNTPVYPVPKPDGR------WRMVLDYREVNKTIPLTAAQNQHSAGILA--TIVRQKYKTTLDLAN
FeSFV     QYHINPKAKPDIQIVINDL-LKQGVLIQ--KESTMNTPVYPVPKPNGR------WRMVLDYRAVNKVTPLIAVQNQHSYGIIG--SLFKGKYKTTIDLSN
HERV-L    KHYHHPGGIAEISATIKDL-KNAGVVIP--ITSLFNSPFWPVQKTDGS------WRMRVDYHKLNQVVTPIAAAIQDVVSLLKQINTSPGTLYAAIDLAN
MuERV-L   KQYCIPGEIAEITATIKDL-KDAGVVVP--TTSPFNSPIWPVQKTDGS------WRMTVDYRKLNQVVTPIAAAVPDVVSLLEQINTSPGTWYAAIDLAN
```

```
MPMV       CFFSIPLHPSDQKRFAFSLPSTNFKEPMQRFQWKVLPQGMANSPTLCQKYVATAIHKVRHAWKQMYIIHYMDDILIAGKDGQQ-VLQCFDQLKQELTAAG
SRV-1      CFFSIPLHPSDQKRFAFSLPSTNFKEPMQRFQWKVLPQRMANSPTLCQKYVATAIHKVRHAWKQMYIIHYMDDILIAGKDGQQ-VLQCFDQLKQELTIAG
SRV-2      CFFTIPLQPVDQKRFAFSLPSTNFKQPMKRYQWKVLPQGMANSPTLCQKYVAAAIEPVRKSWAQMYIIHYMDDILIAGKLGEQ-VLQCFAQLKQALTTTG
TvERV-K1   CFFSIPLHPDDSKRFAFTVPVTNCVGPSPRFQWKVLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDDILIAGPCEQE-LLRITTLLITALKDRG
pTvERV-K2  CFFSIPLHPDDSKRFAFTVPVTNCVGPSPRFQWKVLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDDILIAGPCEQE-LLRITTLLITALKDRG
SMRV-H     CFFTIPLHPEDRPYFAFSVPQINFQSPMPRYQWKVLPQGMANSPTLCQKFVAAAIAPVRSQWPEAYILHYMDDILLACDSAEA-AKACYAHIISCLTSYG
JSRV       CFYTIPLAPQDCKRFAFSLPSVNFKEPMQRYQWRVLPQGMTNSPTLCQKFVATAIAPVRQRFPQLYLVHYMDDILLAHTDEHL-LYQAFSILKQHLSLNG
MMTV       CFFNIKLHPEDCKRFAFSVPSPNFKRPYQRFQWKVLPQGMKNSPTLCQKFVDKAILTVRDKYQDSYIVHYMDDILLAHPSRSI-VDEILTSMIQALNKHG
CHIAP      CFFSIPLFPRDRQRFAFTVPSLNHMEPDKRYQWRVLPQGMANSPTICQLYVQKALEPVRKQFTSMIMIHYMDDILICHRKIEV-LQQAFPMLVAELKQWG
MIAP       CFFSIPLCPRDRPRFAFTIPSINSDEPDNRYQWKVLPQGMSNSPTMCQLYVQEALLPVREQFPSLILLLYMDDILLCHKELTM-LQKAYPFLLKTLSQWG
HERV-K10   CFFTIPLAEQDCEKFAFTIPAINNKEPATRFQWKVLPQGMLNSPTICQTFVGRALQPVREKFSDCYIIHYIDDILCAAETKDK-LIDCYTFLQAEVANAG
RSV        CFFSIPLAEQDREAFAFTLPSVNNQAPARRFQWKVLPQGMTCSPTICQLVVGQVLEPLRLKHPSLCMLHYMDDLLLAASSHDG-LEAAGEEVISTLERAG
LDV        CFFSIPLAERDSERFAFTIPSPNLREPAKRYQWTVLPQGMKNSPYICQQVVAEVIRPIRERFRDAVIIHYMDDILIAAAEERQ-TEVIFEAVKTTCHEKG
Visna      AYFTIPLYEPYRQYTCFTMLSPNNLGPCVRYYWKVLPQGWKLSPAVYQFTMQKILRGWIEEHPMIQFGIYMDDIYIGSDLGLEEHRGIVNELASYIAQYG
OMVV       AYFTIPLYEPYRPYTCFTMLSPNNLGPCTRYYWKVLPQGWKLSPSVYQFTMQEILRDWIAKHPMIQFGIYMDDIYIGSDLDIMKHREIVEELASYIAQYG
CAEV       AYFTIPLYEPYREYTCFTLLSPNNLGPCKRYYWKVLPQGWKLSPSVYQFTMQEILEDWIQQHPEIQFGIYMDDIYIGSDLEIKKHREIVKDLANYIAQYG
FIV        AYFTIPLDPDYAPYTAFTLPRKNNAGPGRRFVWCSLPQGWILSPLIYQSTLDNIIQPFIRQNPQLDIYQYMDDIYIGSNLSKKEHKEKVEELRKLLLWWG
EIAV       AYFTIPLDPEFRPYTAFTIPSINHQEPDKRYVWNCLPQGFVLSPYIYQKTLQEILQPFRERYPEVQLYQYMDDLFVGSNGSKKQHKELIIELRAILLEEG
SIVmac     AYFSIPLDEEFRQYTAFTLPSVNNAEPGKRYIYKVLPQGWKGSPAIFQYTMRHVLEPFRKANPDVTLVQYMDDILIASDRTDLEHDRVVLQLKELLNSIG
HIV-2      AYFSIPLHEDFRPYTAFTLPSVNNAEPGKRYIYKVLPQGWKGSPAIFQHTMRQVLEPFRKANKDVIIIQYMDDILIASDRTDLEHDRVVLQLKELLNGLG
SIVmnd     AYFSIPLDPDYRPYTAFTVPSVNNQAPGKRYMYNVLPQGWKGSPCIFQGTVASLLEVFRKNHPTVQLYQYMDDLFVGSDYTAEEHEKAIVELRALLMTWN
HIV-1      AYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWG
SIVagm     AYYSIPLDPNFRKYTAFTIPTVNNQGPGIRYQFNCLPQGWKGSPTIFQNTAASILEEIKRNLPALTIVQYMDDLWVGSQENEHTHDKLVEQLRTKLQAWG
BIV        AYFTIPLHEDFRPFTAFSVVPVNREGPIERFQWNVLPQGWVCSPAIYQTTTQKIIENIKKSHPDVMLYQYMDDLLIGSNRDD--HKQIVQEIRDKLGSYG
HTLV-1     AFFQIPLPKQFQPYFAFTVPQQCNYGPGTRYAWRVLPQGFKNSPTLFEMQLAHILQPIRQAFPQCTILQYMDDILLASPSHAD-LQLLSEATMASLISHG
HTLV-2     AFFQIPLPKQYQPYFAFTIPQPCNYGPGTRYAWTVLPQGFKNSPTLFEQQLAAVLNPMRKMFPTSTIVQYMDDILLASPTNEE-LQQLSQLTLQALTTHG
BLV        AFFQIPVEDRFRSYFAFTLPTPGGLQPHRRFAWRVLPQGFINSPALFERALQEPLRQVSAAFSQSLLVSYMDDILIASPTEEQ-RSQCYQALAARLRDLG
MoMLV      AFFCLRLHPTSQPLFAFEWRDPEM-GISGQLTWTRLPQGFKNSPTLFDEALHRDLADFRIQHPDLILLQYVDDLLLAATSELD-CQQGTRALLQTLGNLG
FeLV       AFFCLRLHSESQLLFAFEWRDPEI-GLSGQLTWTRLPQGFKNSPTLFDEALHSDLADFRVRYPALVLLQYVDDLLLAAATRTE-CLEGTKALLETLGNKG
GaLV       AFFCLRLHPNSQPLFAFEWKDPEK-GNTGQLTWTRLPQGFKNSPTLFDEALHRDLAPFRALNPQVVLLQYVDDLLVAAPTYED-CKKGTQKLLQELSKLG
BaEV       AFFCLPLAPQSQELFAFEWKDPER-GISGQLTWTRLPQGFKNSPTLFDEALHRDLTDFRTQHPEVTLLQYVDDLLLAAPTKKA-CTQGTRHLLQELGEKG
MuRRS      AFFCLKLHPSSQPIFAFEWRDPDT-GQTGQLTWMW-PQEFKNSPTLFDEALHRDLASFRAENSQVTLLQYVDDLLLAATTEEN-CWQGTKRLLAELGELG
HERV-E     AFFSIRLAPERQKLFAFQWEDPES-GVTTQYTWTQLPQRFKNSPTIFGEALARDLQKFPTRDLGCVLLQYVDDLLLGHPTAVG-WPREQMLYSGTWRTVG
WDSV       AFFSVPIHKDSQYLFAFTFE-------GHQYTWTVLPQGFIHSPTLFSQALYQSLHKIKFKI-SSEICIYMDDVLIASKDRDT-NLKDTAVMLQHLASEG
SFV        GFWAHPITPESYWLTAFTWQ-------GKQYCWTRLPQGFLNSPALFTADVVDLLKEIPN------VQAYVDDIYISHDDPQE-HLEQLEKIFSILLNAG
HSRV       GFWAHPITPESYWLTAFTWQ-------GKQYCWTRLPQGFLNSPALFTADVVDLLKEIPN------VQVYVDDIYLSHDDPKE-HVQQLEKVFQILLQAG
FeSFV      GFWAHPIVPEDYWITAFTWQ-------GKQYCWTVLPQGFLNSPGLFTGDVVDLLQGIPN------VEVYVDDVYISHDSEKE-HLEYLDILFNRLKEAG
HERV-L     AFYPFHKPHQKQFAFSWQGQ---------QCIFTVLLQGYINSPALCHNLIQSDLDHFSLPQDIT-LVHYIDGIMLIGSNEQE-VANTLDLLVRHLHAIG
MuERV-L    AFFSVPVHKDHQKQIAFSWQ-------GQQYTFTVLPQVYINSPALCHNLVRRDLDRLDLPQSIT-LVHYIDDIMLVGPSEQE-VATTLDSLVTHMRIRG

MPMV       LHIAPEKVQL-QDPYTYLGFEL
SRV-1      LHIAPEKIQL-QDPYTYLGFEL
SRV-2      LQIAPEKVQL-QDPYTYLGFQI
TvERV-K1   LYISPDKVQL-ASPFFFLGFEL
pTvERV-K2  LYISPDKVQL-ASPFFFLGFEL
SMRV-H     LKIAPDKVQV-SEPFSYLGFEL
JSRV       LVIADEKIQT-HFPYNYLGFSL
MMTV       LVVSTEKIQK-YDNLKYLGTHI
CHIAP      LEIASEKVQV-SDTGLFLGSVI
MIAP       LQIATEKVQI-SDTGQFLGSVV
HERV-K10   LAIASDKIQT-STPFHYLGMQI
RSV        FTISPDKVQR-EPGVQYLGYKL
LDV        LKINEAKTQR-APEVSYLGWRV
Visna      FMLPEDKRQE-GYPAKWLGFEL
OMVV       FMLPEEKRQE-GYPAKWLGFEL
CAEV       FTLPEEKRQK-GYPAKWLGFEL
FIV        FETPEDKLQE-EPPYTWMGYEL
EIAV       FETPDDKLQE-VPPYSWLGYQL
SIVmac     FSSPEEKFQK-DPPFQWMGYEL
HIV-2      FSTPDEKFQK-DPPYHWMGYEL
SIVmnd     LETPEKKYQK-EPPFHWMGYEL
HIV-1      LTTPDKKHQK-EPPFLWMGYEL
SIVagm     LETPEKKMQK-EPPYEWMGYKL
BIV        FKTPDEKVQ--EERVKWIGFEL
HTLV-1     LPVSENKTQQTPGTIKFLGQII
HTLV-2     LPISQEKTQQTPGQIRFLGQVI
BLV        FQVASEKTRQTPSPVPFLGQMV
MoMLV      YRASAKKAQICQKQVKYLGYLL
FeLV       YRASAKKAQICLQEVTYLGYSL
GaLV       YRVSAKKAQLCQREVTYLGYLL
BaEV       YRASAKKAQICQTKVTYLGYIL
MuRRS      YRASAKKAQLCQMELVYLRYTL
HERV-E     IRCPRKKAQICRQQVCYLGFTI
WDSV       HKVSKKKLQLCQQEVVYLGQLL
SFV        YVVSLKKSEIAQREVEFLGFNI
HSRV       YVVSLKKSEIGQKTVEFLGFNI
FeSFV      YIVSLKKSNIANSIVDFLGFQI
HERV-L     WEINPNKIHAPSTSVKFLGSSG
MuERV-L    WEINPTKIQGPSTSVKFLGVQW
```

Appendix 7 – PR and RT alignment

The amino acid sequences below were manually aligned using the alignment in Herniou *et al.* (1998) as a guide. The overlined sequences were used to construct a phylogenetic tree using CLUSTAL X and the BLOSUM weight matrix, as described in Chapter 2 – Materials and Methods. The resulting tree is shown in FIG. 5.12. Retrovirus names, accession numbers, and references are shown in Appendix 1.

```
pTvERV-K2       TIRPFVVPCLPVNLWGRDILSQMGVIMCSP-PVWIDQWPLPKEKLEAANMLVQQQLTAGHIEPSNSPWNTPIFVIKKKSGA-WRLLHDLRAVNKTMIPMG
TvERV-K1        TIRPFVVPCLPVNLWGRDILSQMGVIMCSP-PVWIDQWPLPKEKLEAALMLVQQQLTAGHIEPSNSPWNTPIFVIKKKSGA-WRLLHDLRAVNKTMIPMG
MPMV            LIKPFVIPNLPVNLWGRDLLSQMKIMMCSP-PVWVDQWPLTNDKLAAAQQLVQEQLEAGHITESSSPWNTPIFVIKKKSGK-WRLLQDLRAVNATMVLMG
SRV-1           LIKPFVIPNLPVNLWGRDLLSQMKIMMCSP-PVWVDQWPLTSEKLAAAQQLVQEQLEAGHITESNSPWNTPIFVIKKKSGK-WRLLQDLRAVNATMVLMG
SERV 23.1       LIKPFVIPYLPVNLWGRDLLAQMKIIMCSP-PVWVDQWPLLNDKLSAAQQLLQEQLEAGHIIESNSPWNTPIFVIKKKSGK-WRLLQDLRAVNITMVLMG
SERV 25.2       LIKPFIIPHLPVNLWGRDLLSQMKIIMCSP-PVWVDQWPLLNDKLSAAQQLVQEQLLAGHITESNSPWNTPIFVIKKKSGK-WRLLQYLRAVNITMVLMG
SRV-2           LIKPFVIPNLPVNLWGRDLLSQMKIMMCSP-PVWVDQWPLTQEKLAAAQQLVQEQLQAGHIIESNSPWNTPIFVIKKKSGK-WRLLQDLRAVNATMVLMG
SMRV-H          HITPYVLPNLPVNLWGRDILSQMKLVMCSP-PVWVDQWPLTYEKTLAAIALVQEQLAAGHIEPTNSPWNTPIFIIKKKSGS-WRLLQDLRAVNKVMVPMG
JSRV            QFKPYILPYLPVNLWGRDILSKMGVYLYSP-PVWVDQWPLTQEKLSAAQQLVQEQLRLGHIEPSTSAWNSPIFVIKKKSGK-WRLLQDLRKVNETMMHMG
ENTV            QFKPYILPHLPVNLWGRDILSKMGVYLYSP-PVWVDQWPLTQEKLSAAQQLVQEQLRLGHIEPSTSAWNSPIFVIKKKSGK-WRLLQDLRKVNETMMHMG
MMTV            IIHPFVIPTLPFTLWGRDIMKDIKVRLMTD-PVWLNQWPLKQEKLQALQQLVTEQLQLGHLEESNSPWNTPVFVIKKKSGK-WRLLQDLRAVNATMHDMG
RVsfdl          IFRPLVIAGLRTSLWGRDLLKTLWTTLNIE-PIWVDQWFLSNEKLTTLKSIIMEQEKLGHIEHSFSAYNSPVFVIRKKEQ--WRML-DLRAVNASMQPMG
HERV-K10        STVQPMITSIPLNLWGRDLLQQWGAEITMP-PVWVNQWPLPKQKLEALHLLANEQLEKGHIEPSFSPWNSPVFVIQKKSGK-WHTLTDLRAVNAVIQPMG
HERV-K-T47D     ELVLFTLYPIPVNLWGRDLLQQWGAQISFP-PVWVDQWPLPKNKLEALHNLVLEQLELGHIEESFSPWNSLVFVIQKKSGK-QRMLTHLRAVNAVLQPLG
Rvbowerbird     AQVKPYVMNTTIWLLGRDALSQMGFCLTNE-PIWIDQWPLSKEKLAHVEELVEDQLRRGHVIPSTSPWNTPIFAIPKKSGK-WRLLHDLRAINSVMQEMG
HRV-5           GNVQPYVSALPITLWGRDILEQLGLTLTNE-PKWIPQWPLTQEKLAAVNDIVLQQLEAGHLQPSTSPWNTPIFVIKKKSGK-YRLLHDLRAVNQQMQPMG
CHIAP           GRFTPYVLPLPVNLWGRDVLQAMGMTLTNE-PVWVSQWPLSSEKLEVVTRLIQEQEQLGHLESSTSPWNSPIFIIKKKSGK-WRLLHDLRAINNQMRPLG
SHIAP           GCFTPYVLPLPVNLWGRDVLQAMGMTLTNE-PLWVSQWPLSSEKLEAVTRLVQEQERLGHLEPSTSPWNTPIFVIKKKSGK-WRLLHDLRAINNQMHLFG
MIAP            GKFIPYVLPLPVDLWGRDIMQHLGLILSDN-PVWVPQWHLSSEKLEAVIQLVEEQLKLGHIDPSTSPWNTPIFVIKKKSGK-WRLLHDLRPINEQMNLFG
RSV             LLLFPAVAMVRGSILGRDCLQGLGLRLTNL-PVWIDQWPLPEGKLVALTQLVEKELQLGHIEPSLSCWNTPVFVIRKASGS-YRLLHDLRAVNAKLVPFG
LDV             IRVSPYIAAIGFNILGREALAQLHCVVSNL-PVWIEQWPLTAQKLDAVQNIIQDLLKDGRIIPSRSQWNSPIFVIQKKDKSKFRMLHDLRAVNALIKDWG
HTLV-I          LTSCLVDTKNNWAIIGRDALQQCQGVLYLP-PPEISQFPLNPERLQALQHLVRKALEAGHIEPYTGPGNNPVFPVKKANGT-WRFIHDLRATNSLTIDLS
HTLV-II         LSSCLLDTHNKWTIIGRDALQQCQGLLYLP-TSPSGPISFKPERLQALNDLVSKALEAGHIEPYSGPGNNPVFPVKKPNGK-WRFIHDLRATNAITTTLT
BLV             IPKILVDTFDKWQILGRDVLSRLQASISIP-PPEVPQFPLNLERLQALQDLVHRSLEAGYISPWDGPGNNPVFPVRKPNGT-WRFVHDLRATNALTKPIP

pTvERV-K2       SLQPGLPAPVAIP-AGFQKIVIDLKDCFFSIPLHPDDSKRFAFTVPVTNCVGPSPRFQWKVLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
TvERV-K1        SLQPGLPAPVAIP-AGFQKVVIDLKDCFFSIPLHPDDSKRFAFTVPVTNCVGPSPRFQWKVLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYIIHYMDD
MPMV            ALQPGLPSPVAIP-QGYLKIIIDLKDCFFSIPLHPSDQKRFAFSLPSTNFKEPMQRFQWKVLPQGMANSPTLCQKYVATAIHKVRHAWKQMYIIHYMDD
SRV-1           ALQPGLPSPVAIP-QGYLKIIIDLKDCFFSIPLHPSDQKRFAFSLPSTNFKEPMQRFQWKVLPQRMANSPTLCQKYVATAIHKVRHAWKQMYIIHYMDD
SERV23.1        ALQPGLPSPVAIP-QKYFKIIIDLKDCFFTIPLHPADQKRFAFSLPSTNFRQPMKRYQWKVLPQGMANSPTLCQKYVAAAIEPVRKTWTQMYIIHYMDN
SERV25.2        ALQPGLPSPVAIP-QKYFKIIIDLKDCFFTIPLHPADQKRFAFSLPSTNFRQPMKRYQWRVLPQGMANSPTLCQKYVAAAIEPVRKIWTQMYIIHYMDD
SRV-2           ALQPGLPSPVAIP-QGYFKIVIDLKDCFFTIPLQPVDQKRFAFSLPSTNFKQPMKRYQWKVLPQGMANSPTLCQKYVAAAIEPVRKSWAQMYIIHYMDD
SMRV-H          ALQPGLPSPVAIP-LNYHKIVIDLKDCFFTIPLHPEDRPYFAFSVPQINFQSPMPRYQWKVLPQGMANSPTLCQKFVAAAIAPVRSQWPEAYILHYMDD
JSRV            ALQPGLPTPSAIP-DKSYIIVIDLKDCFYTIPLAPQDCKRFAFSLPSVNFKEPMQRYQWRVLPQGMTNSPTLCQKFVATAIAPVRQRFPQLYLVHYMDD
ENTV            ALQPGLPTPSAIP-DKSYIIVIDLKDCFYTIPLAPQDCKRFAFSVPSVNFKEPMKRYQWKVLPQGMTNSPTLCQKLVATAIAPVRQRFPQIYLVHYMDD
MMTV            ALQPGLPSPVAVP-KGWEIIIIDLQDCFFNIKLHPEDCKRFAFSVPSPNFKRPYQRFQWKVLPQGMKNSPTLCQKFVDKAILTVRDKYQDSYIVHYMDD
RVsfdunnartl    ALQPGLLSPNMVP-KEYHMVIIDIKDCFYSIPLHSADREKFAFSVPAVNLQAPAPRWQWKVLPQEMANSPILCQVNVDKVLAPVRNLYPNVYMLHYMDD
HERV-K10        PLQPGLPSPAMIP-KDWPLIIIDLKDCFFTIPLAEQDCEKFAFTIPAINNKEPATRFQWKVLPQGMLNSPTICQTFVGRALQPVREKFSDCYIIHYIDD
HERV-K-T47D     TLQSGLPSRSMLA-EYWPLILIDLKDCFFNIPLASQDFEKFAFMVPSLNNVAQATCYYWKVLPQGMLNSPTICQYFVGRVLQPVRDQFPRCYIVHYMDD
RVbowerbird     HLQRGLPSQSMLP-VNWQLLVVDLRDCFFTIPLHEDDSEKFAFSVPSINKSEPTKRYQWVVLPQGMHNSPIMCQLYVAWALAPLRKQYPQYLIYHYMDD
HRV-5           ALQPGLPVPTMIP-KHWPLIVLDLKDCFFSIPLHEQDIQRFAFTVPSINHQGPDKRYEWKVLPQGMTNSPAICQLYVDQAVEPVRQQCPKVQILHYMDD
CHIAP           PVQRGLPLLSALP-QNWKLIIIDIKDCFFSIPLFPRDRQRFAFTVPSLNHMEPDKRYQWRVLPQGMANSPTICQLYVQKALEPVRKQFTSMIMIHYMDD
SHIAP           PVQRGLPLLSALP-QDWKLIIIDIKDCFFSIPLYPRDRPRFAFTIPSLNHMEPDKRFQWKVLPQGMANSPTICQLYVQEALEPIRKQFTSLIVIHYMDD
MIAP            PVQRGLPVLSALP-RGWNLIIIDIKDCFFSIPLCPRDRPRFAFTIPSINSDEPDNRYQWKVLPQGMSNSPTMCQLYVQEALLPVREQFPSLILLLYMDD
RSV             AVQQGAPVLSALP-RGWPLMVLDLKDCFFSIPLAEQDREAFAFTLPSVNNQAPARRFQWKVLPQGMTCSPTICQLVVGQVLEPLRLKHPSLCMLHYMDD
LDV             ALQPGTPWPGAIP-SEWPVIAMDISDCFFSIPLAERDSERFAFTIPSPNLREPAKRYQWTVLPQGMKNSPYICQQVVAEVIRPIRERFRDAVIIHYMDD
HTLV-I          SSSPGPPDLSSLPTTLAHLQTIDLKDAFFQIPLPKQFQPYFAFTVPQQCNYGPGTRYAWRVLPQGFKNSPTLFEMQLAHILQPIRQAFPQCTILQYMDD
HTLV-II         SPSPGPPDLTSLPTALPHLQTIDLTDAFFQIPLPKQYQPYFAFTIPQPCNYGPGTRYAWTVLPQGFKNSPTLFEQQLAAVLNPMRKMFPTSTIVQYMDD
BLV             ALSPGPPDLTAIPTHLPHIICLDLKDAFFQIPVEDRFRSYFAFTLPTPGGLQPHRRFAWRVLPQGFINSPALFERALQEPLRQVSAAFSQSLLVSYMDD
```

Appendix 8 – Pro and Pol alignment

The amino acid sequences below were aligned using CLUSTAL X as described in Chapter 2 – Materials and Methods. The Pro and Pol sequences were aligned separately, and were then concatenated. The overlined sequences were used to construct a neighbour-joining tree using CLUSTAL X and the BLOSUM weight matrix, and a maximum parsimony tree using PAUP*4.0b2, as described in Chapter 2 – Materials and Methods. The resulting trees are shown in FIG. 5.13. Retrovirus names, accession numbers, and references are shown in Appendix 1.

```
          → Pro
MMTV      GSAGLDLSSQKDLILSLEDGVSLVPTLVKGT-LPEGTTGLIIGRSSN-YKKGLEVLPGVIDSDFQGEIKVMVKAAKNAVIIHKGERIAQLLLLPYLKLPN
HERV-K10  GSAAVDLCTIQAVSLLPGEPPQKIPTGVYGP-LPEGTVGLILIGRSSL-NLKGVQIHTSVVDSDYKGEIQLVISSSIPWSASPRD-RIAQLLLLPYIKGGN
JSRV      GSAGLDLCATSYTVLTPEMGVQTLATGVFGP-LPPGTVGLLLGRSSA-SLKGILIHPGVIDSDYTGEIKILASAPNKIIVINAGQRIAQLLLVPLVIQGK
ENTV      GSAGLDLCATSYTVLTPEMGVQTLATGVFGP-LPPGTAGLLLGRSSA-SLKGILIHPGVIDSDYTGEIKILASAPNKIIVINAGQRIAQLLLVPLVIQGK
TvERV-K1  GSAGVDLCATAAAILRPEDGPQILPTGVFGP-PPKGSFFLIIGRAST-TLQGITIHPTIVDNDYEGEIKIIASALKDTVRIQPNQRIAQALPLPMNTTFP
pTvERV-K2 GSAGVDLCATAAAILRPEDGPQILPTGVFGP-PPRGSFFLIIGRAST-TLQGITIHPTIVDNDYAGEIKIIASALKDTVRIQPNQRIAQALPLPMNTTFP
SMRV-H    GSAGLDLCTTTDTILTTQNSPLTLPVGIYGP-LPPQTFGLILAEPAL-PSKGIQVLPGILDNDFEGEIHIILSTTKDLVTIPKGTRLAQIVILPLQQINS
MPMV      GSAGLDLCSTSHTVLTPEMGPQALSTGIYGP-LPPNTFGLILGRSSI-TMKGLQVYPGVIDNDYTGEIKIMAKAVNNIVTVSQGNRIAQLILLPLIETDN
SRV-1     GSAGLDLSSTSHTVLTPEMGPQALSTGIYGP-LPPNTFGLILGRSSI-TIKGLQVYPGVIDNDYTGEIKIMAKAVNNIVTVPQGNRIAQLILLPLIETDN
SRV-2     GSAGLDLCSTTHTVLTPEMGPQTLATGVYGP-LPPNTFGLILGRGST-TVKGLQIYPGVIDNDYTGEFKIMARAISSIITIPQGERIAQLVLLPLLRTAH
SERV 23.1 GSAGLDLSSTSHTVLTPEMGPQTLNTGIYGP-LPPNTFGLLLGRSSV-TMRGLQVLPGVIDNDYEGEIKIMARAIDNIITVPQGVRIAQLILLPLVKTDN
SERV 25.2 GSAGLDLSSTSHTVLTPEMGPQTLNTGIYGP-LPPDTFGLLLGRSSV-TMKGLQVLPGVINNDYEGEIKIMARAINNIITVPQGVRIAQLVLLPLVKTDN
CHIAP     GHSGVDLCAASDFLLMPQMNVQPVPIQSPGP-LPPATIGLILGRGSL-TLQGLIVYPGIVDPYHKEEFQVLCSSPSGVFSIKQGDKIAQLVLLPSPGDRE
SHIAP     RNLRNGLACRSDFLLMPQMNVQPVPVHSPGP-LPPATIGLILGRGSL-TLQGLIIYPGIVDPYHKEEIQVLCSSPRGVFSIKQGDRIAQLVLPPSLGDGE
MIAP      KTHKVDLRAASDFLLMPQMSIQPVPVE-PIPSLPLGTMGLILGRGSASTLQGLVVHPELWIVNIPQKYQVLCSSPKGVFSISKGDRIPQLLLLLPDNTRE

MMTV      PVIKEERGSEGFGST-SHVHWVQEISDSRPMLHIYLNGRRFLGLLDTGADKTCIAGRDWPANWPIHQTESSLQGLGMACGVARSSQPLRWQHED-KSGII
HERV-K10  SEIKRIGGLVSTDPTGKAAYWASQVSENRPVCKAIIQGKQFEGLVDTGADVSIIALNQWPKNWPKQKAVTGLVGIGTASEVYQSMEILHCLGPDNQESTV
JSRV      -TINRDRQDKGFGS--SDAYWVQNVTEARPELELRINANFFRGVLDTGADISVISDKYWPTTWPKQMAISTLQGIGQTTNPEQSSSLLTWKDKDGHTGQF
ENTV      -TINRDRQDKGFGS--SDAYWVQNVTEAXPELELRINXKLFRGVLDTGADISVISDKYWPTTWPKQMAISTLQGIGQTTNPEQSSSLLTWKDKDGHTGQF
TvERV-K1  -ASKNKRGNSTPGS--SDIYWAQALSQERPTLMLKIQGKPFMGLLDSGADSTVISESHWPSTWPLQPSLTHLQGIGQSSNTMQSSQLLQWEDREGNRGTI
pTvERV-K2 -ASKNKRGNSTPGS--SDIYWAQALSQERPTLMLKIQGKSFMGLLDSGADSTVISEAHWPPAWPLQPSLTHLQGIGQSSNTMQSTQLLQWEDREGNRGTI
SMRV-H    NFHKPYRGASAPGS--SDVYWVQQISQQRPTLKLKLNGKLFSGILDTGADATVISYTHWPRNWPLTTVATHLRGIGQATNPQQSAQMLKWEDSEGNNGHI
MPMV      KVQQPYRGQGSFGS--SDIYWVQPITCQKPSLTLWLDDKMFTGLIDTGADVTIIKLEDWPPNWPITDTLTNLRGIGQSNNPKQSSKYLTWRDKENNSGLI
SRV-1     KVQQPYRGQGSFGS--SDIYWVQPITCQKPSLTLWLDDKMFTGLIDTGADVTIIKLEDWPPNWPITDTLTNLRGIGQSNNPKQSSKYLTWRDKENNSGLI
SRV-2     KIQHPYRGDKNFGS--SDIFWVQPITHQKPSLVLWLDGKAFTGLIDTGADVTIIKQEDWPSHWPTTETLTNLRGIGQSNNPRQSSKYLTWKDKENNSGLI
SERV 23.1 NIQHSNRNIKGFGS--SDIYWVQPITNQKPSLTLWLDGKAFTGLIDTGADVTIIKQEDWPSHWPTTETLTHLRGIGQSSNPKQSSKYLTWTDKENNSGLI
SERV 25.2 NIQHSNRNAKGFGS--SNIYWVQPITNQKPSLTLWLDGKAFTGLIDTGADVTIIKQEDWPSHWPTTETLTHLRGIGQSSNPKQSSKYLTWTDKENNSGLI
CHIAP     -NCTSRKRALGSTGN-DSAYLAIPLDE-RPTMKLLVNGKEFEGITDTGADKSIISLHWWPKSWPTVTSSHSLQGLGYQSSPAVSAAALVWRSTEGRQGRF
SHIAP     -TYTLQKRAMGSSGS-DSAYLAIPLNE-RPTIKLLINGKEFEGIMDSGADKSIISLHWWPKSWPTVVSSHSLQGLGYQSSPAISASALTWRDAEGKQGCF
MIAP      KSAGPEIKKMGSSGN-DSAYLVVSLND-RPKLRLKINGKEFEGILDTGADKSIISTHWWPKAWPTTESSHSLQGLGYQSCPTISSVALTWESSEGQQGKF

                                  Pro ← → Pol
MMTV      HPFVIPTLPFTLWGRDIMKDIKVRLMTDSPD-QISWKSDQPVWLNQWPLKQEKLQALQQLVTEQLQLGHLEESNSPWNTPVFVIKKKSGKWRLLQDLRAV
HERV-K10  QPMITS-IPLNLWGRDLLQQWGAEITMPAPL-PLTWKTEKPVWVNQWPLPKQKLEALHLLANEQLEKGHIEPSFSPWNSPVFVIQKKSGKWHTLTDLRAV
JSRV      KPYILPYLPVNLWGRDILSKMGVYLYSPSPT-PIDWKSEEPVWVDQWPLTQEKLSAAQQLVQEQLRLGHIEPSTSAWNSPIFVIKKKSGKWRLLQDLRKV
ENTV      KPYILPHLPVNLWGRDILSKMGVYLYSPSPT-PIDWKSEEPVWVDQWPLTQEKLSAAQQLVQEQLRLGHIEPSTSAWNSPIFVIKKKSGKWRLLQDLRKV
TvERV-K1  RPFVVPCLPVNLWGRDILSQMGVIMCSPSSV-KITWRSETPVWIDQWPLPKEKLEAALMLVQQQLTAGHIEPSNSPWNTPIFVIKKKSGAWRLLHDLRAV
pTvERV-K2 RPFVVPCLPVNLWGRDILSQMGVIMCSPSSV-KITWRSETPVWIDQWPLPKEKLEAANMLVQQQLTAGHIEPSNSPWNTPIFVIKKKSGAWRLLHDLRAV
SMRV-H    TPYVLPNLPVNLWGRDILSQMKLVMCSPNDT-KISWKITDPVWVDQWPLTYEKTLAAIALVQEQLAAGHIEPTNSPWNTPIFIIKKKSGSWRLLQDLRAV
MPMV      KPFVIPNLPVNLWGRDLLSQMKIMMCSPNDI-PITWKSDEPVWVDQWPLTNDKLAAAQQLVQEQLEAGHITESSSPWNTPIFVIKKKSGKWRLLQDLRAV
SRV-1     KPFVIPNLPVNLWGRDLLSQMKIMMCSPSDI-PITWKSDEPVWVDQWPLTSEKLAAAQQLVQEQLEAGHITESNSPWNTPIFVIKKKSGKWRLLQDLRAV
SRV-2     KPFVIPNLPVNLWGRDLLSQMKIMMCSPNDI-PITWKSDEPVWVDQWPLTQEKLAAAQQLVQEQLQAGHIIESNSPWNTPIFVIKKKSGKWRLLQDLRAV
SERV 23.1 KPFVIPYLPVNLWGRDLLAQMKIIMCSPNDI-PITWKSDEPVWVDQWPLLNDKLSAAQQLLQEQLEAGHIIESNSPWNTPIFVIKKKSGKWRLLQDLRAV
SERV 25.2 KPFIIPHLPVNLWGRDLLSQMKIIMCSPNDI-PITWKSDEPVWVDQWPLLNDKLSAAQQLVQEQLLAGHITESNSPWNTPIFVIKKKSGKWRLLQYLRAV
CHIAP     TPYVLP-LPVNLWGRDVLQAMGMTLTNEYSP-PIPWNTEDPVWVSQWPLSSEKLEVVTRLIQEQEQLGHLESSTSPWNSPIFIIKKKSGKWRLLHDLRAI
SHIAP     TPYVLP-LPVNLWGRDVLQAMGMTLTNEYSP-PIPWKTEEPLWVSQWPLSSEKLEAVTRLVQEQERLGHLEPSTSPWNTPIFVIKKKSGKWRLLHDLRAI
MIAP      IPYVLP-LPVNLWGRDIMQHLGLILSNENAP-PIPWKTGDPVWVPQWHLSSEKLEAVIQLVEEQLKLGHIDPSTSPWNTPIFVIKKKSGKWRLLHDLRPI

MMTV      NATMHDMGALQPGLPSPVAVPKGWEIIIIDLQDCFFNIKLHPEDCKRFAFSVPSPNFKRPYQRFQWKVLPQGMKNSPTLCQKFVDKAILTVRDKYQDSYI
HERV-K10  NAVIQPMGPLQPGLPSPAMIPKDWPLIIIDLKDCFFTIPLAEQDCEKFAFTIPAINNKEPATRFQWKVLPQGMLNSPTICQTFVGRALQPVREKFSDCYI
JSRV      NETMMHMGALQPGLPTPSAIPDKSYIIVIDLKDCFYTIPLAPQDCKRFAFSLPSVNFKEPMQRYQWRVLPQGMTNSPTLCQKFVATAIAPVRQRFPQLYL
ENTV      NETMMHMGALQPGLPTPSAIPDKSYIIVIDLKDCFYTIPLAPQDCKRFAFSVPSVNFKEPMKRYQWKVLPQGMTNSPTLCQKLVATAIAPVRQRFPQIYL
TvERV-K1  NKTMIPMGSLQPGLPAPVAIPAGFQKVVIDLKDCFFSIPLHPDDSKRFAFTVPVTNCVGPSPRFQWKVLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYI
pTvERV-K2 NKTMIPMGSLQPGLPAPVAIPAGFQKIVIDLKDCFFSIPLHPDDSKRFAFTVPVTNCVGPSPRFQWKVLPQGMTNSPTLCQKYVAQTIDPFRLQFPQLYI
SMRV-H    NKVMVPMGALQPGLPSPVAIPLNYHKIVIDLKDCFFTIPLHPEDRPYFAFSVPQINFQSPMPRYQWKVLPQGMANSPTLCQKFVAAAIAPVRSQWPEAYI
MPMV      NATMVLMGALQPGLPSPVAIPQGYLKIIIDLKDCFFSIPLHPSDQKRFAFSLPSTNFKEPMQRFQWKVLPQGMANSPTLCQKYVATAIHKVRHAWKQMYI
SRV-1     NATMVLMGALQPGLPSPVAIPQGYLKIIIDLKDCFFSIPLHPSDQKRFAFSLPSTNFKEPMQRFQWKVLPQRMANSPTLCQKYVATAIHKVRHAWKQMYI
SRV-2     NATMVLMGALQPGLPSPVAIPQGYFKIVIDLKDCFFTIPLQPVDQKRFAFSLPSTNFKQPMKRYQWKVLPQGMANSPTLCQKYVAAAIEPVRKSWAQMYI
SERV 23.1 NITMVLMGALQPGLPSPVAIPQKYFKIIIDLKDCFFTIPLHPADQKRFAFSLPSTNFRQPMKRYQWKVLPQGMANSPTLCQKYVAAAIEPVRKIWTQMYI
SERV 25.2 NITMVLMGALQPGLPSPVAIPQKYFKIIIDLKDCFFTIPLHPADQKRFAFSLPSTNFRQPMKRYQWRVLPQGMANSPTLCQKYVAAAIEPVRKIWTQMYI
CHIAP     NNQMRPLGPVQRGLPLLSALPQNWKLIIIDIKDCFFSIPLFPRDRQRFAFTVPSLNHMEPDKRYQWRVLPQGMANSPTICQLYVQKALEPVRKQFTSMIM
SHIAP     NNQMHLFGPVQRGLPLLSALPQDWKLIIIDIKDCFFSIPLYPRDRPRFAFTIPSLNHMEPDKRFQWKVLPQGMANSPTICQLYVQEALEPIRKQFTSLIV
MIAP      NEQMNLFGPVQRGLPVLSALPRGWNLIIIDIKDCFFSIPLCPRDRPRFAFTIPSINSDEPDNRYQWKVLPQGMSNSPTMCQLYVQEALLPVREQFPSLIL
```

```
MMTV      VHYMDDILLAHPSRSIVDEILTSMIQALNKHGLVVSTEKIQKYDNLKYLGTHIQGDSVSYQKLQIRTDKLRTLNDFQKLLGNINWIRPFLKLTTGELKPL
HERV-K10  IHYIDDILCAAETKDKLIDCYTFLQAEVANAGLAIASDKIQTSTPFHYLGMQIENRKIKPQKIEIRKDTLKTLNDFQKLLGDINWIRPTLGIPTYAMSNL
JSRV      VHYMDDILLAHTDEHLLYQAFSILKQHLSLNGLVIADEKIQTHFPYNYLGFSLYPRVYNTQLVKLQTDHLKTLNDFQKLLGDINWIRPYLKLPTYTLQPL
ENTV      VHYMDDILLAHTDEHLLYQAFSILKQHLSLNGLVIADEKIQTHFPYNYLGFSLYPRTYNTQLVKLQTDHLKTLNDFQKLLGDINWIRPYLKLPTYTLQPL
TvERV-K1  IHYMDDILIAGPCEQELLRITTLLITALKDRGLYISPDKVQLASPFFFLGFELHSSHIITQKAQIRTNHLHTLNDFQKLLGDINWLRPYLKITTGDLKPL
pTvERV-K2 IHYMDDILIAGPCEQELLRITTLLITALKDRGLYISPDKVQLASPFFFLGFELHSSHIITQKAQIRTNHLHTLNDFQKLLGDINWLRPYLKITTGDLKPL
SMRV-H    LHYMDDILLACDSAEAAKACYAHIISCLTSYGLKIAPDKVQVSEPFSYLGFELHHQQVFTPRVCLKTDHLKTLNDFQKLLGDIQWLRPYLKLPTSALVPL
MPMV      IHYMDDILIAGKDGQQVLQCFDQLKQELTAAGLHIAPEKVQLQDPYTYLGFELNGPKITNQKAVIRKDKLQTLNDFQKLLGDINWLRPYLKLPTGDLKPL
SRV-1     IHYMDDILIAGKDGQQVLQCFDQLKQELTIAGLHIAPEKIQLQDPYTYLGFELNGPKITNQKAVIRKDKLQTLNDFQKLLGDINWLRPYLKLTTADLKPL
SRV-2     IHYMDDILIAGKLGEQVLQCFAQLKQALTTTGLQIAPEKVQLQDPYTYLGFQINGPKITNQKAVIRRDKLQTLNDFQKLLGDINWLRPYLHLTTGDLKPL
SERV 23.1 IHYMDNILIAGKIGEQVLQCFAQLKQKLTAAGLQIAPEKAQLQDPYTYLGFQINGPKIINQKAVIRHDHLKTLNDFQKLLGDINWLRPYLKLTTGELKPL
SERV 25.2 IHYMDDILIAGQIGDQVLQCFAQLKQELTTAGLQIAPEKIQLQNPYTYLGFQINGPKIINQKAVIHRDQLKTLNDFQKLLGDINWLRPYLKLTTGELKPL
CHIAP     IHYMDDILICHRKIEVLQQAFPMLVAELKQWGLEIASEKVQVSDTGLFLGSVITPTKIIPQKIEIRKDHLRTLNDFQKLLGDINWLRPFLKIPSADLKPL
SHIAP     IHYMDDILICHKELDVLQKAFPMLVAELKQWGLEIASEKVQIADTGLFLGSKITPKNIVPQKIEIRKDHLQTLNDFQKLLGDINWLRPFLKIPSADLKPL
MIAP      LLYMDDILLCHKELTMLQKAYPFLLKTLSQWGLQIATEKVQISDTGQFLGSVVSPDKIVPQKVEIRRDHLHTLNNFQKLLGDINWLRPFLKIPSAELRPL

MMTV      FEILNGDSNPISTRKLTPEACKALQLMNERLSTARVKRLDLSQPWSLCILKTEYTPTACLWQDG---------VVEWIHLPHISPKVITPYDIFCTQLII
HERV-K10  FSILRGDSDLNSQRILTPEATKEIKLVEEKIQSAQINRIDPLAPLQLLIFATAHSPTGIIIQNTD---------LVEWSFLPHSTVKTPTLYLDQIATLIG
JSRV      FDILKGDSDPASPRTLSLEGRTALQSIEEAIRQQQITYCDYQRSWGLYILPTPRAPTGVLYQDK---------PLRWIYLSATPTKHLLPYYELVAKIIA
ENTV      FDILKGDSDPASPRTLSLEGQTALQLIEEAIRRQQITYCDYQRSWGLYILPTPRAPTGVLYQDK---------PLRWIYLSATPTKHLLPYYELVAKIVA
TvERV-K1  FDILKGDSNPTSPRVLTKEGEAALQKINHAISAQKIGYYNPQKKLWFLIFSTSFSPTGLLWQD---------LPLYWIHGPATPTKVLPTYPLLTASLLQ
pTvERV-K2 FDILKGDSNPTSPRVLTKEGEAALQKINHAISAQKIGYYNPQKKLWFLIFSTSFSPTGLLWQD---------LPLYWIHGPATPTKVLPTYPLLTASLLQ
SMRV-H    NNILKGDPNPLSVRALTPEAKQSLALINKAIQNQSVQQISYNLPLVLLLLPTPHTPTAVFWQPNGTDPTKNGSPLLWLHLPASPSKVLLTYPSLLAMLII
MPMV      FDTLKGDSDPNSHRSLSKEALASLEKVETAIAEQFVTHINYSLPLIFLFNTALTPTGLFWQDN---------PIMWIHLPASPKKVLLPYYDAIADLII
SRV-1     FDTLKGDSNPNSHRSLSKEALALLDKVETAIAEQFVTHINYSLPLMFLIFNTALTPTGLFWQNN---------PIMWVHLPASPKKVLLPYYDAIADLII
SRV-2     FDILKGDSNPNSPRSLSEAALASLQKVETAIAEQFVTQIDYTQPLTFLIFNTTLTPTGLFWQNN---------PVMWVHLPASPKKVLLPYYDAIADLII
SERV 23.1 FDILKGDSNPKSPRSITKKTLITLQQVKHAIATQFVTSIDYSQPLIFIIFSTTITPTGLFWQNN---------PIMWVHLPSSPKKVLLPYYDAIADLII
SERV 25.2 FDILKGDSNPKSPRSISKEALTALQQVENPITTQFVTSIDYSQPLIFIIFNTAITPTGLFWQNN---------PIMWVHLPSSPKKVLFPYYDAIADLII
CHIAP     FDILEGEPHISSPRSFTPAACQALQKVEKALQDAQLHRIDETLPFSLCVFKTAKLPTAILWQHG---------PLLWIHPNASPAKIIDWYPDAVVQLAL
SHIAP     FDLLEGEPHISSPRKFTPAAHRALQMVEEALQEAQLQRIEESLPFNLCVFKTAQLPTAVLWQHG---------PLLWIHPNASPAKIIDWYPDAVAQPRS
MIAP      FWYLEGDPHISSPRTLTLAANQALQKVEKALQNAQLQAIEDSQPFSLCVFKTAQLPTAVLWQNG---------PLLWIHPNVSPAKIIDWYPDAIAQLAL

MMTV      KGRHRSKELFSKDPDYIVVPYTKVQFDLLLQEKEDWPISLLGFLGEVHFHLPKDPLLTFTLQTAIIFPHMTSTTPLEKGIVIFTDGSANGRSVTYIQGRE
HERV-K10  QTRLRITKLCGNDPDKIVVPLTKEQVRQAFINSGAWQIGLANFVGLIDNHYPKTKIPQFLKLTTWILPKITRREPLENALTVFTDGSSNGK-AAYTGPKE
JSRV      KGRHEAIQYFGMEPPFICVPYALEQQDWLFQFSDNWSIAFANYPGQITHHYPSDKLLQFASSHAFIFPKIVRRQPIPEATLIFTDGSSNGTAALIINHQT
ENTV      KGRHEAIQYFGMEPPFICVPYALEQQDWLFQFSDNWAIAFANYPGRIIHHYPPDKLLQFASSHAFIFPKIVRRQPIPNATLIFTDGSSNGTAALIINHQT
TvERV-K1  EGGSLSVKFFGKDPDHIVCPYTSAQLSWLQQNDDTWAISSVSYQGILDNHYPPDKLIQFLRDNPVAFPRITKRDPLENARLVFTDGSSSGTAAISIDSNL
pTvERV-K2 EGRSLSVKFFGKDPDHIVCPYTSSQLSWLQQNDDTWAISSVSYQGILDNHYPPDKLIQFLRDNPVAFPRITKRDPLENARLVFTDGSSSGTAAISIDSNL
SMRV-H    KGRYTGRQLFGRDPHSIIIPYTQDQLTWLLQTSDEWAIALSSFTGDIDNHYPSDPVIQFAKLHQFIFPKITKCAPIPQATLVFTDGSSNGIAAYVIDNQP
MPMV      LGRDHSKKYFGIEPSTIIQPYSKSQIDWLMQNTEMWPIACASFVGILDNHYPPNKLIQFCKLHTFVFPQIISKTPLNNALLVFTDGSSTGMAAYTLTDTT
SRV-1     LGRDHSKKYFGIEPSVIIQPYSKSQIDWLMQNTEMWPIACASYVGILDNHYPPNKLIQFCKLHAFIFPQIISKTPLNNALLVFTDGSSTGMAAYTLADTT
SRV-2     LGRDNSKKYFGLEPSTIIQPYSKSQIHWLMQNTETWPIACASYAGNIDNHYPPNKLIQFCKLHAVVFPRIISKTPLDNALLVFTDGSSTGIAAYTFEKTT
SERV 23.1 LGRENSRKYFGIEPSTIIQPYTQSHIHWLLQNTEAWPIACASYTGTIDNHYPPNKLIQFCKLHAFVFPHITSKEPLNDALLIFTDGSSTGLAAYTYNDVI
SERV 25.2 LGRENSRKYFGIEPSTIIQPYTQSHIHWLLQNTEAWPIACASYTGTIDNHYPPNKLIQFCKLHAFVFPHITSKEPLNDALLIFTDGSSTGLAAYTYNNTI
CHIAP     RGIKAAVAHPGRDPHLLVVPYTTAQIQTLTATSNDWAVLVTSFSGKIDNHFPKHPILQFTQNQAIVFPQMTAKHPIPNGTVVYTDGSKTGVGAYVIGNKV
SHIAP     R-IKAAVTHFGRDPDSLIVPYTAAQVQTLAATSSDWAVLVTSFSGQIDNHFPKHPILQFALNQAIVFPQVTAKDPLPDGTVVYTDGSKTGLGAYVVKDRV
MIAP      KGLKAAITHFGRSPYLLIVPYTAAQVQTLAATSNDWAVLVTSFSGKIDNHYPKHPILQFAQNQSVVFPQITVRNPLKNGIVVYTDGSKTGIGAYVANGKV

MMTV      PIIKENTQNTAQQAEIVAVITAFEEVS-QPFNLYTDSKYVTGLFPEIETATL--SPRTKIYTELKHLQRLIHKRQEKFYIGHIRGHTGLPG--PLAQGNA
HERV-K10  RVIKTPYQ-SAQRDELVAVITVLQDFD-QPINIISDSAYVVQATRDVETALIKYSMDDQLNQLFNLLQQTVRKRNFPFYITYIRAHTNLPG--PLTKANE
JSRV      YYAQTSFS-SAQVVELFAVHQALLTVP-TSFNLFTDSSYVVGALQMIETVPIIGTTSPEVLNLFTLIQQVLHCRQHPCFFGHIRAHSTLPG--ALVQGNH
ENTV      YYTQTSFS-SAQVVELFAVHQALLTVP-TSFNLFTDSSYVVGALQMIETVPIIGTTSPEVFNLFTLIQQVLHCRQYPCFFGHIRAHSTLPG--ALVRGNH
TvERV-K1  TSFHTPYK-SAQLVELAAILWVFEHVP-DPMNIYTDSAYIAHSIPELETVPYIKPST-NAFSMFLQIQKLIRERAQPFFICHIRAHSDLPG--PLAKGND
pTvERV-K2 TSFHTPYK-SAQLVELAAILWVFEHVP-DPMNIYTDSAYIAHSIPELETVPYIKPST-NAFSMFLQIQKLIRERAQPFFICHIRAHSDLPG--PLAKGND
SMRV-H    ISIKSPYL-SAQLVELYAILQVFTVLAHQPFNLYTDSAYIAQSVPLLETVPFIKSST-NATPLFSKLQQLILNRQHPFFIGHLRAHLNLPG--PLAEGNA
MPMV      IKFQTNLN-SAQLVELQALIAVLSAFPNQPLNIYTDSAYLAHSIPLLETVAQIKHIS-ETAKLFLQCQQLIYNRSIPFYIGHVRAHSGLPG--PIAQGNQ
SRV-1     IKFQTNLN-SAQLVELQALIAVLSAFPNQPLNIYTDSAYLAHSIPLLETVAQIKHIS-ETAKLFLQCQQLIYNRSIPFYIGHVRAHSGLPG--PIAHGNQ
SRV-2     VRFKTSHT-SAQLVELQALIAVLSAFPHRALNVYTDSAYLAHSIPLLETVSHIKHIS-DTAKFFLQCQQLIYNRSIPFYLGHIRAHSGLPG--PLSQGNH
SERV 23.1 VKFQTTYT-SAQLVKLQAIIAALSAFPCQPLNIYTDSAYLAHSIPLLETVPQIKHIS-DTANLFLQCQQLIQKRTTPFFLGHIRAHSGLPG--PLTQGNA
SERV 25.2 VKFQTTYT-SAQLVELQAIIAALSAFPCQPLNIYTDSAYLAHSIPLLETVPQIKHIS-NTANLFLQCQQLIQKSTTPFFLGHIRAHSGLPG--PLAQGNS
CHIAP     VSKQFNET-SPQIVECQVVLEVLEAFP-GPLNIVSDSSYVVNAVNLLETAGIIRPSS-RVAGIFQKIQITLSNRRFPVFVTHVRAHSGLPG--PMSSGND
SHIAP     ISKQYNET-SPQVVECLIVLEVLEAFP-GPLNIVSDSSYVVNAVNLLEIAGIIRSSS-RVANIFQKIQAALLNRRFPVFITHVRAHSGLPG--PMSLGND
MIAP      VSKQYNEN-SPRMVECLVVLEVLKTFL-EPLNIVSDSCYVVNAVNLLEVAGVIKPSS-RVANIFQQIQLVLLSR-SPVYITHVRAHSGLPTSAPWLSGND

MMTV      YADSLTRILTALES--------AQESHAL-HHQNAAALRFQFHITREQAREIVKLCPNCPDWGHAP--QLGVNPRGLKPRVLWQMDVTHVSEFGKLKYVH
HERV-K10  QADLL--VSSALIK--------AQELHAL-THVNAAGLKNKFDVTWKQAKDIVQHCTQCQVL-HLPTQEAGVNPRGLCPNALWQMDVTHVPSFGRLSYVH
JSRV      TADVLTKQVFFQS-----AIDAARKSHDL-HHQNSHSLRLQFKISREAARQIVKSCSTCPQFFVLP--QYGVNPRGLRPNHLWQTDVTHIPQFGRLKYVH
ENTV      TADALTKQVFFQS-----AIDAARKSHDL-HHQNSHSLRLQFKISREAARQIVKSCSTCPQFFVLP--QYGVNPRGLRPNHLWQTDVTHIPQFGRLKYVH
TvERV-K1  TVDAASKHLVFASVLQPSPIETATKAHEL-HHLNAHSLSQIYKITREQARQIVKQCPGCLVLLPEP--HLGVNPRGLVPGELWQMDVTHFTPFGKLKYIH
pTvERV-K2 TVDAASKHLVFASVLQPSPIETAIKAHEL-HHLNAHSLSQIYKITREQARQIVKQCPGCLVLLPEP--HLGVNPRGLVPGELWQMDVTHFTPFGKLKYIH
SMRV-H    LADAATQIFPIIS----DPIHEATQAHTL-HHLNAHTLRLLYKITREQARDIVKACKQCVVATPVP--HLGVNPRGLVPNAIWQMDVTHFTPFGKQRFVH
MPMV      RADLATKIVASNIN---TNLESAQNAHTL-HHLNAQTLRLMFNIPREQARQIVKQCPICVTYLPVP--HLGVNPRGLFPNMIWQMDVTHYSEFGNLKYIH
SRV-1     KADLATKTVASNIN---TNLESAQNAHTL-HHLNAQTLKLMFNIPREQARQIVRQCPICATYLPVP--HLGVNPRGLLPNMIWQMDVTHYSEFGNLKYIH
SRV-2     ITDLATKVVATTLT---TNLTEAQTAHAL-HHLNAQSLRLMFKITREQARQIVKQCPTCVTYLPIP--HFGVNPKGLVPNMLWQMDVTHYSEFGKLKYVH
SERV 23.1 TADTATKTIATVTI---DNLQQAQKAHAL-HHLNAQTLRLMFKLTREQARQIVKQCANCITYLPVP--HLGVNPRGLIPNEIWQMDVTHHSEFGQLKYIH
SERV 25.2 TADMATKTIATVTT---DNLQQAQKAHAL-HHLNAQTLRLMFKITREQARQIVKQCANCITYLPVP--HLGVNPRGLIPNEIWQMDVTHHSEFGQLKYIH
CHIAP     LADRATKLMAAALS---TQIQAAQEFHQR-FHVTAETLRRQFALTKQEARQIVTQCKNCCEFLPAP--HVGINPHGIRPLQMWQMDVTHVASFGKLQYVH
SHIAP     LADKATKLVATALS---THAQAAKEFHKR-FHVTAETLRRRFALSRKEAREIVTQCQNCCEFLPTP--HMGINPRGIRPLQMWQMDVTHIPSFGRLQYVH
MIAP      LADKAT-SGGCSLS---SPVEAAQEIFITTFHVTAEHYRSRNSLTRKEARDIVTQCQSCCEFLPVP--HVGINPRGIRPLQVWQMDVTHVSSFGKLQYLH
```

```
MMTV        VTVDTYSHPTFATARTGEATKDVLQHLAQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARWKISHVTGIPYNPQGQAIVERTHQNIKAQLNKLQKAG----
HERV-K10    VTVDTYSHFIWATCQTGESTSHVKKHLLSCFAVMGVPEKIKTDNGPGYCSKAFQKFLSQWKISHTTGIPYNSQGQAIVERTNRTLKTQLVKQKEGGDS--
JSRV        VSIDTFSNFLMASLHTGESTRHCIQHLLFCFSTSGIPQTLKTDNGPGYTSRSFQRFCLSFQIHHKTGIPYNPQGQGIVERAHQRIKHQLLKQKKGN---E
ENTV        VSIDTFSNFLMASLHTGESTRHCIQHLLFCFSTSGIPQTLKTDNGPGYTSRSFQRFCLSFQIQHKTGIPYNPQGQGIVERAHQRLKHQLLKQKKGN---E
TvERV-K1    VSIDTFSGFLIATLQMGEATKHVINHIVASLAMGPKPKILKTDNGPGYVSSSFKRFCAYMNIRNITSIPYNPQGQGIVERAHQTLKRMISKLQSGSEKLY
pTvERV-K2   VSIDTFSGFLIATLQMGEATKHVINHIVASLAMGPKPKILKTDNGPGYVSSSFKQFCAYMNIRNITGIPYNPQGQGIVERAHQTLKRMISKLQSGSEKLY
SMRV-H      VTVDTFSGFILATPQTGEASKNVISHVIHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQIKHVTGIPYNPQGQGVVERAHQTLKNALNRLARSP--LG
MPMV        VSIDTFSGFLLATLQTGETTKHVITHLLHCFSIIGLPKQIKTDNGPGYTSKNFQEFCSTLQIKHITGIPYNPQGQGIVERAHLSLKTTIEKIKKGE--WY
SRV-1       VSIDTFSGFLLATLQTGETTKHVITHLLHCFSIIGLPKQIKTDNGPGYTSKNFQEFCSTLQIKHVTGIPYNPQGQGIVERAHLSLKTTIEKIKKGE--WY
SRV-2       VSIDTFSGFLVATLQTGEATKHVIAHLLHCFSIIGQPIHIKTDNGPGYTSSNFRAFCSKLHIKHTPGIPYNPQGQGIVERAHLSLKNTLEKIKKGE--WY
SERV 23.1   VCIDTYSGFIIATLQTGEATKHVIAHLLHCSSILGIPKQIKTDNGPGYIAKAFLQFCNTLQIKHTTGIPYNPQGQGIVERAHLSLKTVITKLKGGS--WY
SERV 25.2   VCIDTYSGFIIATLQTGEATKHVITHLLHCFSILGIPKQIKTDNGPGYIAKTFLQFCNTLQIKHITGIPYNPQGQGIVERAHLSLKTTIEKIKGGS---Y
CHIAP       VSVDTCSGIICATPLTGEKAAHVIQHCLEAWGAWGKPHILKTDNGPAYTSQKFQHFCRQMEITHLTGLPYNPQGQGIVECAHRTLKSYLIKQKEGM--GA
SHIAP       VSVDTCSGVMFATPLTGEKASYVIQHCLEAWSAWGKPRI-KTDNGPAYTSQKFRQFCRQMDVTHLTGLPYNPQGQGIVERAHRTLKSYLIKQKGSI--ED
MIAP        VSIDTCSGIMFASPLTGEKASHVIQHCLEAWSAWGKPRLLKTDNGPAYTSQKFQQFCRQMDVTHLTGLPYNPQGQGIVERAHRTLKTYLIKQKREL--EE

                                                                                                          Pol ←
MMTV        -KYYTPHHLLAHALFVLNHVNMD-NQGHTAAERHWGPISADPKP--MVMWKDLLTGSWKGPDVLITAGRG
HERV-K10    KECTTPQMQLNLALYTLNFLNIYRNQTTTSAEQHLTGKKNSPHEGKLIWWKDNKNKTWEIG-KVITWGRG
JSRV        LYSPSPHNALNHALYVLNFLTLD-TEGNSAAQRFWGERSSCKKP--LVRWKDPLTNLWYGPDPVLIWGRG
ENTV        LYSPSPHNVLNHALYVLNFLTLD-AEGNSAAQRFWGERSSCKKP--LVRWKDPLTNLWYGPDPVLIWGRG
TvERV-K1    IRANNHRTLLNHALFVLNFLTLD-AHGKSAADRLWHPSSAHDYA--QVKWKDPLTHQWHGPDPVLIWGKG
pTvERV-K2   IRANNHRTLLNHALFVLNFLTLD-AHGKSAADRLWHPSSAHDYA--QVKWKDPLTHQWHGPDPVLIWGKG
SMRV-H      FSMQQPRNLLSHALFQLNFLQLD-SQGRSAADRLWHPQTSQQHA--TVMWRDPLTSVWKGPDPVLIWGRG
MPMV        PRKGTPRNILNHALFILNFLNLD-DQNKSAADRFWHNNPKKQFA--MVKWKDPLDNTWHGPDPVLIWGRG
SRV-1       PTKGTPRNILNHALFILNFLNLD-DQNHSAADRFWHSNPRKQFA--MVKWKDPLDNTWPWPDPVIIWGRG
SRV-2       PTQGSPRNILNHALFILNFLNLD-AQNKSAADRFWHTSSKKEYA--MVKWKDPLDNTWHGPDPVLIWGRG
SERV 23.1   PVKGTPRNILNHALFILNFLNLD-SHGKSAADRFWHPESQKQFA--MVKWKDPLDNSWHGPDPVLIWGRG
SERV 25.2   LVKGTPRNILNHALFILNFLNLD-SHGKSAADHFWHPESQKQFA--MVKWKDPLDNSRHGPDPVLIWGRG
CHIAP       SLPSVPRVAISMALFTLNFLNTD-AQGHTAAKRHTS-EPERSKE--MVKWKDVLTGLWRGPDPILIRSRG
SHIAP       VLPSVPRVAVSMALFTLNFLNCD-AQGHTAADRHSL-EPDRPKE--MVKWKDVLTDLWKGPDPILIRSRG
MIAP        ILP-QHQESLSMALFTLNFLNID-VHGHTAAERQCS-EPDRPNE--MVKWKNVLHNKWYGPDPILIRSRG
```

References

Adkins, B., T. Hunter, and K. Beemon. 1982. Expression of the PRC II avian sarcoma virus genome. *Journal of Virology.* 41:767–780.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology.* 215:403–410.

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

Anderssen, S., E. Sjøttem, G. Svineng, and T. Johansen. 1997. Comparative analyses of LTRs of the ERV-H family of primate-specific retrovirus-like elements isolated from marmoset, African green monkey, and man. *Virology.* 234:14–30.

Andersson, M. -L., M. Lindeskog, P. Medstrand, B. Westley, F. May, and J. Blomberg. 1999. Diversity of human endogenous retrovirus class II-like sequences. *Journal of General Virology.* 80:255–260.

Arnold, B. A., R. W. Hepler, and P. M. Keller. 1998. One-step fluorescent probe product-enhanced reverse transcriptase assay. *BioTechniques.* 25:98–106.

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, eds. 1998. *Current Protocols in Molecular Biology.* John Wiley and Sons, Inc.

Bai, J., J. V. Bishop, J. O. Carlson, and J. C. DeMartini. 1999. Sequence comparison of JSRV with endogenous proviruses: envelope genotypes and a novel ORF with similarity to a G-protein-coupled receptor. *Virology.* 258:333–343.

Bai, J., R. -Y. Zhu, K. Stedman, C. Cousens, J. Carlson, J. M. Sharp, and J. C. DeMartini. 1996. Unique long terminal repeat U3 sequences distinguish exogenous Jaagsiekte sheep retroviruses associated with ovine pulmonary carcinoma from endogenous loci in the sheep genome. *Journal of Virology.* 70:3159–3168.

Baltimore, D. 1970. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature.* 226:1209–1211.

Baltimore, D. 1985. Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. *Cell.* 40:481–482.

Barré-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 1983. Isolation of a T-lymphotropic

retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science.* 220:868–871.

Battula, N. and L. A. Loeb. 1976. On the fidelity of DNA replication. Lack of exodeoxyribonuclease activity and error-correcting function in avian myeloblastosis virus DNA polymerase. *Journal of Biological Chemistry.* 251:982–986.

Bebenek, K. and T. A. Kunkel. 1993. The fidelity of retroviral reverse transcriptases. In *Reverse Transcriptase* (ed. A. M. Skalka and S. P. Goff), pp. 85–102. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Bénit, L., N. De Parseval, J. -F. Casella, I. Callebaut, A. Cordonnier, and T. Heidmann. 1997. Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a *gag* coding sequence closely related to the *Fv1* restriction gene. *Journal of Virology.* 71:5652–5657.

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

Berkhout, B., M. Jebbink, and J. Zsíros. 1999. Identification of an active reverse transcriptase enzyme encoded by a human endogenous HERV-K retrovirus. *Journal of Virology.* 73:2365–2375.

Berkowitz, R., J. Fisher, and S. P. Goff. 1996. RNA packaging. *Current Topics in Microbiology and Immunology.* 214:177–218.

Bernhard, W. 1960. The detection and study of tumor viruses with the electron microscope. *Cancer Research.* 20:712–727.

Best, S., P. Le Tissier, G. Towers, and J. P. Stoye. 1996. Positional cloning of the mouse retrovirus restriction gene *Fv1. Nature.* 382:826–829.

Best, S., P. R. Le Tissier, and J. P. Stoye. 1997. Endogenous retroviruses and the evolution of resistance to retroviral infection. *Trends in Microbiology.* 5:313–318.

Bittner, J. J. 1936. Some possible effects of nursing on the mammary gland tumour incidence in mice. *Science.* 84:162.

Blin, N. and D. W. Stafford. 1976. A general method for isolation of high molecular weight DNA from eukaryotes. *Nucleic Acids Research.* 3:2303–2308.

Boeke, J. D. and J. P. Stoye. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 343–435. Cold Spring Harbor Laboratory Press, New York.

Boller, K., H. König, M. Sauter, N. Mueller-Lantzsch, R. Löwer, J. Löwer, and R. Kurth. 1993. Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. *Virology.* 196:349–353.

Böni, J., H. Pyra, and J. Schüpbach. 1996. Sensitive detection and quantification of particle-associated reverse transcriptase in plasma of HIV-1-infected individuals by the product-enhanced reverse transcriptase (PERT) assay. *Journal of Medical Virology.* 49:23–28.

Borson, N. D., W. L. Salo, and L. R. Drewes. 1992. A lock-docking oligo(dT) primer for 5' and 3' RACE PCR. *PCR Methods and Applications.* 2:144–148.

Breathnach, R. and P. Chambon. 1981. Organization and expression of eucaryotic split genes coding for proteins. *Annual Review of Biochemistry.* 50:349–83.

Breslauer, K. J., R. Frank, H. Blöcker, and L. A. Marky. 1986. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences U.S.A.* 83:3746–3750.

Brown, J., A. L. Matthews, P. A. Sandstorm, and L. E. Chapman. 1998. Xenotransplantation and the risk of retroviral zoonosis. *Trends in Microbiology.* 6:411–415.

Brown, P. O. 1997. Integration. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 161–203. Cold Spring Harbor Laboratory Press, New York.

Bucher, P. 1990. Weight matirx descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology.* 212:563–578.

Canfield, P. J., J. M. Sabine, and D. N. Love. 1988. Virus particles associated with leukaemia in a koala. *Australian Veterinary Journal.* 65:327–328.

Chakrabarti, L., M. Guyader, M. Alizon, M. D. Daniel, R. C. Desrosiers, P. Tiollais, and P. Sonigo. 1987. Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses. *Nature.* 328:543–547.

Chevet, E., G. Lemaître, and M. D. Katinka. 1995. Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. *Nucleic Acids Research.* 23:3342–3344.

Chiu, I. M., P. R. Andersen, S. A. Aaronson, and S. R. Tronick. 1983. Molecular cloning of the unintegrated squirrel monkey retrovirus genome: organization and distribution of related sequences in primate DNAs. *Journal of Virology.* 47:434–441.

Chiu, I. -M., R. Callahan, S. R. Tronick, J. Schlom, and S. A. Aaronson. 1984. Major *pol* gene progenitors in the evolution of oncoviruses. *Science.* 223:364–370.

Chiu, I. -M. and S. F. Skuntz. 1986. Nucleotide sequence analysis of squirrel monkey retrovirus reveals a novel primer-binding site for $tRNA^{Lys}_{1,2}$. *Journal of Virology.* 1986:983–987.

Chomczynski, P. and N. Sacchi. 1987. Single-step method of RNA isolation by guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry.* 162:156–159.

Chong, H., W. Starkey, and R. G. Vile. 1998. A replication-competent retrovirus arising from a split-function packaging cell line was generated by recombination events between the vector, one of the packaging constructs, and endogenous retroviral sequences. *Journal of Virology.* 72:2663–2670.

Chung, C. T. and R. H. Miller. 1993. Preparation and storage of competent *Escherichia coli* cells. *Methods in Enzymology.* 218:621–627.

Church, G. M. and W. Gilbert. 1984. Genomic sequencing. *Proceedings of the National Academy of Sciences U.S.A.* 81:1991–1995.

Coffin, J. 1982. Endogenous viruses. In *Molecular Biology of Tumor Viruses,* 2nd Ed., Vol. 1: *RNA Tumor Viruses* (ed. R. Weiss, N. Teich, H. Varmus, and J. Coffin), pp. 1109–1203. Cold Spring Harbor Laboratory Press, New York.

Coffin, J. M. 1979. Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *Journal of General Virology.* 42:1–26.

Coffin, J. M. 1992a. Structure and classification of retroviruses. In *The Retroviridae* (ed. J. A. Levy), Vol. 1, pp. 19–49. Plenum Press, New York.

Coffin, J. M. 1992b. Superantigens and endogenous retroviruses: a confluence of puzzles. *Science.* 255:411–413.

Coffin, J. M. 1993. Reverse transcription and evolution. In *Reverse Transcriptase* (ed. A. M. Skalka and S. P. Goff), pp. 445–479. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Coffin, J. M., S. H. Hughes, and H. E. Varmus, eds. 1997a. *Retroviruses.* Cold Spring Harbor Laboratory Press, New York. 843 pp.

Coffin, J. M., S. H. Hughes, and H. E. Varmus. 1997b. The interactions of retroviruses and their hosts. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 335–341. Cold Spring Harbor Laboratory Press, New York.

Cordonnier, A., J. F. Casella, and T. Heidmann. 1995. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of Virology.* 69:5890–5897.

Cousens, C., E. Minguijon, R. G. Dalziel, A. Ortin, M. Garcia, J. Park, L. Gonzalez, J. M. Sharp, and M. de las Heras. 1999. Complete sequence of enzootic nasal tumor virus, a retrovirus associated with transmissible intranasal tumors of sheep. *Journal of Virology.* 73:3986–3993.

Cowan, P. E. 1990. Brushtail possum. In *The Handbook of New Zealand Mammals* (ed. C. M. King), pp. 68–98. Oxford University Press, Auckland.

Cowan, P. E. 1996. Possum biocontrol: prospects for fertility regulation. *Reproduction, Fertility and Development.* 8:655–660.

Cowan, P. E. and C. H. Tyndale-Biscoe. 1997. Australian and New Zealand mammal species considered to be pests or problems. *Reproduction, Fertility and Development.* 9:27–36.

Craigie, R. 1992. Hotspots and warm spots: integration specificity of retroelements. *Trends in Genetics.* 8:187–190.

Craven, R. C., A. E. Leure-duPree, R. A. Weldon Jr., and J. W. Wills. 1995. Genetic analysis of the major homology region of the Rous sarcoma virus Gag protein. *Journal of Virology.* 69:4213–4227.

Craven, R. C. and L. J. Parent. 1996. Dynamic interactions of the Gag polyprotein. *Current Topics in Microbiology and Immunology.* 214:65–94.

Crittenden, L. B., A. M. Fadly, and E. J. Smith. 1982. Effect of endogenous leukosis virus genes on response to infection with avian leukosis and reticuloendotheliosis viruses. *Avian Diseases.* 26:279–294.

Crittenden, L. B., E. J. Smith, and A. M. Fadly. 1984. Influence of endogenous viral (ev) gene expression and strain of exogenous avian leukosis virus (ALV) on mortality and ALV infection and shedding in chickens. *Avian Diseases.* 28:1037–1056.

Cullen, B. R. 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. *Microbiological Reviews.* 56:375–394.

Davies, J. F. I., Z. Hostomska, Z. Hostomsky, S. R. Jordan, and D. A. Matthews. 1991. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. *Science.* 252:88–95.

DeHaven, J. E., D. A. Schwartz, M. W. Dahm, E. S. Hazard III, R. Trifiletti, E. R. Lacy, and J. S. Norris. 1998. Novel retroviral sequences are expressed in the epididymis and uterus of Syrian hamsters. *Journal of General Virology.* 79:2687–2694.

Delarue, M., O. Poch, N. Tordo, D. Moras, and P. Argos. 1990. An attempt to unify the structure of polymerases. *Protein Engineering.* 3:461–467.

Delassus, S., P. Sonigo, and S. Wain-Hobson. 1989. Genetic organization of gibbon ape leukemia virus. *Virology.* 173:205–213.

Dhellin, O., J. Maestre, and T. Heidmann. 1997. Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. *The EMBO Journal.* 16:6590–6602.

Dickson, C., R. Eisenman, H. Fan, E. Hunter, and N. Teich. 1982. Protein biosynthesis and assembly. In *Molecular Biology of Tumor Viruses,* 2nd Ed., Vol. 1: *RNA Tumor Viruses* (ed. R. Weiss, N. Teich, H. Varmus, and J. Coffin), pp. 513–648. Cold Spring Harbor Laboratory Press, New York.

Donahue, P. R., E. A. Hoover, G. A. Beltz, N. Riedel, V. M. Hirsch, J. Overbaugh, and J. I. Mullins. 1988. Strong sequence conservation among horizontally transmissible, minimally pathogenic feline leukemia viruses. *Journal of Virology.* 62:722–731.

Donehower, L. A., R. C. Bohannon, R. J. Ford, and R. A. Gibbs. 1990. The use of primers from highly conserved *pol* regions to identify uncharacterized retroviruses by the polymerase chain reaction. *Journal of Virological Methods.* 28:33–46.

Doolittle, R. F. and D. -F. Feng. 1992. Tracing the origin of retroviruses. *Current Topics in Microbiology and Immunology.* 176:195–211.

Doolittle, R. F., D. -F. Feng, M. S. Johnson, and M. A. McClure. 1989. Origins and evolutionary relationships of retroviruses. *The Quarterly Review of Biology.* 64:1–30.

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

Dorner, A. J., F. Bonneville, R. Kriz, K. Kelleher, K. Bean, and R. J. Kaufman. 1991. Molecular cloning and characterization of a complete Chinese hamster provirus related to intracisternal A particle genomes. *Journal of Virology.* 65:4713–4719.

Einfeld, D. 1996. Maturation and assembly of retroviral glycoproteins. *Current Topics in Microbiology and Immunology.* 214:133–176.

Elder, J. H., D. L. Lerner, C. S. Hasselkus-Light, D. J. Fontenot, E. Hunter, P. A. Luciw, R. C. Montelaro, and T. R. Phillips. 1992. Distinct subsets of retroviruses encode dUTPase. *Journal of Virology.* 66:1791–1794.

Ellerman, V. and O. Bang. 1908. Experimentelle leukamie bei huhnern. *Zentralbl. Bakteriol. Parasitenkd. Infectionskr. Hyg. Abt. Orig.* 46:595–609.

Enders, J. and T. Peebles. 1954. Propagation in tissue cultures of cytopathogenic agents from patients with measles. *Proceedings of the Society of Experimental Biology and Medicine.* 86:277–287.

Evans, L. H. and M. W. Cloyd. 1985. Friend and Moloney murine leukemia viruses specifically recombine with different endogenous retroviral sequences to generate minke cell focus-forming viruses. *Proceedings of the National Academy of Sciences U.S.A.* 82:459–463.

Favor, J. and C. Morawetz. 1992. Insertional mutations in mammals and mammalian cells. *Mutation Research.* 284:53–74.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution.* 17:368–76.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution.* 39:783–791.

Fine, D. L., G. C. Clarke, and L. O. Arthur. 1979. Characterization of infection and replication of Mason-Pfizer monkey virus in human cell cultures. *Journal of General Virology.* 44:457–469.

Fine, D. and G. Schochetman. 1978. Type D primate retroviruses: a review. *Cancer Research.* 38:3123–3139.

Fitch, W. M. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology.* 20:406–416.

Frankel, W. N., J. P. Stoye, B. A. Taylor, and J. M. Coffin. 1990. A linkage map of endogenous murine leukemia proviruses. *Genetics.* 124:221–236.

French, N. S. and J. D. Norton. 1997. Structure and functional properties of mouse VL30 retrotransposons. *Biochimica et Biophysica Acta.* 1352:33–47.

Frohman, M. A. 1995. Rapid amplification of cDNA ends. In *PCR Primer: A Laboratory Manual* (ed. C. W. Dieffenbach and G. S. Dveksler), pp. 381–409. Cold Spring Harbor Laboratory Press, New York.

Frohman, M. A., M. K. Dush, and G. R. Martin. 1988. Rapid amplification of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences U.S.A.* 85:8998–9002.

Fukasawa, M., T. Miura, A. Hasegawa, S. Morikawa, H. Tsujimoto, K. Miki, T. Kitamura, and M. Hayami. 1988. Sequence of simian immunodeficiency virus from African green monkey, a new member of the HIV/SIV group. *Nature.* 333:457–461.

Gallo, R. C., S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, B. Safai, G. White, P. Foster, and P. D. Markham. 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science.* 224:500–503.

Ganslosser, U. and R. Etter-Ganslosser. 1990. Phalangers. In *Grzimek's Encyclopedia of Mammals* (ed. B. Grzimek), Vol. 1, pp. 305–311. McGraw-Hill, Inc., New York.

Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature.* 397:436–441.

Gao, F., L. Yue, A. T. White, P. G. Pappas, J. Barchue, A. P. Hanson, B. M. Greene, P. M. Sharp, G. M. Shaw, and B. H. Hahn. 1992. Human infection by genetically diverse SIV_{SM}-related HIV-2 in west Africa. *Nature.* 358:495–499.

Gardner, M. B., C. A. Kozak, and S. J. O'Brien. 1991. The Lake Casitas wild mouse: evolving genetic resistance to retroviral disease. *Trends in Genetics.* 7:22–27.

Garvey, K. J., M. S. Oberste, J. E. Elser, M. J. Braun, and M. A. Gonda. 1990. Nucleotide sequence and genome organization of biologically active proviruses of the bovine immunodeficiency-like virus. *Virology.* 175:391–409.

Goubau, P., A. M. Vandamme, and J. Desmyter. 1996. Questions on the evolution of primate T-lymphotropic viruses raised by molecular and epidemiological studies of divergent strains. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology.* 13(Suppl 1):S242–S247.

Grand, R. J. A. 1989. Acylation of viral and eukaryotic proteins. *Biochemical Journal.* 258:625–638.

Griffiths, D. J., P. J. W. Venables, R. A. Weiss, and M. T. Boyd. 1997. A novel exogenous retrovirus sequence identified in humans. *Journal of Virology.* 71:2866–2872.

Guntaka, R. V. 1993. Transcription termination and polyadenylation in retroviruses. *Microbiological Reviews.* 57:511–521.

Guyader, M., M. Emerman, P. Sonigo, F. Clavel, L. Montagnier, and M. Alizon. 1987. Genome organization and transactivation of the human immunodeficiency virus type 2. *Nature.* 326:662–669.

Hamilton, R. C., A. MacGregor, and D. Pye. 1979. A marsupial oncovirus? *Journal of General Virology.* 44:535–539.

Hardy, W. D. J. 1991. General principles of retrovirus immunodetection tests. *Journal of the American Veterinary Medicine Association.* 199:1282–1287.

Harris, J. M., R. H. Haynes, and E. M. McIntosh. 1997. A consensus sequence for a functional human endogenous retrovirus K (HERV-K) dUTPase. *Biochemistry and Cell Biology.* 75:143–151.

Hatfield, D. and S. Oroszlan. 1990. The where, what and how of ribosomal frameshifting in retroviral protein synthesis. *Trends in Biochemical Sciences.* 15:186–190.

Heberling, R. L., S. T. Barker, S. S. Kalter, G. C. Smith, and R. J. Helmke. 1977. Oncornavirus: isolation from a squirrel monkey (*Saimiri sciureus*) lung culture. *Science.* 195:289–292.

Hecht, S. J., K. E. Stedman, J. O. Carlson, and J. C. DeMartini. 1996. Distribution of endogenous type B and type D sheep retrovirus sequences in ungulates and other mammals. *Proceedings of the National Academy of Sciences U.S.A.* 93:3297–3302.

Hendy, M. D. and D. Penny. 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematics and Biosciences.* 59:277–290.

Heneine, W., S. Yamamoto, W. M. Switzer, T. J. Spira, and T. M. Folks. 1995. Detection of reverse transcriptase by a highly sensitive assay in sera from persons infected with human immunodeficiency virus type 1. *The Journal of Infectious Diseases.* 171:1210–1216.

Heneine, W., W. M. Switzer, P. Sandstrom, J. Brown, S. Vedapuri, C. A. Schable, A. S. Khan, N. W. Lerche, M. Schweizer, D. Neumann-Haefelin, L. E. Chapman, and T. M. Folks. 1998. Identification of a human population infected with simian foamy viruses. *Nature Medicine.* 4:403–407.

Henikoff, S. and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences U.S.A.* 89:10915–10919.

Herniou, E., J. Martin, K. Miller, J. Cook, M. Wilkinson, and M. Tristem. 1998. Retroviral diversity and distribution in vertebrates. *Journal of Virology.* 72:5955–5966.

Herrmann, M. and J. R. Kalden. 1994. PCR and reverse dot hybridization for the detection of endogenous retroviral transcripts. *Journal of Virological Methods.* 46:333–348.

Hillis, D. M., B. K. Mable, A. Larson, S. K. Davis, and E. A. Zimmer. 1996. Nucleic acids IV: sequencing and cloning. In *Molecular Systematics* (ed. D. M. Hillis, C. Moritz, and B. K. Mable), pp. 321–381. Sinauer Associates, Inc., Sunderland, Massachusetts.

Holmes, D. S. and M. Quigley. 1981. A rapid boiling method for the preparation of bacterial plasmids. *Analytical Biochemistry.* 114:193–197.

Hu, W. -S., V. K. Pathak, and H. M. Temin. 1993. Role of reverse transcriptase in retroviral recombination. In *Reverse Transcriptase* (ed. A. M. Skalka and S. P. Goff), pp. 251–274. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Hui, E. K. -W., P. -C. Wang, and S. J. Lo. 1998. Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. *Cellular and Molecular Life Sciences.* 54:1403–1411.

Hunter, E. 1997. Viral entry and receptors. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 71–119. Cold Spring Harbor Laboratory Press, New York.

Hunter, E. and R. Swanstrom. 1990. Retrovirus envelope glycoproteins. *Current Topics in Microbiology and Immunology.* 157:187–253.

Jacobo-Molina, A. and E. Arnold. 1991. HIV reverse transcriptase structure-function relationships. *Biochemistry.* 30:6351–6361.

Janke, A., X. Xu, and U. Arnason. 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proceedings of the National Academy of Sciences U.S.A.* 94:1276–1281.

Jenkins, N. A., N. G. Copeland, B. A. Taylor, and B. K. Lee. 1981. Dilute (*d*) coat colour mutation of DBA/2J mice is associated with the integration of an ecotropic MuLV genome. *Nature.* 293:370–374.

Jolly, S. E. 1993. Biological control of possums. *New Zealand Journal of Zoology.* 20:335–339.

Junghans, R. P., L. R. Boone, and A. M. Skalka. 1982. Retroviral DNA H structures: displacement-assimilation model of recombination. *Cell.* 30:53–62.

Jurka, J. 1998. Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology.* 8:333–337.

Kato, S., K. Matsuo, N. Nishimura, N. Takahashi, and T. Takano. 1987. The entire nucleotide sequence of baboon endogenous virus DNA: A chimeric genome structure of murine type C and type D retroviruses. *Japan Journal of Genetics.* 62:127–137.

Katz, R. A., J. P. G. Mack, G. Merkel, J. Kulkosky, Z. Ge, J. Leis, and A. M. Skalka. 1992. Requirement for a conserved serine in both processing and joining activities of retroviral integrase. *Proceedings of the National Academy of Sciences U.S.A.* 89:6741–6745.

Katz, R. A. and A. M. Skalka. 1990. Generation of diversity in retroviruses. *Annual Review of Genetics.* 24:409–445.

Katz, R. A. and A. M. Skalka. 1994. The retroviral enzymes. *Annual Review of Biochemistry.* 63:133–173.

Katzman, M., J. P. G. Mack, A. M. Skalka, and J. Leis. 1991. A covalent complex between retroviral integrase and nicked substrate DNA. *Proceedings of the National Academy of Sciences U.S.A.* 88:4695–4699.

Kaufman, S. L., R. C. Gallo, and N. R. Miller. 1979. Detection of virus-specific RNA in simian sarcoma-leukemia virus-infected cells in in situ hybridization to viral complementary DNA. *Journal of Virology.* 30:637–641.

Kelleher, C. D. and J. J. Champoux. 1998. Characterization of RNA strand displacement synthesis by Moloney murine leukemia virus reverse transcriptase. *The Journal of Biological Chemistry.* 273:9976–9989.

Keshet, E. and H. M. Temin. 1978. Sites of integration of reticuloendotheliosis virus DNA in chicken DNA. *Proceedings of the National Academy of Sciences U.S.A.* 75:3372–3376.

Khan, E., J. P. G. Mack, R. A. Katz, J. Kulkosky, and A. M. Skalka. 1991. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. *Nucleic Acids Research.* 19:851–860.

Kitamura, Y., T. Ayukawa, T. Ishikawa, T. Kanda, and K. Yoshiike. 1996. Human endogenous retrovirus K10 encodes a functional integrase. *Journal of Virology.* 70:3302–3306.

Klikova, M., S. S. Rhee, E. Hunter, and T. Ruml. 1995. Efficient in vivo and in vitro assembly of retroviral capsids from Gag poteins expressed in bacteria. *Journal of Virology.* 69:1093–1098.

Koo, H. -M., S. Parthasarathi, Y. Ron, and J. P. Dougherty. 1994. Pseudotyped REV/SRV retroviruses reveal restrictions to infection and host range within members of the same receptor interference group. *Virology.* 205:345–351.

Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Research.* 12:857–872.

Krieg, A. M., M. F. Gourley, and A. Perl. 1992. Endogenous retroviruses: potential etiologic agents in autoimmunity. *The FASEB Journal.* 6:2537–2544.

Kuff, E. L. and K. K. Lueders. 1988. The intracisternal A-particle gene family: structure and functional aspects. *Advances in Cancer Research.* 51:184–276.

Kulkosky, J., K. S. Jones, R. A. Katz, J. P. G. Mack, and A. M. Skalka. 1992. Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Molecular and Cellular Biology.* 12:2331–2338.

Kulkosky, J. and A. M. Skalka. 1994. Molecular mechanism of retroviral DNA integration. *Pharmac. Ther.* 61:185–203.

Lynch, R., ed. 1998. *Biological control of possums.* RSNZ Miscellaneous Series Vol. 45. The Royal Society of New Zealand, Wellington. 129 pp.

Mackintosh, C. G., J. L. Crawford, E. G. Thompson, B. J. McLeod, J. M. Gill, and J. S. O'Keefe. 1995. A newly discovered disease of the brushtail possum: wobbly possum syndrome. *New Zealand Veterinary Journal.* 43:126.

Mager, D. L. and N. L. Goodchild. 1989. Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. *American Journal of Human Genetics.* 45:848–854.

Mang, R., J. Goudsmit, and A. C. van der Kuyl. 1999. Novel endogenous type C retrovirus in baboons: complete sequence, providing evidence for baboon endogenous virus *gag-pol* ancestry. *Journal of Virology.* 73:7021–7026.

Marchuk, D., M. Drumm, A. Saulino, and F. S. Collins. 1991. Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *Nucleic Acids Research.* 19:1154.

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

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

Martinelli, S. C. and S. P. Goff. 1990. Rapid reversion of a deletion mutation in Moloney murine leukemia virus by recombination with a closely related endogenous provirus. *Virology.* 174:135–144.

Marx, P. A., D. H. Maul, K. G. Osborn, N. W. Lerche, P. Moody, L. J. Lowenstine, R. V. Henrickson, L. O. Arthur, R. V. Gilden, M. Gravell, W. T. London, J. L. Sever, J. A. Levy, R. J. Munn, and M. J. Gardner. 1984. Simian AIDS: isolation of a type D retrovirus and transmission of the disease. *Science.* 223:1083–1086.

Masahito, P., M. Nishioka, H. Ueda, Y. Kato, I. Tamazaki, K. Nomura, H. Sugano, and T. Kitagawa. 1995. Frequent development of pancreatic carcinomas in the *Rana nigromaculata* group. *Cancer Research.* 55:3781–3784.

Maudru, T. and K. Peden. 1997. Elimination of background signals in a modified polymerase chain reaction-based reverse transcriptase assay. *Journal of Virological Methods.* 66:247–261.

Mayer, J., E. Meese, and N. Mueller-Lantzsch. 1998. Human endogenous retrovirus K homologous sequences and their coding capacity in Old World primates. *Journal of Virology.* 72:1870–1875.

McClure, M. A. 1993. Evolutionary history of reverse transcriptase. In *Reverse Transcriptase* (ed. A. M. Skalka and S. P. Goff), pp. 425–444. Cold Spring Harbor Laboratory Press, New York.

McClure, M. A., M. S. Johnson, and R. F. Doolitlle. 1987. Relocation of a protease-like gene segment between two retroviruses. *Proceedings of the National Academy of Sciences U.S.A.* 84:2693–2697.

McClure, M. A., M. S. Johnson, D. -F. Feng, and R. F. Doolittle. 1988. Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. *Proceedings of the National Academy of Sciences U.S.A.* 85:2469–2473.

McGeoch, D. J. 1990. Protein sequence comparisons show that the 'pseudoproteases' encoded by poxviruses and certain retroviruses belong to the deoxyuridine triphosphatase family. *Nucleic Acids Research.* 18:4105–4110.

Medstrand, P. and D. L. Mager. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. *Journal of Virology.* 72:9782–9787.

Medstrand, P. and J. Blomberg. 1993. Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues. *Journal of Virology.* 67:6778–6787.

Meers, J., M. Perrott, M. Rice, and C. Wilks. 1998. The detection and isolation of viruses from possums in New Zealand. In *Biological control of possums* (ed. R. Lynch), pp. 25–28. RSNZ Miscellaneous Series Vol. 45. The Royal Society of New Zealand, Wellington.

Mietz, J. A., Z. Grossman, K. K. Lueders, and E. L. Kuff. 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.

Miller, A. D. 1997. Development and applications of retroviral vectors. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 437–473. Cold Spring Harbor Laboratory Press, New York.

Moore, R., M. Dixon, R. Smith, G. Peters, and C. Dickson. 1987. Complete nucleotide sequence of a milk-transmitted mouse mammary tumor virus: two frameshift suppression events are required for translation of gag and pol. *Journal of Virology.* 61:480–490.

Morgan-Capner, P. and J. R. Pattison. 1985. Techniques in clinical virology. In *Virology: a practical approach* (ed. B. W. J. Mahy), pp. 237–258. IRL Press Limited, Oxford.

Morris, R. S. and D. U. Pfeiffer. 1995. Directions and issues in bovine tuberculosis epidemiology and control in New Zealand. *New Zealand Veterinary Journal.* 43:256–265.

Mueller-Lantzsch, N., M. Sauter, A. Weiskircher, K. Kramer, B. Best, M. Buck, and F. Grässer. 1993. Human endogenous retroviral element K10 (HERV-K10) encodes a full-length Gag homologous 73-kDa protein and a functional protease. *AIDS Research and Human Retroviruses.* 9:343–350.

Murphy, J. E. and S. P. Goff. 1994. Forced integration of Moloney murine leukemia virus DNA with a mutant integration site occurs through recombination with VL30 DNA. *Virology.* 204:458–461.

Myers, G. 1997. Retroviral sequences. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 709–755. Cold Spring Harbor Laboratory Press, New York.

Nermut, M. V. and M. V. Hockley. 1996. Comparative morphology and structural classification of retroviruses. *Current Topics in Microbiology and Immunology.* 214:1–24.

Nowak, R. M. 1991. *Walker's Mammals of the World.* 5th ed. Vol. 1. The Johns Hopkins University Press, Baltimore, Maryland.

Oda, T., S. Ikeda, S. Watanabe, M. Hatsushika, K. Akiyama, and F. Mitsunobu. 1988. Molecular cloning, complete nucleotide sequence, and gene structure of the provirus genome of a retrovirus produced in a human lymphoblastoid cell line. *Virology.* 167:468–476.

O'Keefe, J. and M. Wickstrom. 1998. Viruses of the brushtail possum in New Zealand. In *Biological control of possums* (ed. R. Lynch), pp. 23–24. RSNZ Miscellaneous Series Vol. 45. The Royal Society of New Zealand, Wellington.

O'Keefe, J. S., W. L. Stanislawek, and D. D. Heath. 1997. Pathological studies of wobbly possum disease in New Zealand brushtail possums (*Trichosurus vulpecula*). *Veterinary Record.* 141:226–229.

Olmsted, R. A., V. M. Hirsch, R. H. Purcell, and P. R. Johnson. 1989. Nucleotide sequence analysis of feline immunodeficiency virus: genome organization and relationship to other lentiviruses. *Proceedings of the National Academy of Sciences U.S.A.* 86:8088–8092.

O'Neill, R. J. W., M. J. O'Neill, and J. A. M. Graves. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature.* 393:68–72.

Ono, M., H. Toh, T. Miyata, and T. Awaya. 1985. Nucleotide sequence of the Syrian hamster intracisternal A-particle gene: close evolutionary relationship of type A particle gene to types B and D oncovirus genes. *Journal of Virology.* 55:387–394.

Ono, M., T. Yasunaga, T. Miyata, and H. Ushikubo. 1986. Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *Journal of Virology.* 60:589–598.

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

Page, R. D. M. 1998. TreeView. Version 1.5.2.

Palmarini, M., C. Cousens, R. G. Dalziel, J. Bai, K. Stedman, J. C. DeMartini, and J. M. Sharp. 1996. The exogenous form of Jaagsiekte retrovirus is specifically associated with a contagious lung cancer of sheep. *Journal of Virology.* 70:1618–1623.

Palmarini, M., J. M. Sharp, M. de las Heras, and H. Fan. 1999. Jaagsiekte sheep retrovirus Is necessary and sufficient to induce a contagious lung cancer in sheep. *Journal of Virology.* 73:6964–6972.

Patience, C., G. R. Simpson, A. A. Colletta, H. M. Welch, R. A. Weiss, and M. T. Boyd. 1996. Human endogenous retrovirus expression and reverse transcriptase activity in the T47D mammary carcinoma cell line. *Journal of Virology.* 70:2654–2657.

Patience, C., D. A. Wilkinson, and R. A. Weiss. 1997. Our retroviral heritage. *Trends in Genetics.* 13:116–120.

Perl, A., J. D. Rosenblatt, I. S. Chen, J. P. DiVincenzo, R. Bever, B. J. Poiesz, and G. N. Abraham. 1989. Detection and cloning of new HTLV-related endogenous sequences in man. *Nucleic Acids Research.* 17:6841–6854.

Perron, H., J. A. Garson, F. Bedin, F. Beseme, G. Paranhos-Baccala, F. Komurian-Pradel, F. Mallet, P. W. Tuke, C. Voisset, J. L. Blond, B. Lalande, J. M. Seigneurin, B. Mandrand, and the Collaborative Research Group on Multiple Sclerosis. 1997. Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. *Proceedings of the National Academy of Sciences U.S.A.* 94:7583–7588.

Petropoulos, C. 1997. Retroviral taxonomy, protein structures, sequences, and genetic maps. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 757–805. Cold Spring Harbor Laboratory Press, New York.

Poch, O., I. Sauvaget, M. Delarue, and N. Tordo. 1989. Identification of four conserved motifs among the RNA-dependent polymerase coding elements. *The EMBO Journal.* 8:3867–3874.

Poiesz, B. J., F. W. Ruscetti, A. F. Gazdar, P. A. Bunn, J. D. Minna, and R. C. Gallo. 1980. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proceedings of the National Academy of Sciences U.S.A.* 77:7415–7419.

Polard, P. and M. Chandler. 1995. Bacterial transposases and retroviral integrases. *Molecular Microbiology.* 15:13–23.

Poulet, F. M., P. R. Bowser, and J. W. Casey. 1994. Retroviruses of fish, reptiles, and molluscs. In *The Retroviridae* (ed. J. A. Levy), Vol. 3, pp. 1–38. Plenum Press, New York.

Power, M. D., P. A. Marx, M. L. Bryant, M. B. Gardner, P. J. Barr, and P. A. Luciw. 1986. Nucleotide sequence of SRV-1, a type D simian acquired immune deficiency syndrome virus. *Science.* 231:1567–1572.

Pulsinelli, G. A. and H. M. Temin. 1994. High rate of mismatch extension during reverse transcription in a single round of retrovirus replication. *Proceedings of the National Academy of Sciences U.S.A.* 91:9490–9494.

Pyra, H., J. Böni, and J. Schüpbach. 1994. Ultrasensitive retrovirus detection by a reverse transcriptae assay based on product enhancement. *Proceedings of the National Academy of Sciences U.S.A.* 91:1544–1548.

Querat, G., G. Audoly, P. Sonigo, and R. Vigne. 1990. Nucleotide sequence analysis of SA-OMVV, a visna-related ovine lentivirus: phylogenetic history of lentiviruses. *Virology.* 175:434–447.

Raba, M., K. Limburg, M. Burghagen, J. R. Katze, M. Simsek, J. E. Heckman, U. L. Rajbhandary, and H. J. Gross. 1979. Nucleotide sequence of three isoaccepting lysine tRNAs from rabbit liver and SV40-transformed mouse fibroblasts. *European Journal of Biochemistry.* 97:305–318.

Rabson, A. B. and B. J. Graves. 1997. Synthesis and processing of viral RNA. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 205–261. Cold Spring Harbor Laboratory Press, New York.

Rao, J. K. M., J. W. Erickson, and A. Wlodawer. 1991. Structural and evolutionary relationships between retroviral and eucaryotic aspartic proteinases. *Biochemistry.* 30:4663–4671.

Rasko, J. E. J., J. -L. Battini, R. J. Gottschalk, I. Mazo, and A. D. Miller. 1999. The RD114/simian type D retrovirus receptor is a neutral amino acid transporter. *Proceedings of the National Academy of Sciences U.S.A.* 96:2129–2134.

Rein, A., L. E. Henderson, and J. G. Levin. 1998. Nucleic-acid-chaperone activity of retroviral nucleocapsid proteins: significance for viral replication. *Trends in Biochemical Sciences.* 23:297–301.

Rein, A., N. Rice, S. Simek, M. Cohen, and R. J. Mural. 1982. In situ hybridization: general infectivity assay for retroviruses. *Journal of Virology.* 43:1055–1060.

Repaske, R., P. E. Steele, R. R. O'Neill, A. B. Rabson, and M. A. Martin. 1985. Nucleotide sequence of a full-length human endogenous retroviral segment. *Journal of Virology.* 54:764–772.

Rice, M. and C. R. Wilks. 1996. Virus and virus-like particles observed in the intestinal contents of the possum, Trichosurus vulpecula. *Archives of Virology.* 141:945–950.

Ringold, G., E. Y. Lasfargues, J. M. Bishop, and H. E. Varmus. 1975. Production of mouse mammary tumor virus by cultured cells in the absence and presence of hormones: assay by molecular hybridization. *Virology.* 65:135–147.

Ristevski, S., D. F. J. Purcell, J. Marshall, D. Campagna, S. Nouri, S. P. Fenton, D. A. McPhee, and G. Kannourakis. 1999. Novel endogenous type D retroviral particles expressed at high levels in a SCID mouse thymic lymphoma. *Journal of Virology.* 73:4662–4669.

Roberts, J. D., K. Bebenek, and T. A. Kunkel. 1988. The accuracy of reverse transcriptase from HIV-1. *Science.* 242:1171–1173.

Robinson, W. S., A. Pitkanen, and H. Rubin. 1965. The nucleic acid of the Bryan strain of Rous sarcoma virus: purification of the virus and isolation of the nucleic acid. *Proceedings of the National Academy of Sciences U.S.A.* 54:137–144.

Rosenberg, N. and P. Jolicoeur. 1997. Retroviral pathogenesis. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 475–585. Cold Spring Harbor Laboratory Press, New York.

Rous, P. 1911. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *Journal of Experimental Medicine.* 13:397–411.

Rustigian, R., P. Johnston, and H. Reihart. 1955. Infection of monkey kidney tissue cultures with virus-like agents. *Proceedings of the Society of Experimental Biology and Medicine.* 88:8–16.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution.* 4:406–425.

Saltarelli, M., G. Querat, D. A. Konings, R. Vigne, and J. E. Clements. 1990. Nucleotide sequence and transcriptional analysis of molecular clones of CAEV which generate infectious virus. *Virology.* 179:347–364.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: a laboratory manual.* 2nd Edition. Cold Spring Harbor, New York: CSH Laboratory Press.

Sanchez, G., D. Gautheret, X. Xu, A. -L. Chenine, and I. Hirsch. 1998. Relative amplification efficiency of differently sized templates by long-distance PCR. *BioTechniques.* 24:400–402.

Sarid, R., A. Chajut, E. Gak, Y. Kim, C. V. Hixson, S. Oroszlan, S. R. Tronick, A. Gazit, and A. Yaniv. 1994. Genome organization of a biologically active

molecular clone of the lymphoproliferative disease virus of turkeys. *Virology.* 204:680–691.

Sarngadharan, M. G., M. Robert-Guroff, and R. C. Gallo. 1978. DNA polymerases of normal and neoplastic mammalian cells. *Biochimica et Biophysica Acta.* 516:419–487.

Sauter, M., S. Schommer, E. Kremmer, K. Remberger, G. Dölken, I. Lemm, M. Buck, B. Best, D. Neumann-Haefelin, and N. Mueller-Lantzsch. 1995. Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. *Journal of Virology.* 69:414–421.

Schlom, J. and S. Spiegelman. 1971. Simultaneous detection of reverse transcriptase and high molecular weight RNA unique to oncogenic RNA viruses. *Science.* 174:840–843.

Schmidt, M., T. Wirth, B. Kroger, and I. Horak. 1985. Structure and genomic organization of a new family of murine retrovirus-related DNA sequences (MuRRS). *Nucleic Acids Research.* 13:3461–3470.

Schmidt, M. F. G. 1989. Fatty acylation of proteins. *Biochimica et Biophysica Acta.* 988:411–426.

Schommer, S., M. Sauter, H. -G. Kräusslich, B. Best, and N. Mueller-Lantzsch. 1996. Characterization of the human endogenous retrovirus K proteinase. *Journal of General Virology.* 77:375–379.

Screaton, G. R., C. R. M. Bangham, and J. I. Bell. 1993. Direct sequencing of single primer PCR products: a rapid method to achieve short chromosomal walks. *Nucleic Acids Research.* 21:2263–2264.

Seifarth, W., C. Baust, A. Murr, H. Skladny, F. Krieg-Schneider, J. Blusch, T. Werner, R. Hehlmann, and C. Leib-Mösch. 1998. Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles. *Journal of Virology.* 72:8384–8391.

Seifarth, W., H. Skladny, F. Krieg-Schneider, A. Reichart, R. Hehlmann, and C. Leib-Mösch. 1995. Retrovirus-like particles released from the human breast cancer cell line T47-D display type B- and C-related endogenous retroviral sequences. *Journal of Virology.* 69:6408–6416.

Seperack, P. K., M. C. Strobel, D. J. Corrow, N. A. Jenkins, and N. G. Copeland. 1988. Somatic and germ-line reverse mutation rates of the retrovirus-induced dilute coat-color mutation of DBA mice. *Proceedings of the National Academy of Sciences U.S.A.* 85:189–192.

Shih, A., R. Misra, and M. G. Rush. 1989. Detection of multiple, novel reverse transcriptase coding sequences in human nucleic acids: relation to primate retroviruses. *Journal of Virology.* 63:64–75.

Shimotohno, K., Y. Takahashi, N. Shimizu, T. Gojobori, D. W. Golde, I. S. Chen, M. Miwa, and T. Sugimura. 1985. Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: an open reading frame for the protease gene. *Proceedings of the National Academy of Sciences U.S.A.* 82:3101–3105.

Silver, J., T. Maudru, K. Fujita, and R. Repaske. 1993. An RT-PCR assay for the enzyme activity of reverse transcriptase capable of detecting single virions. *Nucleic Acids Research.* 21:3593–3594.

Simpson, G. R., C. Patience, R. Löwer, R. R. Tönjes, H. D. M. Moore, R. A. Weiss, and M. T. Boyd. 1996. Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. *Virology.* 222:451–456.

Skalka, A. M. and S. P. Goff, eds. 1993. *Reverse Transcriptase.* Cold Spring Harbor Laboratory Press, New York. 492 pp.

Smit, A. F. A. 1996. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development.* 6:743–748.

Sonigo, P., M. Alizon, K. Staskus, D. Klatzmann, S. Cole, O. Danos, E. Retzel, P. Tiollais, A. Haase, and S. Wain-Hobson. 1985. Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. *Cell.* 42:369–382.

Sonigo, P., C. Barker, E. Hunter, and S. Wain-Hobson. 1986. Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. *Cell.* 45:375–385.

Steinhuber, S., M. Brack, G. Hunsmann, H. Schwelberger, M. P. Dierich, and W. Vogetseder. 1995. Distribution of human endogenous retrovirus HERV-K genomes in humans and different primates. *Human Genetics.* 96:188–192.

Stephens, R. M., J. W. Casey, and N. R. Rice. 1986. Equine infectious anemia virus gag and pol genes: relatedness to visna and AIDS virus. *Science.* 231:589–594.

Stoye, J. P., S. Fenner, G. E. Greenoak, C. Moran, and J. M. Coffin. 1988. Role of endogenous retroviruses as mutagens: the hairless mutation of mice. *Cell.* 54:383–391.

Stoye, J. P., C. Moroni, and J. M. Coffin. 1991. Virological events leading to spontaneous AKR thymomas. *Journal of Virology.* 65:1273–1285.

Sverdlov, E. D. 1998. Perpetually mobile footprints of ancient infections in the human genome. *FEBS Letters.* 428:1-6.

Swain, A. and J. M. Coffin. 1992. Mechanism of transduction by retroviruses. *Science.* 255:841–845.

Swanstrom, R. and J. W. Wills. 1997. Synthesis, assembly, and processing of viral proteins. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 263–334. Cold Spring Harbor Laboratory Press, New York.

Swofford, D. L. 1998. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b2. Sinauer Assoc. Inc. Sunderland, Massachusetts.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. In *Molecular Systematics* (ed. D. M. Hillis, C. Moritz, and B. K. Mable), pp. 407–514. Sinauer Associates, Inc., Sunderland, Massachusetts.

Tchenio, T. and T. Heidmann. 1992. Defective retroviruses can disperse in the human genome by intracellular transposition. *Journal of Virology.* 65:2113–2118.

Tchenio, T. and T. Heidmann. 1992. High-frequency intracellular transposition of a defective mammalian provirus detected by an in situ colorimetric assay. *Journal of Virology.* 66:1571–1578.

Teich, N. 1982. Taxonomy of retroviruses. In *Molecular Biology of Tumor Viruses,* 2nd Ed., Vol. 1: *RNA Tumor Viruses* (ed. R. Weiss, N. Teich, H. Varmus, and J. Coffin), pp. 25–207. Cold Spring Harbor Laboratory Press, New York.

Telesnitsky, A. and S. P. Goff. 1997. Reverse transcriptase and the generation of retroviral DNA. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 121–160. Cold Spring Harbor Laboratory Press, New York.

Temin, H. M. 1964. Homology between RNA from Rous sarcoma virus and DNA from Rous sarcoma virus-infected cells. *Proceedings of the National Academy of Sciences U.S.A.* 52:323–329.

Temin, H. M. 1989. Retrovirus variation and evolution. *Genome.* 31:17–22.

Temin, H. M. 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *Proceedings of the National Academy of Sciences U.S.A.* 90:6900–6903.

Temin, H. M. and D. Baltimore. 1972. RNA-directed DNA synthesis and RNA tumor viruses. *Advances in Virus Research.* 17:129–186.

Temin, H. M. and S. Mizutani. 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature.* 226:1211–1213.

Temin, H. M. and H. Rubin. 1958. Characteristics of an assay for Rous sarcoma virus and Rous sarcoma cells in tissue culture. *Virology.* 6:669–688.

ten Dam, E., I. Brierley, S. Inglis, and C. Pleij. 1994. Identification and analysis of the pseudoknot-containing gag-pro ribosomal frameshift signal of simian retrovirus-1. *Nucleic Acids Research.* 22:2304–2310.

Thayer, R. M., M. D. Power, M. L. Bryant, M. B. Gardner, P. J. Barr, and P. A. Luciw. 1987. Sequence relationships of type D retroviruses which cause simian acquired immunodficiency syndrome. *Virology.* 157:317–329.

Thenius, E. 1990. Phylogeny. In *Grzimek's Encyclopedia of Mammals* (ed. B. Grzimek), Vol. 1, pp. 219–231. McGraw-Hill, Inc., New York.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research.* 25:4876–4882.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.

Todaro, G. J. 1975. Evolution and modes of transmission of RNA tumor viruses. *American Journal of Pathology.* 81:590–606.

Toh, H., R. Kikuno, H. Hayashida, T. Miyata, W. Kugimiya, S. Inouye, S. Yuki, and K. Saigo. 1985. Close structural resemblance between putative polymerase of a *Drosophila* transposable genetic element 17.6 and *pol* gene product of Moloney murine leukaemia virus. *The EMBO Journal.* 4:1267–1272.

Tönjes, R. R., C. Limbach, R. Löwer, and R. Kurth. 1997. Expression of human endogenous retrovirus type K envelope glycoprotein in insect and mammalian cells. *Journal of Virology.* 71:2747–2756.

Tönjes, R. R., R. Löwer, K. Boller, J. Denner, B. Hasenmeier, H. Kirsch, H. Konig, C. Korbmacher, C. Limbach, R. Lugert, R. C. Phelps, J. Scherer, K. Thelen, J. Löwer, and R. Kurth. 1996. HERV-K: the biologically most active human endogenous retrovirus family. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology.* 13(Suppl 1):S261–S267.

Tristem, M. 1996. Amplification of divergent retroelements by PCR. *BioTechniques.* 20:608–612.

Tristem, M., E. Herniou, K. Summers, and J. Cook. 1996. Three retroviral sequences in amphibians are distinct from those in mammals and birds. *Journal of Virology.* 70:4864–4870.

Tsujimoto, H., A. Hasegawa, N. Maki, M. Fukasawa, T. Miura, S. Speidel, R. W. Cooper, E. N. Moriyama, T. Gojobori, and M. Hayami. 1989. Sequence

of a novel simian immunodeficiency virus from a wild-caught African mandrill. *Nature.* 341:539–541.

Urnovitz, H. B. and W. H. Murphy. 1996. Human endogenous retroviruses:nature, occurrence, and clinical implications in human disease. *Clinical Microbiology Reviews.* 9:72–99.

van der Kuyl, A. C., R. Mang, J. T. Dekker, and J. Goudsmit. 1997. Complete nucleotide sequence of simian endogenous type D retrovirus with intact genome organization: evidence for ancestry to simian retrovirus and baboon endogenous virus. *Journal of Virology.* 71:3666–3676.

Varmus, H. and P. Brown. 1989. Retroviruses. In *Mobile DNA* (ed. D. E. Berg and M. M. Howe), pp. 53–108. American Society for Microbiology, Washington, D. C.

Vogt, P. K. 1997a. Historical introduction to the general properties of retroviruses. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 1–25. Cold Spring Harbor Laboratory Press, New York.

Vogt, V. M. 1996. Proteolytic processing and particle maturation. *Current Topics in Microbiology and Immunology.* 214:95–131.

Vogt, V. M. 1997b. Retroviral virions and genomes. In *Retroviruses* (ed. J. M. Coffin, S. H. Hughes, and H. E. Varmus), pp. 27–69. Cold Spring Harbor Laboratory Press, New York.

Voo, K. S. and B. M. Jacobsen. 1998. Rapid resuspension of pelleted bacterial cells for miniprep plasmid DNA isolation. *BioTechniques.* 24:240–243.

Weiss, R. 1982. Experimental biology and assay of RNA tumor viruses. In *Molecular Biology of Tumor Viruses,* 2nd Ed., Vol. 1: *RNA Tumor Viruses* (ed. R. Weiss, N. Teich, H. Varmus, and J. Coffin), pp. 25–207. Cold Spring Harbor Laboratory Press, New York.

Weiss, R. A. and L. N. Payne. 1971. The heritable nature of the factor in chicken cells which acts as a helper virus for Rous sarcoma virus. *Virology.* 45:508–515.

Weiss, R. A., W. S. Mason, and P. K. Vogt. 1973. Genetic recombinants and heterozygotes derived from endogenous and exogenous avian RNA tumor viruses. *Virology.* 52:535–552.

Whiting, S. H. and J. J. Champoux. 1994. Strand displacement synthesis capability of moloney murine leukemia virus reverse transcriptase. *Journal Of Virology.* 68:4747–4758.

Whiting, S. H. and J. J. Champoux. 1998. Properties of strand displacement synthesis by Moloney murine leukemia virus reverse transcriptase: mechanistic implications. *Journal of Molecular Biology.* 278:559–577.

Wichman, H. A. and R. A. Van Den Bussche. 1992. In search of retrotransposons: exploring the potential of the PCR. *BioTechniques.* 13:258–264.

Wilkinson, D. A., D. L. Mager, and J. -A. C. Leong. 1994. Endogenous human retroviruses. In *The Retroviridae* (ed. J. A. Levy), Vol. 3, pp. 465–535. Plenum Press, New York.

Wilton, S. and L. Lim. 1996. Long-range PCR: synthesis of products independent of size. *Trends in Genetics.* 12:458.

Worley, M., B. Rideout, A. Shima, and D. Janssen. 1993. Opportunistic infections, cancer and hematologic disorders associated with retrovirus infection in the koala. In *Proceedings of the American Association of Zoo Vetrinarians.* pp. 181–182. Saint Louis:.

Xiong, Y. and T. H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal.* 9:3353–3362.

Xu, H. and J. D. Boeke. 1987. High-frequency deletion between homologous sequences during retrotransposition of Ty elements in *Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences U.S.A.* 84:8553–8557.

Yamamoto, S., T. M. Folks, and W. Heneine. 1996. Highly sensitive qualitative and quantitative detection of reverse transcriptase activity: optimization, validation, and comparative analysis with other detection systems. *Journal of Virological Methods.* 61:135–143.

York, D. F., R. Vigne, D. W. Verwoerd, and G. Querat. 1992. Nucleotide sequence of the Jaagsiekte retrovirus, an exogenous and endogenous type D and B retrovirus of sheep and goats. *Journal of Virology.* 66:4930–4939.

Yoshida, M., I. Miyoshi, and Y. Hinuma. 1982. Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. *Proceedings of the National Academy of Sciences U.S.A.* 79:2031–2035.

Zhou, A., X. Jiang, and X. Xu. 1997. Improved alkaline lysis method for rapid isolation of plasmid DNA. *BioTechniques.* 23:592–594.