Chapter 1

Background and literature review

Image on reverse: Australian Old Endemic rodent, *Conilurus penicillatus*.
Modified image from the private collection of Assoc. Prof. Bill Breed

Chapter 1

Background and literature review

1.1 Background

Surrounding mammalian oocytes is an extracellular matrix, the zona pellucida (ZP), which is secreted by (although not exclusively in some species) developing oocytes and, after fertilization, remains in place until dissolving shortly before implantation (Yanagimachi 1994). This glycoprotein matrix has elastic properties similar to that of a gel (Green 1997). The three-dimensional structure of the ZP reflects its multiple functions as it protects and encloses the ovulated oocyte as well as providing a level of molecule permeability. However, the same properties that protect the oocyte, and later the developing conceptus, also pose a challenge to the fertilising spermatozoon ("the sperm"). The sperm, which has had to navigate its way along the female reproductive tract to reach the ampulla of the oviduct, must first bind to the ZP, undergo the acrosome reaction, and only then can penetration of the matrix occur and hence binding to and fusion with the plasma membrane (Yanagimachi 1994).

In the laboratory mouse ("the mouse") the series of steps of sperm-ZP interaction that leads to successful fertilization have been extensively investigated and a model has been proposed that is generally accepted (Yanagimachi 1994). The capacitated sperm that has reached the oviduct must first penetrate the cumulus oophorus, a matrix of cumulus cells and hyaluronic acid, by releasing hyaluronidase, which enzymatically dissolves the oophorus (Lin *et al.* 1994; Talbot *et al.* 2003). Once the sperm has reached the ZP, the receptors on the plasma membrane overlying the sperm head bind to ligands (either sugar or protein) on the surface of the ZP (Jungnickel *et al.* 2003) The sperm attaches firstly in a non-covalent manner and then subsequently binds in an irreversible process. This event is often referred to as primary sperm-ZP binding (Bleil & Wassarman 1983). The molecules that are

involved in this process have recently become the subject of much debate with alternative models being proposed.

Once the sperm has bound to the ZP (either to sugar ligands or the matrix itself), a series of catalytic events, the acrosome reaction (AR), occurs. This reaction causes the release of lytic enzymes from the acrosome, a large vesicle positioned between the plasma membrane of the sperm head and the sperm nucleus (Bleil *et al.* 1988). The AR must occur before the sperm can penetrate the zona matrix as the reaction exposes secondary receptors on the inner membrane of the acrosome allowing the sperm to adhere to, and then penetrate, the matrix. This event, in the mouse model, is known as secondary sperm-ZP binding (Bleil *et al.* 1988). Regardless of the molecules involved in secondary binding, the sperm must remain associated with the ZP and yet be able to move through the matrix to reach the perivitelline space between the ZP and the oocyte membrane (the oolemma).

Upon entering the perivitelline space the sperm can bind to, and then fuse, with the oolemma of the oocyte. Fertilization has then taken place. Shortly after fusion, the cortical reaction occurs whereby granules positioned beneath the surface of the oocyte release their contents into the perivitelline space, causing a change in the ZP, often referred to as zona hardening, and rending the ZP refractory to continued sperm binding. Poorly understood changes to the cell membrane of the oocyte also occur, preventing any other sperm in the perivitelline space from fusing with the oolemma (Gardner & Evans 2006). Figure 1.1 provides a diagrammatic representation of these events.

As the developing embryo divides and commences to move down the oviduct towards the uterus, the zona matrix continues to surround the embryo, preventing attachment of the blastomeres to the oviductal epithelium (Yanagimachi 1994). Prior to implantation of the blastocyst to the uterine endometrium, the ZP is proteolytically dissolved.

NOTE: This figure is included on page 3 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.1. Diagrammatic representation of the events that occur at mammalian fertilization (Redrawn from Wassarman 1988).

1.2 Literature Review

1.2.1 Fertilisation and the role of the ZP

The importance of primary sperm-ZP binding in the events that lead to successful fertilisation is evident as, unless this occurs, the sperm cannot reach the membrane of the oocyte, the oolemma. Extensive research has been conducted on this aspect of fertilisation with most research using the laboratory mouse as the model system. Hence, much of what we know of these events is from this species. The difficulty has been in extrapolating the hypothetical models based on mouse data to other species such as humans, where fertility is of intense interest. A considerable amount of the mouse research is conducted using *in vitro* procedures which may be unreliable indicators of the actual *in vivo* events of fertilization (Hunter & Rodriquez-Martinez 2002). While the relative advancement of mouse genetics has enabled researchers to use transgenic mice to unravel the events of fertilization the actual biochemical or molecular processes involved in primary sperm-ZP binding*, in vivo*, have proven elusive. Furthermore, due to the ethical prohibition on using human oocytes, even the *in vitro* mouse experiments have not been replicated using human oocytes.

While the simplest model for primary sperm-ZP binding involves the sperm receptors binding to oligosaccharides on the ZP, the inability to identify the specific mouse sperm receptor suggests that this event is much more complex. Certainly there is evidence that there are both high and low binding affinities involved in fertilization, and it is possible that there are a number of different types of receptor-ligand binding events taking place (Thaler & Cardullo 1996; Castle 2002). It has also proven difficult to definitively identify the sugars that bind sperm receptors. In addition, the actual chemical process of primary sperm-ZP binding is currently being revised with active debate continuing. Two groups of researchers have emerged and their papers, including 2007 publications, provide alternative and contrasting models. The first group, Wassarman and colleagues, have been investigating primary sperm-ZP binding for over two decades and from their experiments proposed that, in mice, ZP3 provided

the *O*-linked oligosaccharide essential for sperm-ZP binding. The second group, Rankin, Dean and colleagues, have developed a model that is carbohydrate independent and, together with mass spectrometry data on the glycosylation of mZP3, assert that primary sperm-ZP binding involves the supramolecular structure or zona scaffold of the matrix rather than one specific glycan or glycoprotein. Other groups, such as Miller and Shur, have developed models that include other molecules, either on the sperm head (β4GalTase) (Miller *et al.* 1993; Shur *et al.* 1998) or accessory to either the sperm or the ZP (SED1) (Ensslin & Shur 2003). Both the Wassarman model and the Miller and Shur models involve mZP3 as the sperm receptor.

As stated previously, the main disagreement appears to involve the actual molecular interactions that occur at fertilization, such as which zona protein is involved rather than the actual occurrence of primary sperm-ZP binding. Therefore, it has been generally accepted that, in mammalian fertilization, sperm must first bind, or adhere, to the zona pellucida matrix, undergo the AR, then undergo secondary binding and penetrate the ZP, enter the perivitelline space and finally fuse with the oolemma.

The ZP of the mouse is comprised of 3 sulfated glycoproteins, named ZP1, ZP2 and ZP3. A fourth ZP glycoprotein has been found in the genomes of several species including the rat and the human (Conner *et al.* 2005). However, in the mouse this fourth gene is present in the form of a pseudogene, where a nonsense mutation results in a prematurely truncated protein and no protein products have been found (Conner *et al.* 2005). Therefore, most of the models of fertilization based on the mouse, do not take into account any role the protein products of *Zp4* may play in either the structure of the zona matrix or primary sperm-ZP binding.

The nomenclature of ZP genes has historically been confusing with some researchers naming their newly sequenced ZP cDNAs using a numerical system based on the relative mass of each protein shown on 2D gel electrophoresis (ZP1, ZP2, ZP3) (Bleil & Wassarman 1980) while other researchers

used an alphabetical system based on the relative size of the cDNAs (ZPB. ZPA, ZPC) (Harris *et al.* 1994). Hence, ZP1 is ZPB, ZP2 is ZPA and ZP3 is ZPC. The discovery of a fourth ZP gene in humans (Lefievre *et al.* 2004) and in the rat (Conner *et al.* 2005, Hoodbhoy *et al.* 2005), named *ZpB*1 (renamed ZP4), has only added to this confusion. Recently, the National Center for Biotechnology Information (NCBI) Nomenclature Commitee determined that the numerical system of naming ZP genes should be used (Conner *et al.* 2005) and therefore, to reduce confusion and introduce consistency, this nomenclature of ZP genes will be followed in this thesis.

1.2.2 Structure of the ZP

The model for the three dimensional structure of the ZP that has prevailed for the past twenty years provides that heterodimers of ZP2 and ZP3 form filaments that are cross linked with ZP1 (Fig. 1.2) (Greve & Wassarman 1985; Green 1997).

NOTE: This figure is included on page 7 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.2. Diagrammatic representation of the arrangements of filaments of the ZP matrix. **(**Redrawn from Wassarman *et al.* 2004a)

Scanning electron microscopy (SEM) has revealed that the external surface of the zona has a “swiss cheese” appearance with numerous fenestrations forming “pores” that vary in diameter between species (Sinowatz *et al.* 2001). The smallest pores are found in the cow and the largest occur in the ZP of the mouse and cat (Sinowatz *et al.* 2001), although the significance of pore size for species specificity of sperm penetration does not appear to be known. Extensive SEM studies on human and mouse ZP have shown that, despite the presence of a fourth ZP glycoprotein expressed at similar levels to ZP2 and ZP3 in the human (Lefievre *et al.* 2004), there is little structural difference in the ZP between that species and the mouse (Familiari *et al.* 2006) (Fig. 1.3.)

NOTE: This figure is included on page 8 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.3. A) SEM image of the ZP of an unfertilised mouse oocyte, showing the fine networks of filaments and the spongy structure of the ZP (x50,000). B) SEM image at the same magnification of a human oocyte showing different patterns of filament aggregation but nevertheless having similar structural appearance (x50,000).
(Both images taken from Familiari *et al.* 2006).

The thickness of the ZP also varies between species, with the matrix being 7 µm thick in the mouse, 13 µm in humans and up to 27 µm thick in the cow (Dunbar *et al.* 1994).

The glycosylated content of the ZP matrix has been estimated to be approximately 50% w/v (Green 1997; Hoodbhoy & Dean 2004). Both *N*-linked oligosaccharides (sugar chains attached to asparagine

residues in the conformation NXS/T, with X being any amino acid but proline) and *O*-linked oligosaccharide (sugar chains attached to either serine or threonine residues) structures are involved. A recent study of the ZP glycans in the mouse (Easton *et al.* 2000) suggested that there are over 25 possible *N*-linked glycans and over 12 *O*-linked glycans. The *N*-glycans are found to have a range of different structures, with both high-mannose and complex types. Easton *et al.* (2000) also identified bi-, tri- and tetra-antennary *N*-glycan structures which were terminated with multiple antennae, including N-acetylglucosamine. The *O*-linked glycans were composed of Core 2 types but terminal N-acetylglucosamine was not identified on *O*-glycans (Easton *et al.* 2000). This last result contrasts to findings by Nagdas *et al.* (1994) who reported *O*-linked terminal N-acetylglucosamine.

Genetic manipulations of mouse ZP glycoproteins have provided insights into the structure of the ZP. Transgenic mouse lines were produced whereby each ZP gene in turn was 'knocked out' or prevented from being expressed, and the ZP of oocytes from each transgenic mouse were studied. *Zp*1 null mice (those mice which lack a *Zp1* functional gene) produced a zona matrix, albeit loosely organised, and were fertile with reduced fecundity (Rankin *et al.* 1999). Mice that lacked a *Zp2* gene produced a thin zona matrix in early follicular development which was not sustained around oocytes of large follicles (Rankin *et al.* 2001). Mice that did not have a functional *Zp3* gene failed to produce a zona pellucida and were infertile (Rankin *et al.* 1996; Liu *et al.* 1996). These results have led to a revised model of ZP structure whereby it has been suggested that the zona filaments comprise interspersed heterodimers of ZP1/ZP3 and ZP2/ZP3 (Hoodbhoy & Dean 2004, see Fig.1.4). The revised model differs from the Wassarman model in that the filaments are composed of ZP proteins that interact via their ZP domains and these filaments are in turn cross-linked by the N-terminal domains of ZP1 and ZP2. ZP1, by forming intermolecular disulfide bonds may provide structural support (Dean 2003).

NOTE: This figure is included on page 10 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.4 A) Diagrammatic representation of the three dimensional structure of the zona pellucida matrix, showing how the three glycoproteins interact with each other and are crosslinked via the N-terminal domains of ZP1 and ZP2. ZP1 also crosslinks with itself, forming dimers, providing structural support. B) Diagrammatic representation of the results of experiments whereby transgenic mice were produced that lacked one of the glycoproteins
(Redrawn from Dean 2003).

1.2.3 ZP glycoproteins

Of the three mouse glycoproteins, ZP1 has the largest mass and ZP3 the smallest. The molecular weight of ZP1 is 200,000 Da, ZP2 is 120,000 Da and ZP3 is 83,000 Da (Bleil & Wassarman 1980). Removing the extensive glycosylation of the glycoproteins reduces the mass of the nascent proteins to 75,000 Da (ZP1), 81,000 Da (ZP2) and 44,000 Da (ZP3) (Wassarman 1988). cDNAs of the mouse ZP genes, and hence the predicted polypeptides, have been determined. The lengths of each glycoprotein vary: ZP1 is 623 residues in length, ZP2 is the largest with 713 residues, and ZP3 is the smallest in length at 424 residues (McLeskey *et al.* 1998).

All three mouse glycoproteins contain a large ZP domain. This domain is common to other secreted glycoproteins such as some growth factors (Bork & Sander 1992) and is thought to be involved in the assembly of the matrix. Each ZP glycoprotein contains a N-terminus signal sequence (for directing proteins to the secretory pathway) and a large hydrophobic membrane-spanning domain near the C-

terminus. Situated N-terminus to the transmembrane domain is a conserved furin cleavage site (CFCS) thought to be where the transmembrane domain is cleaved from the secreted protein (Fig. 1.5).

NOTE: This figure is included on page 11 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.5 Diagrammatic representation of the primary structure of the three mouse ZP glycoproteins.
(Redrawn from Wassarman *et al.* 2004a).

1.2.4 mZP3

Due to the apparent importance ZP3 plays in the assembly of the zona matrix, research has tended to focus on the role each domain of this glycoprotein plays in the secretion and assembly of the mouse zona matrix. Using epitope-tagged mutated *Zp3* cDNAs micro-injected into the nucleus of growing mouse oocytes (Qi *et al.* 2002) or transfecting either mouse embryonal carcinoma cells (EC) or Chinese Hamster ovary (CHO) cells (Williams & Wassarman 2001), it was shown that mutation of the furin cleavage site prevented secretion of the ZP3 nascent glycoprotein and resulted in accumulation of ZP3 in the ER. However, deletion of the transmembrane domain did not affect secretion but ZP3 failed to become incorporated into the zona matrix (Jovine *et al.* 2002). Furthermore, when the ZP domain was removed from ZP3, the mutated protein was still secreted but failed to be incorporated into the matrix (Jovine *et al.* 2002). It therefore appears that both the ZP domain and the transmembrane domain are

not necessary for the secretion but essential for assembly of the zona matrix (see Fig. 1.6 for primary structure of mZP3).

Fig. 1.6. Diagrammatic representation of the *Zp3* gene in the mouse, showing the exons and identified domains and regions. The secreted protein is cleaved at the conserved furin cleavage site (CFCS).
(Redrawn from Williams *et al.* 2006)

The role each glycoprotein plays in fertilization has also been investigated. Early research, using isolated and purified ZP glycoproteins, together with sperm-inhibition assays, determined that in the mouse only mZP3 bound capacitated sperm and induced the AR (Bleil & Wassarman 1983). Those sperm that had undergone spontaneous AR failed to bind to ZP3, but were able to bind to ZP2 (Bleil *et al.* 1988). ZP1 appeared to play no role in sperm-ZP binding in the mouse, although there was some evidence that it did in the pig (Yurewicz *et al.* 1998). However, due to nomenclature problems, proteins identified as ZP1 in some organisms, such as the pig, have now been renamed ZP4 (Conner *et al.* 2005), and in the light of this, it would appear that ZP4 may play a role in fertilization in species other than the mouse.

Removal of *O*-linked oligosaccharide chains from solubilised ZP3 abolished sperm-ZP binding whereas removal of *N*-linked oligosaccharide chains did not (Florman & Wassarman 1985), suggesting sperm bind to *O*-linked oligosaccharides on ZP3. The identification of which glycosides are involved in sperm-
ZP binding has proven to be difficult due to the number of candidate molecules. Galactose at the nonreducing terminal of *O*-glycans (Bleil & Wassarman 1988; Litscher *et al.* 1995) has been implicated as well as N-acetylglucosamine (Miller *et al.* 1992), fucose (Johnston *et al.* 1998) and mannose (Tulsiani *et al.* 1992).

While the Wassarman model of sperm-ZP binding involves protein-carbohydrate interaction, there is a growing body of evidence that suggests it is the ZP3 protein itself that binds sperm. Chapman *et al.* (1998) demonstrated that prokaryotically expressed ZP3 with no glycosylation, bound human sperm and induced the AR. In addition, similarly expressed bonnet monkey ZP3, without glycosylation, also bound sperm (Patra *et al.* 2000; Gahlay *et al.* 2002). Recently, non-glycosylated prokaryotically expressed mouse ZP3 molecular constructs containing different regions of the protein were used in experiments to inhibit *in vitro* sperm-ZP binding (Li *et al.* 2007). Only the fragment containing the carboxyl region of ZP3 was able to inhibit sperm-ZP binding *in vitro*. It was therefore concluded that the polypeptide backbone of the carboxyl region of ZP3 is involved in sperm-ZP binding in the mouse.

Researchers, using mouse embryonal carcinoma (EC) cells transfected with mouse ZP3 (mZP3), produced biologically active ZP3 which could act like purified wild type ZP3 when binding sperm (Kinloch *et al.* 1995). This finding enabled researchers to experiment with mutated constructs of ZP3. Coupled with the non-biologically active EC generated hamster ZP3, exon-swapping experiments established that the C-terminus region of mZP3 was involved in sperm-ZP binding (Kinloch *et al.* 1995). The region encoded by exon 7 of *Zp3* provides the conserved furin cleavage site and after secretion becomes the C-terminus of the protein. Recently, it was demonstrated that removal of exon 7 from EC generated constructs abolished sperm-ZP binding (Williams *et al.* 2006).

In the mouse, ZP3 has also been implicated in the acrosome reaction (AR). It has been shown that solubilized mZP3 expressed from mammalian expression systems such as EC cells, induced the AR (Bleil & Wassarman 1983; Kinloch *et al.* 1991; Wassarman *et al.* 2001). However, oligosaccharides isolated from ZP3 or other small glycoproteins were unable to do so (Florman *et al.* 1984; Leyton & Saling 1989). In addition, the polypeptide backbone of recombinant human ZP3 was able to induce the AR (Chapman *et al.* 1998). Miller, Shur and associates have investigated the sperm protein β1,4 galactosyltransferase (GalTase), which binds terminal N-acetylglucosamine on ZP3 *O*-linked

oligosaccharides in a non-catalytic reaction (Miller *et al.* 1992, 1993), and causes sperm to undergo the AR. Although genetic manipulation experiments have tended not to support the role of GalTase in sperm-ZP binding, it appears to induce the AR by activating a G-protein response (Lu & Shur 1997; Lu *et al.* 1997). Miller and Shur have resolved the difficulties in their model by suggesting that GalTase "sees" the ZP3 oligosaccharides in the context of other sperm surface proteins that are responsible for the initial binding or tethering of sperm to the ZP (Shur 1998).

1.2.4.1 Exon 7 of mZP3

It now appears to be generally accepted that the ZP domain (encoded by exons 2-6), the conserved furin cleavage site (CFCS) and the transmembrane domain (encoded by exon 8) of mZP3 are involved in both secretion and assembly of mZP3 into the zona matrix (Litscher *et al.* 1999; Williams & Wassarman 2001; Qi *et al.* 2002; Jovine *et al.* 2002; Jovine *et al.* 2004). The region encoded by exon 7 does not appear to be involved in this process.

The region encoded by exon 7 is 46 amino acids in length. At the C-terminus end is the CFCS where the nascent protein is cleaved at position 351 (asparagine) (Boja *et al.* 2003). This region in the mouse contains four cysteine residues all considered to be involved in disulfide bonding, a large number of serine residues (seven) and is overall hydrophilic (Boja *et al.* 2003).

Researchers using EC expressed constructs (see above) were also able to narrow down the specific area within the exon 7 coding region that provides the *O*-linked oligosaccharides involved in sperm-ZP binding (Kinloch *et al.* 1995). In particular, two serine residues (potential sites for linkage to *O*-glycans) were found to be essential for sperm-ZP binding (Ser-332 and Ser-334) (Chen *et al.* 1998). These two serine residues are contained within a stretch of amino acids the Wassarman group named the 'combining-site for sperm' (Wassarman & Litscher 1995).

1.2.4.2 The combining-site for sperm

The term 'combining-site for sperm' used by the Wassarman group to identify the site within the exon 7 coding region that is involved in sperm-ZP binding is not generally used by other researchers. A specific region of 16 amino acids (mZP3-328 to mZP3-343) was initially identified as the combining-site for sperm with the aid of immunochemistry although the selection of the epitope used in the experiment was not explained and appears to be arbitrary (Rosiere & Wassarman 1992). Since identification of the specific combining-site for sperm, the actual stretch of amino acids involved varies from publication to publication. Table 1.1 demonstrates the lack of consensus of the exact location of the putative combining-site for sperm within the Wassarman group. Recent papers by Wassarman and colleagues refer only to the region encoded by exon 7 of *Zp3* (Williams *et al.* 2006). This region, encoded by exon 7, is situated between the C-terminus end of the ZP domain (encoded by exons 2 to 6) and the transmembrane domain (encoded by exon 8). The furin cleavage site is encoded by the last few residues of exon 7 (mZP3 350 to 353). It therefore appears that the combining-site for sperm refers to a stretch of amino acids within the exon 7 coding region and not to specific residues as identified by Rosiere and Wassarman (1992).

Table 1.1. Specific residues within the exon 7 coding region of *Zp3* identified as the combining-site for sperm. In the right hand column are the references where the specific regions are referred to.

Residue position according to mZP3	Reference
328 to 343	Rosiere & Wassarman, 1992
318 to 353	Wassarman & Litscher 1995
329 to 334	Wassarman 1999a
	Wassarman 1999b
	Wassarman *et al.* 1999
	Chen *et al.* 1999
319 to 339	Wassarman & Litscher 2001
309 to 354 (all of exon 7)	Jovine *et al.* 2002,
	Wassarman *et al.* 2004b
326 to 338	Williams *et al.* 2003
309 to 349	Jovine *et al.* 2004

This region shows a lack of sequence conservation between species. For example, between the mouse and human, the proportion of amino acids that are conserved within the region encoded by exon 7 is 55% and between mouse and rat, 83%. Wassarman has speculated that the high level of divergence within this region may determine the species specificity seen between heterologous gametes (Wassarman & Litscher 1995: see Fig. 1.7). This proposal has been put forward repeatedly by the Wassarman group during the past ten years (Kinloch *et al.* 1995; Wassarman *et al.* 1996; Chen *et al.* 1998; Wassarman *et al.* 1999; Wassarman 1999b; Wassarman & Litscher 2001; Wassarman 2002; Williams *et al.* 2003; Wassarman *et al.* 2004b; Williams *et al.* 2006). Furthermore, codons within exon 7 have been found to be evolving under positive Darwinian selection (Swanson *et al.* 2001, 2003; Jansa *et al.* 2003; Turner & Hoekstra 2006) and this finding has been used by Wassarman and colleagues in support of their proposal.

NOTE: This figure is included on page 17 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.7. Comparison of the putative combining site region of the mouse, hamster, human and marmoset. The alignment (using single letter amino acid code) demonstrates the lack of sequence conservation within the region 318 to 353. The vertical lines above the residues indicate the sites of potential *O*-linked glycosylation, while the residues underlined are those sites that are potentially *N*-linked glycosylated.
(Redrawn from Wassarman & Litscher 1995).

1.2.5 Supramolecular or zona scaffold model

The model proposed by Wassarman and colleagues has been the dominant model for two decades. However, recently, another model has been proposed which is carbohydrate independent. Rankin and colleagues performed a number of experiments using transgenic mice where *Zp2* and *Zp3* were 'knocked out' and human *Zp2* (hu*Zp2*) and *Zp3* (hu*Zp3*) were 'knocked in' (Rankin *et al.* 1998, 2003). These "rescued" mice (so called rescued because hu*Zp2* and hu*Zp3* rendered mice fertile and were therefore considered to be "rescued" from the infertile *Zp2* null and *Zp3* null state) produced oocytes with intact zona pellucida comprised of mouse ZP1 (mZP1), huZP2 and huZP3. The expectation had been that human sperm would be able to bind to the chimeric zona but in fact, this did not occur. Only mouse sperm bound to the zona matrix of the transgenic mice. Rates of ovulation and fecundity were observed to be lower in mice with huZP2 and huZP3 and *in vitro* assays showed a decreased rate of fertilization, suggesting zona developmental abnormalities. The molecular mass of the transgenic zona pellucida was distinct from wild type mouse zona, suggesting that the human ZP glycoproteins were posttranslationally modified as native human ZP2 and ZP3. These results suggested to the researchers that the primary structure of ZP2 and ZP3 alone could not account for 'taxon-specific' binding of sperm to the ZP as, despite the differences in amino acid sequence, mouse sperm still bound to the human sequence while human sperm did not (Rankin *et al.* 1998).

The observation that led to the formulation of a 'supramolecular' model of sperm-ZP binding was that sperm continued to bind to the ZP after cortical granule exocytosis following sperm fusion with the mouse oolemma (Rankin *et al.* 2003). In wild type ZP, mouse sperm do not continue to bind to the zona of fertilized eggs, as the zona becomes refractory to continued binding. It had been hypothesized by Wassarman and colleagues that this occurs due to the removal of oligosaccharides that bind ZP3 and a cleaving of ZP2 (Bleil & Wassarman 1988). Rankin and colleagues observed that in the transgenic mice the huZP2 did not undergo proteolytic cleaving after cortical granule exocytosis (Rankin *et al.* 2003). They concluded that it is the cleavage of ZP2 after fertilization that provides the block to continued binding of sperm. Even though sperm continued to bind after fertilization there was no evidence of polyspermy, presumably due to changes to the egg plasma membrane that are independent to the zona glycoproteins. Furthermore, it could not possibly be sugars that bound the sperm to the ZP, as they would have been cleaved after fertilization in the ZP2 rescue mice, had the Wassarman model held true (Dean 2003). A sugar that could bind the sperm but not be available for cleaving was not plausible, they claimed, and that any model of fertilization must take this into account (Dean 2003). Rankin and colleagues proposed a model that sperm bind to the supramolecular structure of the ZP, not to one specific ZP glycoprotein or to any specific oligosaccharide (Fig. 1.8).

Following successful fertilization proteases cleave ZP2, altering the supramolecular structure and rendering it unable to support sperm binding (Rankin *et al.* 2003; Dean 2003). With the transgenic mice, this did not occur, as the cleavage site may have been buried due to a different configuration of the matrix in the mouse background.

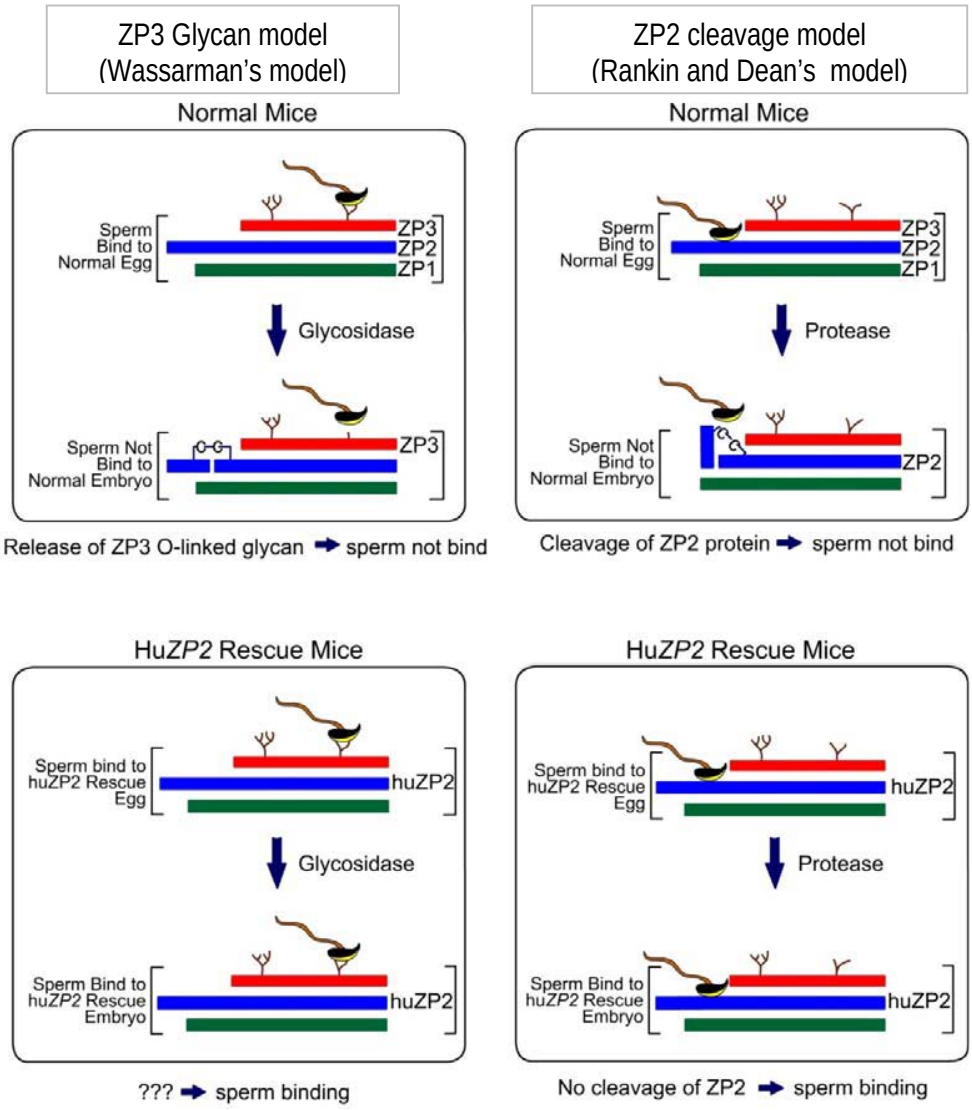


Fig. 1.8. Diagrammatic representation of the two opposing models of fertilization and ZP2 cleaving post fertilization. On the left is the model according to the Wassarman group and on the right is the model proposed by Rankin and Dean (Redrawn from Dean 2003; Hoodbhoy & Dean 2004).

In addition to the transgenic mouse experiments, mass spectrometry has been performed on mouse ZP glycoproteins, and the presence of *O*-linked oligosaccharides within the region encoded by exon 7 of m*Zp3* could not be detected (Boja *et al.* 2003). It was concluded that these findings were inconsistent with the model whereby serine residues within the exon 7 coding region were essential for primary sperm-ZP binding. The Wassarman group refuted the significance of these results (Williams *et al.* 2006), suggesting that ZP3 may only be glycosylated on the outer surface of the ZP and the relatively small amounts of *O*-linked oligosaccharides may account for the lack of detection.

Recently, the Rankin and Dean group extended their supramolecular model, renaming it the 'zona scaffold' model (Baibakov *et al.* 2007). This revised model proposed that it is the cleavage status of ZP2 that regulates the three dimensional structure of the ZP matrix, making it either permissive (ZP2-intact) or refractory (ZP2-cleaved) for sperm binding. Furthermore, after experimenting with polycarbonate filters where sperm were induced to undergo the AR when passing through pores less than 1.2 μm, the 'scaffold model' was extended to also include the three dimensional structure of the ZP regulating the AR. It was reasoned that initial penetration of sperm into the matrix triggered a 'mechanosensory' signal that was sufficient for AR induction. Hence, the zona scaffold model of sperm-ZP binding and induction of the AR is not only carbohydrate independent, it is inconsistent with the ligand-receptor model of the Wassarman group.

1.2.6 Concluding comments on sperm-ZP binding models

The transgenic mouse experiments have indeed cast doubt over the role of ZP3. These experiments have used intact ZP, and have viewed fertilization from the aspect of the whole ZP matrix *in vivo*. Nevertheless, they have not been able to conclusively show that ZP3 is not directly involved in sperm-ZP binding. It also raises doubts about the involvement of glycans in this process but has again not conclusively shown that glycans are not involved. The ZP matrix contains a considerable amount of sugar and, while extensive investigations have been carried out on what type of oligosaccharide structures comprise the ZP, it is still not known how these sugars influence the folding of the glycoproteins or the three-dimensional structure of the matrix. Sugars may influence how the sperm recognise the ZP. The failure of human sperm to bind to humanized ZP matrix could still be due to differential glycosylation despite the evidence that huZP2 and huZP3 retain their molecular mass even when expressed in a mouse background (Rankin *et al.* 2003). This could be due to the mouse glycosylation machinery possibly changing the order and configuration of the oligosaccharides, although mass spectrometry of both mouse and human ZP glycoproteins show nearly identical *O*-glycan patterns

(Dell *et al.* 2003). The recently discovered fourth human ZP glycoprotein may influence the ability of human sperm to bind to human ZP (Lefievre *et al.* 2004, Conner *et al.* 2005). ZP3 functions within a complex glycosylated matrix and the introduction of a fourth glycoprotein may change the matrix so that binding sites normally presented to the sperm are hidden when ZP4 is missing.

Absent from the literature is the obvious experiment which would possibly settle the controversy over models. No results from transgenic mouse lines lacking exon 7 from ZP3 have been published. Rankin and colleagues used in support of their model a publication of the Wassarman group where transgenic mice lacking linkage sites for *O*-linked oligosaccharides remained fertile, citing Liu *et al.*1995 (Rankin *et al.* 2003). However, these cited experiments only produced 'mosaic' mice whereby up to 50% of oocytes in these mice expressed the mutated inactive ZP3 (lacking serine residues within the exon 7 coding region). When the mutated ZP3 was purified it did not have any sperm binding abilities. The authors of these papers (including Wassarman) concluded only that the numbers of functional sperm receptors in the ZP can be reduced by more than 50% without affecting fertilization *in vivo* (Liu *et al.* 1995). It is puzzling why transgenic mice lines with the exon 7 coding region of mZP3 deleted have not been established.

Wassarman has refuted the claims of Rankin and colleagues in two papers published in 2005 and again recently in 2006 (Wassarman 2005; Wassarman *et al.* 2005; Williams *et al.* 2006). He reaffirmed his hypothesis that mZP3 is the glycoprotein responsible for providing the sugars in both sperm-binding and the acrosome reaction, and cited the large literature in support of this. He also pointed out that glycosylation differences may account for the failure of human sperm to bind to the humanized mouse ZP. More importantly though, he pointed out the failure of Rankin and colleague to establish whether the sperm bound to the ZP of fertilized eggs were acrosome-reacted. He considered that the sperm bound to the zona could have been acrosome reacted sperm bound to ZP2, which had failed to cleave, and therefore was permissive for continual binding. In this scenario there is no evidence to suggest, he

claimed, that sugars attached to ZP3 were not available for primary sperm-ZP binding. The Rankin and Dean group recently published data indicating that the sperm that remained attached to the zona after fertilization were acrosome reacted (Baibakov *et al.* 2007).

It is difficult to find common ground between the two opposing models for primary sperm-ZP binding. It is clear that mZP3 is an important glycoprotein in the zona pellucida as removal prohibits formation of the matrix itself. Certainly, in a purified state, there is sufficient evidence to conclude that mZP3 is the only glycoprotein that can bind sperm and induce the acrosome reaction. However, Wassarman has not been able to demonstrate that mZP3 has this role when embedded in a matrix. Alternatively, Rankin and colleagues have not been able to definitively find that mZP3, either the protein backbone or linked sugars, does not bind sperm.

The evidence accumulated over the past fifteen years appears to support the proposal that the region encoded by exon 7 of ZP3 is involved in primary sperm-ZP binding in the mouse. Due to the high level of conservation of the majority of ZP3, exon 7 would appear to be an ideal region to study the evolution of this gene with the implication that as this region is involved in sperm-ZP binding in the mouse, at least, amino acid substitutions may provide further information on its role. If the primary structure of the region encoded by exon 7 of *Zp3* is involved in determining species specificity then it could be predicted that the changes to the primary structure may be seen even between closely related species, or between those species that occur sympatrically, in order to limit the potential for hybridisation.

1.2.7 Species specificity of sperm-ZP binding

One often claimed function of the zona pellucida is to prevent interspecies fertilization. This is said to occur when sperm from one species will not recognize and bind to the zona matrix of an oocyte from a different species (O'Rand 1988). Support for this proposal can be found in those experiments whereby removing the zona removes the barrier to interspecies fertilization (Yanagimachi 1978). It has been suggested that co-evolution of gametes may occur when changes to, for example, proteins on the sperm

surface drive changes to the glycoproteins that comprise the zona pellucida (Bedford 1991). In externally fertilizing animals, such as free spawning aquatic species, the egg coats may well act as a barrier to cross-fertilization in order to prevent hybridization and limit wastage of oocytes (Vacquier 1998). However, in internally fertilizing animals, such as mammals, the zona pellucida as an interspecies fertilization barrier would probably not rank very high in a hierarchy of isolating mechanisms (O'Rand 1988). Other mechanisms, such as behavioural, anatomical or chemical barriers may prevent sperm from one species ever entering the ampulla of the oviduct of another species (O'Rand 1988).

The evidence for primary sperm-ZP binding being a species specific event comes primarily from *in vitro* experiments where heterologous gametes are combined. However, an analysis of the results of such experiments fails to provide unambiguous support for the claim of general species specificity of mammalian sperm-ZP binding. Although it is stated, by Wassarman, that there is ample evidence for the existence of a species specific ZP barrier (Wassarman, 1999a; Wassarman, 1999b), there is actually not much support for this in the literature.

Considerable research was conducted on sperm-ZP binding on rodents in the 1970's, mostly with zona-free eggs and provided reasonably consistent results. Mouse sperm were able to bind to the oolemma of rat, hamster or rabbit oocytes but rat, hamster or guinea pig sperm were not able to bind to the mouse oolemma (Hanada & Chang 1972, 1976, 1978; Yanagimachi 1972: Yanagimachi *et al.* 1976). The only truly promiscuous oolemma was that of the hamster which could be penetrated by the sperm of six other species (Hanada & Chang 1978). Guinea pig sperm could not bind to the oolemma of any other species. The authors of these papers concluded that generally there was no specificity seen at the membrane level (Hanada & Chang 1972, 1976, 1978). Yanagimachi is often quoted as a reference when other authors discuss species specificity of sperm-ZP binding and he is the researcher who first used zona-free hamster eggs to test the fertilizing capacity of human sperm (Yanagimachi *et al.* 1976).

However, in 1972 he used hamster eggs and guinea pig sperm and found that both the ZP and the oolemma exhibited strong specificity (Yanagimachi 1972).

Zona-intact sperm-ZP binding experiments are not supportive of the existence of the ZP barrier. All experiments listed below have employed various washing techniques to remove loosely bound sperm from the ZP. In the 1970's, it was found that hamster sperm did not bind to the ZP of rat or mouse (Hartmann *et al.* 1972) nor did mouse sperm bind to the ZP of hamster eggs (Hartmann & Hutchison 1974). However, in 1980 it was found that both hamster and mouse sperm bound to the ZP of both hamster and mouse oocytes in equal amounts whereas guinea pig sperm could not bind to the ZP of hamster or mouse oocytes (Schmell &Gulyas 1980). Rabbit sperm could also bind to the ZP of pig, mouse, human but human and guinea pig sperm could not bind to ZP of any of the other species (Swenson & Dunbar 1982). Sperm from two species of *Mus* could not bind to the other's ZP (Lambert 1984). Another experiment showed that sperm from a ram (ovine) was able to bind to the ZP of rat oocytes (Fournier-Delpech *et al.* 1982), and sperm from the domestic cat was able to bind to the ZP of the cheetah (Donoghue *et al.* 1992). There was no cross-binding of sperm to the ZP in the Chinese and Syrian hamsters (Roldan & Yanagimachi 1989), but the ZP from two species of Macaque monkeys showed no specificity although sperm from either monkey could not bind to hamster oocytes (Vandevoort *et al.* 1992). Human sperm were found to bind to gorilla ZP only and yet sperm from the horse and marmoset bound to the ZP of human salt-stored oocytes (Liu *et al.* 1991; Lanzendorf *et al.* 1992).

There was no specificity found at the zona level between bovine and ram (Slavik *et al.* 1990) and in a series of experiments using several *Bos* species (including endangered antelopes), it was found that while there was no interspecies barrier at the ZP (McHugh & Rutledge 1998; Roth *et al.* 1998; Sinowatz *et al.* 2003), the oolemma presented the major barrier (Kouba *et al.* 2001).

From this literature, a few clear points emerge. Within the order of Rodentia, mouse sperm will bind to the zona pellucida and the oolemma of most other species. However, the sperm from many species, including the rat, will not bind to the ZP of the mouse or, after removal of the ZP, to the oolemma. Human sperm will only bind to human or gorilla ZP, and guinea pig sperm appears to be able to bind only to guinea pig ZP. Hamster ZP and the oolemma can bind the sperm of a number of species, excluding the guinea pig and human, but removing the ZP allows only human sperm and not guinea pig sperm to bind to, and fuse with, hamster oocytes and enter the cytoplasm. With the limited experimental work that has been done on human oocytes, there appears to be little specificity at the ZP level. There also appears to be little or no specificity at the ZP level in larger mammals, such as primates or *Bos* species.

It would appear, therefore, that species specificity is not a general rule in primary sperm-ZP binding but does exist when gametes from some species are placed together in culture. There is a paucity of evidence relating to heterologous sperm-ZP binding in rodents other than the mouse, rat, hamster and guinea pig. While a few species of Australian endemic *Rattus* have produced hybrid offspring in the laboratory, it is not known if heterologous combinations of gametes between other species of Australian murine rodents are possible. Although there are laboratory bred species available, such as *Notomys alexis* (Spinifex hopping mouse) and *Pseudomys australis* (Plains rat), super-ovulation of these species in order to harvest enough oocytes to perform repeat sperm-ZP assays is difficult.

Despite the lack of evidence, the claim that sperm-ZP binding is species-, taxa- or order specific continues to be made. However, species specificity has only been tested on a limited number of, usually, laboratory or domesticated species. It is not known how widespread species specificity of sperm-ZP binding is or, more importantly, how it occurs. If species specificity does exist between some species then a signal of positive selection may be detected in the amino acid sequence of one of the ZP genes.

1.2.8 Positive selection

In support of the proposition that sequence divergence within the coding region of exon 7 of *Zp3* contributes to species specific binding of sperm to the zona pellucida, Wassarman has referred to evidence that this region may be evolving under positive selection (Williams *et al.* 2003, Wassarman *et al.* 2004b; Williams *et al.* 2006).

Nucleotide substitutions, or mutations, that reduce the fitness of an organism are deleterious and are selected against, and do not become fixed within the population. This type of selection is called purifying selection. On the other hand, positive selection occurs when the mutation increases the fitness of the organism and is fixed within the population. Those mutations that neither add nor detract from the organism's fitness are due to neutral evolution (Graur & Li 2000).

One method to detect the presence of positive selection is to consider the ratio of amino acid changing substitutions (nonsynonymous) to silent or nonreplacing substitutions (synonymous) within a gene among a group of species. Where at a given loci, there are more nonsynonymous substitutions (d_N) to synonymous substitutions (d_S) the loci is considered to be evolving under positive selection ($d_N/d_S > 1$). If the d_N/d_S ratio is = 1 then the loci is considered to be evolving under neutral evolution and if the ratio is less than 1, purifying selection is assumed (Yang 1998). Generally, when the d_N/d_S value is averaged across the whole of the gene, the value is invariably less than 1, possibly due to the structural constraints placed on the protein products (Yang 1998). Reproductive proteins, both male and female, are thought to evolve at a higher rate than other genes, due to the role they may play in speciation events, sperm competition or sexual selection (Wyckoff *et al.* 2000). However, when the method of averaging the d_N/d_S ratio was applied, only a few male reproductive proteins showed a strong signal (Wyckoff *et al.* 2000). When the ratio of d_N/d_S was averaged across the whole of the *Zp3* gene the d_N/d_S value was less than 1 (0.40) suggesting that *Zp3* is evolving under purifying selection (Wyckoff *et al.* 2000).

A method of detecting positive selection within genes, at the codon level, was developed in 1998 (Nielsen & Yang 1998). This codon-substitution method, using maximum likelihood and Bayesian statistics (posterior probabilities), is able to detect positive selection ($d_N/d_S = \omega > 1$) at only a few codon sites, even if the gene as a whole is evolving under purifying selection (see Chapter 2.10.2 for theory).

The maximum likelihood method of detecting adaptive evolution has been used to analyse the *Zp3* coding sequence among a range of mammals: i.e. mouse, rat, marmoset, bonnet monkey, dog, cat and pig (Swanson *et al.* 2001). It was found that while the gene, as a whole, was evolving under purifying selection ($\omega = 0.27$), a number of codon sites (7.6%) were evolving under positive selection ($\omega = 1.7$). Of those sites identified, seven were within the exon 7 coding region, although only two sites had posterior probabilities of > 0.90. Swanson *et al.* concluded that, as this region is involved in species-specific sperm-ZP binding, the selective pressure for *Zp3* to evolve and adapt may relate to sperm-egg interactions (Swanson *et al.* 2001; Swanson & Vacquier 2002).

ZP3 was again the subject of maximum likelihood analysis to test for positive selection between closely related species of the genus of *Mus* (Jansa *et al.* 2003). The coding sequence of exon 6 and 7 of *Zp3* from 13 species of *Mus* and 2 more distantly related species was determined. The analysis used *Rattus norvegicus* as an outgroup. Several amino acid sites were found to be evolving under positive selection ($\omega = 2.4$), with three sites having a posterior probability of > 0.95. Two of these three sites were located within the exon 7 coding region although they were not the same as those sites identified by Swanson *et al.* (2001) (Jansa *et al.* 2003). It was speculated that changes in residues near the sperm-binding region may affect glycosylating patterns of the serine residues within this region and consequently may affect interspecific sperm recognition (Jansa *et al.* 2003).

Both the Swanson *et al.* (2001) and Jansa *et al.* (2003) conclusions have been subsequently questioned in two studies: one published in 2005 (Berlin & Smith) and other in 2006 (Turner & Hoekstra). Berlin and

Smith (2005) reanalyzed the mammalian data of Swanson *et al.* (2001), using the maximum likelihood method but with different parameters (those that had been implemented in the PAML program since 2001). The analysis included the same eight sequences used by Swanson *et al.* (2001) plus sequences from seven additional mammalian species: i.e. cow, rabbit, fox, ferret, lemming, Brandt's vole and hamster. Contrary to the results of Swanson *et al.* (2001), Berlin and Smith found no convincing evidence of positive selection. They claimed that the group of models (M7-M8) used were unreliable indicators of positive selection as they can generate false positives (Berlin & Smith 2005). A fact not noticed by either Swanson *et al.* (2001) or Berlin and Smith (2005) was that the human sequence used in both studies was not the correct *Zp3* human gene as it contained a single nucleotide insertion within exon 8 that resulted in a frame shift, truncating the protein prematurely (NCBI accession number X56777: Van Duin *et al*. 1992, Kipersztok *et al*. 1994). Those sites within the exon 8 coding region, found by Swanson *et al.* (2001) to be evolving under positive selection, were directly attributable to the error using this sequence for their analyses.

Turner and Hoekstra (2006) investigated the evolution of exon 6 and 7 of *Zp3* among the genus of a North American cricetid rodent, *Peromyscus*. Using the models M1a/M2a, M7/M8 and M8/M8a (see Chapter 2.10.2 for details of various models), the authors found that three sites were under positive selection ($\omega = 2.09$), but one (mZP3-346) had a posterior probability of greater than 0.95. They also found that there was *intra*-species differences with exon 6 and 7 as well as *inter*-species differences. Turner and Hoekstra also reanalysed the Jansa *et al.* (2003) data to see if there was still evidence for positive selection when the *Rattus* outgroup was excluded. With the *Rattus* outgroup removed from the data set, no support for positive selection was found. The authors concluded that, contrary to the *Peromyscus* genus, there was no evidence for positive selection occurring in exon 7 within the *Mus* genus (Turner & Hoekstra 2006).

It would appear, therefore, that there is conflicting evidence as to whether or not positive selection has occurred within the exon 7 coding region among mammalian groups other than within the genus *Peromyscus*. There is now no strong evidence that it has occurred within the family Muridae, of which *Mus* and *Rattus* are members (subfamily: Murinae). It is possible that positive selection has occurred in the lineage leading from *Rattus* to *Mus*, but this has not been tested.

Other than the Jansa *et al.* (2003) study on the *Mus* genus, and the Turner and Hoekstra (2006) study there has been no study of other murid rodent species and especially no other study on the highly speciose sub-family of Murinae, of which the laboratory mouse and rat are members. There is evidence that the genomes of species within this subfamily evolve more rapidly than in other mammalian groups (Berry & Scriven 2005), and it would appear to be an ideal group of mammals to study the evolution of the exon 7 coding region of *Zp3*. In particular, in sampling from a broad range of species, a signal for positive selection could be detected where it may not be when comparing species that are either too closely or too distantly related. The sub-family Murinae consists of a large group of rodent species indigenous to Africa, Europe, Middle East, North Eurasia, Indomalayan region, Sri Lanka, Taiwan, Japan, Philippine Islands and archipelagos of the Sunda shelf, Australia and New Guinea. This study has limited the range of species to those endemic to Africa, South-east Asia, Indonesia, Malaysia, New Guinea and Australia. The following section details the biogeographical history and phylogenetic relationships of these species.

1.2.9 Murine rodent species

The oldest murine fossil remains have been found in the Siwalik beds in Northern Pakistan (Watts & Baverstock 1994a). Remains of the now extinct *Antemus* have been dated at between 15 and 16 millions years and, as no obvious ancestor has been found, it is presumed that *Antemus* evolved elsewhere and migrated to Pakistan (Watts & Baverstock 1995). In sediment dated at around 12 million years, *Progonomys* fossils have been found and it is thought that this species possibly evolved from

Mus, as did *Apodemus* (Watts & Baverstock 1995). Around 10 million years ago *Karnimata* appears in the fossil record and it is possible that this lineage gave rise to *Rattus* and the African Arvicanthis lineage that includes *Aethomys*. Several lineages, therefore, evolved in Northern Pakistan, including at least some of the *Mus*, *Apodemus*, African and South-east Asian clades (Watts & Baverstock 1995). Furthermore, ancestors of the South-east Asian and Australasian murines possibly reached South-east Asia around 8 to 10 million years ago and there began a further period of rapid speciation (Watts & Baverstock 1994a).

Australia and New Guinea have a large number of murine species, many of which evolved in either New Guinea or Australia. Based on fossil evidence, morphological and chromosomal factors, it has been hypothesized that murines colonized New Guinea during a period of lowered sea levels and emergent land around 6 to 8 million years ago when New Guinea was a collection of islands (Aplin 2005). These ancestral murines possibly entered New Guinea via a northern route from Sulawesi through the Moluccas (Watts & Baverstock 1992). Steppan *et al.* (2005) found a close relationship between Philippine Old Endemic murines (such as *Apomys* and *Archboldomys*) and the Australian/New Guinea Old Endemic murines, suggesting a relatively recent dispersal route between these two areas. Once in New Guinea, these murines underwent an initial radiation from around 5-6 million years ago, moving from island to island, occupying diverse ecological niches, and gave rise to the *Lorentzimys* and *Pogonomys* group of murines (Old Endemic New Guinean murines) as well as to the stem lineages of *Melomys*, *Uromys* and *Xeromys*. It has been postulated (Aplin 2005) that the ancestral Australian murine probably entered Australia around 5 million years ago, giving rise to six major lineages: *Xeromys* and *Uromys* lineages, *Hydromys*, and three groups subsumed into the *Pseudomys* division (Musser & Carleton 2005). Possibly the explosive radiation of the later group originated from a single ancestor, occurring in an increasingly arid/dry environment as demonstrated by low diversity of rain-forest adapted species and the abundance of semi-arid adapted species (Aplin 2005). Around 1.5 million years ago a

two-way process of faunal exchange commenced during glacial periods, when lower sea levels exposed land across theTorres Strait providing avenues of savannah and rainforest for migration from New Guinea to Australia. Thus species from the *Rattus*, *Uromys*, *Hydromys* and *Xeromys* divisions probably crossed into Australia from New Guinea at this time (Aplin 2005). Furthermore, two species of murines believed to have evolved in Australia (*Pseudomys delicatulus* and *Conilurus penicillatus*) are now found in the South west of New Guinea (Watts & Baverstock 1995).

1.2.9.1 Taxonomic groupings of Murine rodents

Murine taxonomy has undergone frequent revision in recent years due to emergent technologies such as DNA sequencing. In 1995, using microcomplement fixation of albumin data, Watts and Baverstock classified murine rodents into the following clades: New Guinea clade comprising of what have been referred to as the New Guinea Old Endemics, including *Macruromys*, *Pogonomys*, *Anisomys*, *Chiruromys* and *Mallomys*; an African clade including *Rhabdomys*, *Praomys*, *Hylomyscus*, *Aethomys* and *Lemniscomys*; a *Dasymys* clade comprising only one genus; a *Mus* clade consisting solely of the *Mus* genus; an *Apodemus* clade consisting solely of the *Apodemus* genus; an Australasian clade comprising a large group of species endemic to both Australia and New Guinea, excluding those species placed in the New Guinea clade (Australasian Old Endemics); a South-east Asian clade consisting of, but not solely, *Rattus*, *Bandicota*, *Bunomys*, *Leopoldomys*, *Maxomys*, *Niviventer* and *Paruromys*.

Steppan *et al.* (Steppan *et al.* 2005), using DNA sequences from one mitochondrial and three nuclear genes, developed a phylogeny that largely agreed with the Watts and Baverstock clade of South-east Asian murines. However, in contrast to Watts and Baverstock, they combined all Australian and New Guinea species into one clade that they referred to as the Australian/New Guinea clade or the Australo-Papuan clade, combining both the Old Endemics and the Watts and Baverstock's Australasian clade into one. However, their sampling of the New Guinea species was restricted to *Anisomys* and *Hyomys* and did not include a number of genera such as *Crossomys* or *Leptomys*. Their phylogeny also split the

African clade into two clades: one comprising *Mastomys* and *Hylomyscus*, that had radiated relatively recently, and an older clade comprising of *Lemniscomys*, *Rhabdomys* and *Aethomys* (*Dasymys* was not sampled). Steppan *et al.*, it should be noted, hesitated to formalize a taxonomy as they considered more work needed to be done (Steppan *et al.* 2005).

Musser and Carleton (2005), in their extensive revision of murine taxonomy, proposed a number of divisions as set out in Table 1.2. As this is the most recent, and extensive, analysis of the murine taxonomy, it has been adopted in this study. Briefly, Musser and Carleton (2005) divided the African clade into four divisions: Dasymys, Aethomys, Arvicanthis (*Lemniscomys* and *Rhabdomys*, among others) and Stenocephalemys (*Hylomyscus* and *Mastomys*, among others). *Mus* and *Apodemus* were placed in their own divisions. The South-east Asian clade has been divided into the Dacnomys division (*Leopoldamys* and *Niviventer,* among others), Maxomys division (comprising solely of *Maxomys*), and the Rattus division (comprising of *Bandicota*, *Bunomys*, *Paruromys* and *Rattus,* among others). They placed *Maxomys* in a division of its own, as it is considered a separate lineage not close to any other.

The genus *Rattus* contains a large number of species and Musser and Carleton (2005) suggested that, as *Rattus* was a heterogenous accumulation of species, they should be divided into six species groupings, as follows:

- *Rattus norvegicus*.
- *Rattus exulans*.
- *Rattus rattus*.
- *Rattus fuscipes*, (consisting *R. colletti*, *R. fuscipes*, *R. lutreolus*, *R. sordidus*, *R. tunneyi* and *R. villosissimus*, all of which are Australian, with one (*R. sordidus*) also occurring in New Guinea).

- *Rattus leucopus*, (consisting of species indigenous to New Guinea, such as *R*. *leucopus*, *R*. *mordax*, *R*. *niobe*, *R*. *praetor*, *R*. *steini* and *R*. *verecundus*. *R*. *leucopus* also occurs in north-east Australia, whereas *R*. *niobe* and *R*. *verecundus* have previously been placed in the Stenomys genus).
- *Rattus xanthurus*, (containing a species that is native to Sulawesi).

A seventh group contained *Rattus* species that have unresolved phylogenetic affinities, of which none are included in the present study.

Musser and Carleton (2005) placed all the New Guinea Old Endemics (Watts and Baverstock's New Guinean clade) into the Pogonomys division, with the exception of *Lorentzimys* which was placed in a division of its own. They divided the Australasian clade into the Hydromys and the Xeromys divisions (species that evolved in New Guinea and predominantly occur on that land mass with the exception of *Hydromys chrysogaster* and *Xeromys myoides* which are also present in Australia), the Uromys division (species that probably evolved in New Guinea but several also occur in both Australia and New Guinea), and the Pseudomys division (species that evolved, or predominately occur, in Australia) (Table 1.2).

Table 1.2. Divisions of murine genera according to Musser and Carleton (2005), adopted in this study as being the most recent taxonomic analysis. Only those genera used in this study are included.

DIVISION	GENERA	DIVISION	GENERA
Aethomys	*Aethomys*	Pseudomys	*Conilurus*
	Micaelamys		*Leggadina*
			Leporillus
Apodemus	*Apodemus*		*Mastacomys*
			Mesembriomys
Arvicanthis	*Lemniscomys*		*Notomys*
	Rhabdomys		*Pseudomys*
			Zyzomys
Dacnomys	*Leopoldamys*		
	Niviventer	Rattus	*Bandicota*
			Bunomys
Dasymys	*Dasymys*		*Paruromys*
			Rattus
Hydromys	*Crossomys*		
	Hydromys	Stenocephalemys	*Hylomyscus*
	Parahydromys		*Mastomys*
Lorentzimys	*Lorentzimys*	Uromys	*Melomys*
			Paramelomys
Maxomys	*Maxomys*		*Solomys*
			Uromys
Mus	*Mus*		
		Xeromys	*Leptomys*
Pogonomys	*Anisomys*		*Pseudohydromys*
	Chiruromys		*Xeromys*
	Coccymys		
	Hyomys		
	Macruromys		
	Mallomys		
	Mammelomys		
	Pogonomelomys		
	Pogonomys		

1.2.9.2 Phylogenetics

The most recent hypothesis for the phylogeny of Old World murines was published by Steppan *et al.* (2005), using nucleotide sequences from nuclear and mitochondrial genes. Steppan *et al.* sampled a broad range of murines and proposed a number of phylogenies derived from different DNA sequences. However, *Steppan et al.* only used one nuclear gene (AP5) when sampling most of the Australian genera (excluding *Hydromys* and *Melomys*) (Fig. 1.9). Furthermore, only two New Guinea species were sampled. Hence, the combined phylogram is not representative of all genera. Watts and Baverstock (1994) used microcomplement fixation of albumin data (MCFA) to propose a phylogeny for New Guinea Old Endemic species (Fig. 1.10) and their relationship to Australian and other murines (Watts & Baverstock 1995) (Fig. 1.11). To date this is the only published phylogeny on the New Guinea murines. Using nuclear and mitochondrial data, Ford (2006) has recently proposed a phylogeny for the Australian Old Endemic murines (excluding *Rattus* and New Guinea species) (Fig. 1.12).

1.2.9.2.1 African murines

Watts and Baverstock (1995) place all African genera within one clade. On the other hand, Steppan *et al.* (2005) separate the African genera into two groups, with a long intervening branch between the two groups, as did LeCompte *et al.* (2005). In one group are the genera *Mastomys* and *Hylomyscus* (Stenocephalemys division) and the other group consists of *Lemnisomys* (Arvicanthis division) and *Aethomys* (Aethomys division).

1.2.9.2.2 South-east Asian murines

Watts and Baverstock (1994) placed *Rattus* into a large South-east Asian clade containing, among others, *Maxomys*, *Leopoldomys*, *Niviventer*, *Bunomys*, *Bandicota* and *Paruromys*. Steppan *et al.* (2005) did not sample all of the above species but placed *Maxomys*, *Niviventer* and *Rattus* into the South-east Asian clade.

1.2.9.2.3 New Guinea species

The only recent phylogenetic tree that has been proposed for species that occur predominately in New Guinea is by Watts and Baverstock (1994), based on MCFA data. Steppan *et al.* (2005) used only two species (*Anisomys* and *Hyomys*). Watts and Baverstock's phylogeny split the New Guinea occurring species into three groups: those species that occur predominately in New Guinea, those species that occur in Australian, New Guinea and other islands, and *Lorentzimys*. Steppan *et al.* (2005) did not sample *Lorentzimys* which is unfortunate as it appears to be difficult to place this genus into context of other New Guinea species.

1.2.9.2.4 Australasian group

The Watts and Baverstock MCFA phylogeny (1995), Steppan *et al.* AP5 phylogeny (2005) and Ford's (2006) generic tree of the phylogenetic relationships of the Australian Old Endemic murines (2006) (Fig. 1.13) all agree in respect of the monophyletic grouping of *Pseudomys* (including *Mastacomys*) and *Notomys*, and *Conilurus*, *Mesembriomys* and *Leporillus*. The genus *Uromys* (with species that occur in both Australia and New Guinea) either forms a monophyletic group with *Melomys* (Ford 2006) or with *Conilurus* (Steppan *et al.* 2005).

A problematic genus has been *Leggadina*. Watts and Baverstock (1992, 1995) placed *Leggadina* at the base of the Australian radiation, proposing that this genus is more closely related to *Hydromys* than to other Australian murines. However, in contrast, Steppan *et al.* (2005) placed *Leggadina* with the *Pseudomys*/*Notomys* group as did Ford (2006). Musser and Carleton (2005) placed *Leggadina* within the large division of Pseudomys acknowledging the difficulty in resolving *Leggadina's* relationships to other Australian Old Endemic murines. The genus *Zyzomys* has also proven to be difficult to place. Watts and Baverstock (1992) placed *Zyzomys* within the *Conilurus*, *Mesembriomys* and *Leporillus* group while Steppan *et al.* (2005) placed *Zyzomys* with the *Pseudomys/Notomys/Leggadina* group. Ford (2006) placed *Zyzomys* between *Pseudomys* and *Leporillus*.

NOTE: This figure is included on page 37 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.9. Most recently proposed phylogeny based on the nucleotide sequence from the AP5 gene of the African, South-east Asian and Australian/New Guinean clades (redrawn from Steppan *et al.* 2005).

NOTE: This figure is included on page 38 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.10. Proposed phylogeny of the New Guinean Old Endemic murines, based on microcomplement fixation of albumin data (redrawn from Watts & Baverstock 1994b).

NOTE: This figure is included on page 39 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.11. Proposed phylogeny of the African, South-east Asian and Asian, New Guinean and Australasian Old Endemic murine species based on microcomplement fixation of albumin (redrawn from Watts & Baverstock 1995).

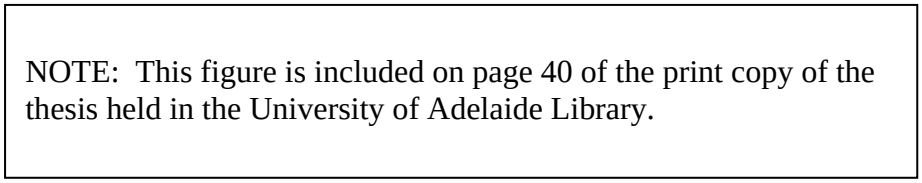


Fig. 1.12. Proposed phylogeny of Australian Old Endemic murine species based on mitochondrial nucleotide sequence data
(redrawn from Ford 2006).

NOTE: This figure is included on page 41 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 1.13 Genetic phylogeny of Australian Old Endemic murine species based on nucleotide and mitochondrial sequence data
(redrawn from Ford 2006).

1.2.10 Summary

Despite intensive studies over the last thirty years, the molecular events that take place during fertilization in mammals are still poorly understood. A considerable amount of research into these events has focused on gametes from the laboratory mouse due to the ease of obtaining oocytes and the suitability of the mouse for genetic manipulation. From these mouse studies a model has been proposed that, prior to penetrating the zona pellucida of oocytes, sperm bind to sugars linked to amino acids encoded by exon 7 of *Zp3* and then undergo the acrosome reaction (Kinloch *et al.* 1995; Chen *et al.* 1998). It has frequently been suggested that this process is a species specific event. The observation that there is a high level of amino acid divergence within the region encoded by exon 7 of *Zp3*, together with suggestive evidence of positive selection, has lead some researchers to propose that this region may contribute to species specificity of sperm-ZP binding (Wassarman *et al.* 1999; Wassarman 1999b; Wassarman & Litscher 2001). If species specificity of sperm-ZP binding occurs, it is likely that there would be a high level of sequence divergence of amino acid residues within the region of exon 7 that facilitate primary sperm-ZP binding. As sperm-ZP binding is likely to involve protein-carbohydrate interactions, it is possible that differential glycosylation of the zona pellucida, due to amino acid divergence, may provide the barrier to interspecific fertilization. Alternatively, there is some evidence to suggest that it is the polypeptide backbone itself that mediates sperm-ZP binding and changes to the primary sequence of ZP3 may influence the three-dimensional structure that the sperm recognise and bind to, inducing species specific sperm-ZP binding. Evidence of positive selection occurring within the exon 7 coding region of *Zp3* in North American rodents has been presented. However, no convincing evidence has been found of positive selection occurring within other groups of mammals. The present study has, as its main aim, the exploration of species differences in the sperm-ZP binding region of a large group of murine rodents (subfamily Murinae) by using DNA and protein sequence comparisons. Phylogenetic analysis, using maximum likelihood models of codon substitution,

of the region of the *Zp3* gene that may be involved in sperm-ZP binding (exon 7 coding region) has then been carried out in selected murine rodent species in order to determine if positive selection, during the evolution of murines, has occurred.

1.3 Hypotheses

The hypotheses to be tested in this study are as follows:

- That there is a high level of sequence divergence within the region encoded by exon 7 of *Zp3* between closely related species which contributes to potential species specific sperm-ZP binding.

- That the region encoded by exon 7 of *Zp3* has undergone rapid evolution due to positive selection.

1.4 Aims

The first hypothesis predicts that the amino acid sequence of the exon 7 coding region of *Zp3* between closely related species will show high levels of divergence.

Aim 1: To compare the nucleotide and predicted amino acid sequences from a broad range of murine rodents.

The second hypothesis predicts that evidence of rapid evolution, driven by positive selection, of the exon 7 coding region of *Zp3* will be detected between closely related species.

Aim 2: Use maximum likelihood phylogenetic analyses of models of codon substitution to determine whether there is evidence of positive selection within a broad range of murine rodents.

1.5 Structure of thesis

The above aims are addressed in three chapters. The results of this study, detailed in these three chapters, are then placed into a broader context by comparing the amino acid sequence, and rate of evolution, of the exon 7 coding region within other groups of mammalian species to that of the murine rodents.

Chapter 1: Background and literature review.

Fertilization and the proposed role ZP3 plays in this event was discussed together with the sequence divergence of the region encoded by exon 7 of *Zp3*, evidence of positive selection and the suggestion of species specificity. In addition, the taxonomy and biogeographic history of Old World murine rodents was discussed in the context of proposed phylogenetics that will be used in the analyses to detect evidence of positive selection.

Chapter 2: Material and Methods.

This chapter provides details on the materials and methods used in this study.

Chapter 3: Evolution of exon 6 and 7 of Zp3 within New Guinean and Australasian Old Endemic murine rodents.

The nucleotide sequence and the predicted primary amino acid structure of the region encoded by exon 7 of *Zp3* is compared across New Guinean and Australasian non-*Rattus* ('Old Endemic') murine rodent species. DNA extraction, PCR amplification and sequencing were used to determine both nucleotide and amino acid sequences. These sequences are then compared, using current hypothetical phylogenetic divisions, in order to determine the rate of evolution of this region of the *Zp3* gene. The significance of the sequence differences in relation to the proposal that species specific binding of sperm to the ZP, as determined by sequence divergence, is then discussed.

Chapter 4: Evolution of exon 6 and exon 7 of Zp3 within African, Eurasian and South-east Asian murine rodents.

The nucleotide and primary amino acid structure of the region encoded by exon 7 of *Zp3* is compared across African, Eurasian and South-East Asian murine rodent species, including several species of the genus *Rattus* that are endemic to New Guinea and Australia. DNA extraction, PCR amplification and sequencing were used to determine the nucleotide and amino acid sequence and these sequences are then compared to determine the rate and pattern of evolution. The results obtained are then compared to those from the New Guinea and Australasian Old Endemic species investigated in Chapter 3. The significance of the sequence differences in relation to the proposal, that species specific binding of sperm to the ZP as determined by sequence divergence, is discussed.

Chapter 5: Detection of positive selection occurring within exon 7 of Zp3 within murine rodents.

Phylogenetic analysis using maximum likelihood and codon substitution models, employing the software PAML (Phylogenetic Analysis using Maximum Likelihood), was applied to the region encoded by exon 6 and exon 7 of *Zp3* from selected murine species from Africa, Eurasia, South-east Asia, New Guinea and Australia, using the current phylogenetic hypotheses of species relationships. The significance of these results in relation to the hypothesis that this region has undergone rapid evolution due to positive selection is discussed.

Chapter 6: Comparison of amino acid sequences of the exon 7 coding region, and detection of positive selection, across mammalian species.

The hypothesis of positive Darwinian selection occurring within the region encoded by exon 7 of *Zp3* of mammals is tested. Amino acid sequences of ZP3, available from GenBank, from a large and diverse range of mammalian species are compared and discussed. Previous studies of positive selection using the software PAML (Phylogenetic Analysis using Maximum Likelihood) are analysed and repeated using additional sequences, updated algorithms and software. The significance of the evidence of positive

selection is discussed in relation to the proposal that high levels of divergence seen with exon 7 of *Zp*3 may contribute to species specificity of sperm-ZP binding.

Chapter 7: General discussion

The results from this study are summarized and their significance to the hypotheses of sperm-ZP binding, in particular species specificity and the controversy over the role of ZP3, are discussed. The results detailed in this thesis will hopefully provide some further understanding of the evolution of the region encoded by exon 7 of *Zp3* and whether the amino acid sequence has any relevance to species specificity of sperm-ZP binding in these animals.

Chapter 2

Materials and Methods

Image on reverse: Australian Old Endemic rodent, *Pseudomys shortridgei.*
Image modified from the private collection of Assoc. Prof. Bill Breed.

Chapter 2

Materials and Methods

2.1 Extraction of DNA

2.1.1 Numbering of samples

Each PCR tube was numbered consecutively, and any subsequent PCR sequencing reaction retained this original number. Hence, for example, PCR tube 35 became sequencing reaction Z35, with forward reaction (Z35F) or reverse reaction (Z35R).

2.1.2. Source of DNA

The majority of tissue used to extract DNA was obtained from the Australian Biological Tissue Collection (ABTC) maintained by the South Australian Museum (SAM). DNA was also provided by either the University of Pretoria, South Africa (*Aethomys* and *Micaelamys* spp.), or from Fred Ford, James Cook University (specified in Table 2.1). All DNA used in this study, and its source, is presented in Table 2.1. Tissue provided by the ABTC was either frozen or stored in alcohol.

Table 2.1. List of all species from which DNA has been extracted in this study. Included in the table are the PCR sequence number, genus and species names, South Australian Museum tube name (where relevant), voucher number and Australian Biological Tissue Collection (ABTC) number, source, extraction date and the location where the specimen was collected.

Seq No.	Genus and species name	Tube Name	Voucher No.	ABTC No.	Source of DNA	Extraction date	Location
Z321	*Aethomys chrysophilus*	Ai3			Sth Africa	30/09/2005	South Africa
Z198	*Aethomys ineptus*	Ai1			Sth Africa	24/02/2005	Sth Africa
Z293	*Aethomys ineptus*	Ai3			Sth Africa	24/06/2005	South Africa
Z322	*Aethomys ineptus*	Ai3			Sth Africa	30/09/2005	South Africa
Z326	*Aethomys ineptus*		D0999	65829	SAM	30/09/2005	40K N Pretoria South Africa
Z199	*Micaelamys namaquensis*		An1		Sth Africa	24/02/2005	Sth Africa
Z263	*Anisomys imitator*		AMSM13770	42758	SAM	07/04/2005	Sol River PNG
Z320	*Apodemus chevrieri*		F-E4 A2	13966	SAM	30/09/2005	Yunnan Prov, China
Z324	*Bandicota indica*		F-E40 HS15	69090	SAM	30/09/2005	Hatsua Village Laos
Z167	*Bunomys andrewsi*	BP8			SAM		
Z148	*Chiruromys vates*		AMSM14662	43096	SAM	16/12/2003	Yuro PNG
Z311	*Chiruromys vates*		AMSM18590	45611	SAM	30/09/2005	Bobole SHP

Seq No.	Genus and species name	Tube Name	Voucher No.	ABTC No.	Source of DNA	Extraction date	Location
Z319	*Coccymys ruemmleri*		AMSM16745	47187	SAM	30/09/2005	Sol River PNG
Z74	*Conilurus penicillatus*	# 4		07431	SAM	12/12/1997	No location
Z133	*Crossomys moncktoni*		AMSM17098	46614	SAM	20/10/2003	Bobole
Z165	*Dasymys incomtus*			65735	SAM	16/12/2003	Africa
Z131	*Hydromys chrysogaster*			07432	SAM	20/10/2003	29k SE Innisfail
Z145	*Hyomys goliath*		AMSM13707	42697	SAM	16/12/2003	Ofekaman PNG
Z312	*Hyomys goliath*		AMSM18487	45613	SAM	30/09/2005	Bobole SHP
Z115	*Leggadina forresti*	41838	SAMAM18344	41838	SAM		2 KAQ Pipalyatjara
Z142	*Leggadina lakedownensis*			07441	SAM	20/10/2003	Qld
Z285	*Lemniscomys griselda*		D1005	65835	SAM	24/06/2005	40K N Pretoria South Africa
Z286	*Leopoldamys edwardsi*		D3084	67574	SAM	24/06/2005	Nam Trang Vietnam
Z287	*Leopoldamys sabanus*		D3082	67572	SAM	24/06/2005	Kayon Mentarang Res Indonesia
Z60	*Leporillus conditor*	lepC1		13331	Fred Ford		Franklin Island West
Z107	*Leptomys elegans*	2 A4			SAM		
Z146	*Lorentzimys nouhuysi*		AMSM13679	42732	SAM	16/12/2003	Sol River PNG
Z310	*Lorentzimys nouhuysi*		AMSM17060	45251	SAM	30/09/2005	Bobole SHP
Z150	*Macruromys major*		AMSM14730	43909	SAM	16/12/2003	Mt Karimui PNG
Z161	*Mallomys aroensis*	L14	AMSM17893	45750	SAM		Bobole
Z307	*Mallomys aroaensis*		AMSM16693	45196	SAM	30/09/2005	Bobole SHP
Z160	*Mallomys rothschildi*		AMSM17362	47402	SAM	16/12/2003	Miptigin PNG
Z308	*Mallomys rothschildi*		AMSM17096	45199	SAM	30/09/2005	Bobole SHP
Z153	*Mallomys rothschildi**		AMSM17362	47402	SAM	16/12/2003	Miptigin PNG
Z179	*Mammelomys lanosus*	D39	AMSM17804	64875	SAM		S of Tifalmin W Sepik Prov
Z315	*Mammelomys lanosus*		AMSM16759	47208	SAM	30/09/2005	Sol River PNG
Z181	*Mammelomys rattoides*	Z253	AMSM15863	44170	SAM		Wigote
Z316	*Mammelomys rattoides*		AMSM16778	47282	SAM	30/09/2005	Munbil SP PNG
Z57	*Mastacomys fuscus*	maf 1	SAMAM10928	07354	Fred Ford		Kosciuszko NP
Z289	*Mastomys natalensis*		D1012	65842	SAM	24/06/2005	40K N Pretoria South Africa
Z265	*Maxomys bartelsii*		AMSM17388	48059	SAM	07/04/2005	Cibodas Forest Indonesia
Z288	*Maxomys hellwaldii*		D0930	65760	SAM	24/06/2005	Tangoa Sulawesi
Z151	*Mayermys ellermani*		AMSAM14827	43919	SAM	16/12/2003	Mt Karimui PNG
Z144	*Melomys burtoni*		WAMM21590	08239	SAM	25/06/2003	Mitchell Plateau
Z129	*Melomys capensis*		SAMAM11379	08349	SAM	20/10/2003	Captain Billy Creek
Z130	*Melomys cervinipes*			08336	SAM	20/10/2003	Wyong, Olney State Forest
Z143	*Melomys rubicola*		SAMAM13256	08416	SAM	20/10/2003	Bramble Cay
Z43	*Melomys rufescens*	A13	No voucher		SAM	23/07/1997	
Z138	*Mesembriomys gouldii*			07449	SAM	20/10/2003	1 mile S Mareeba
Z137	*Mesembriomys macrurus*			07451	SAM	20/10/2003	No location
Z72	*Mesembriomys macrurus**	A51		18112	SAM	16/09/1994	Mitchell Plateau
Z109	*Mesembriomys macrurus**	A51		18112	SAM	16/09/1994	Mitchell Plateau
Z264	*Niviventer fulvescens*		AMSM17394	48010	SAM	07/04/2005	Cibodas Forest Indonesia
ZP1/ZP2	*Notomys alexis*				B.Breed	26/11/2002	captive breed
ZP7/ZP8	*Notomys aquilo*		AMS32455	18253	SAM		Groote Eylandt, 3 ks
Z76	*Notomys cervinus*	16888	SAMAM16888	27132	SAM	20/05/1998	Emerald Rr mouth Sandringham Stn
ZP5/ZP6	*Notomys fuscus*	NT 5368	SAMAM17506	34069	SAM		Montecollina Bore
Z56	*Notomys mitchellii*	nm1	SAMAM18497	27066	Fred Ford		Konettra Downs SA
Z152	*Parahydromys asper*		AMSM17099	17099	SAM	20/10/2003	Magidobo PNG
Z296	*Paramelomys levipes*		AMSM12648	42506	SAM	24/06/2005	Near Kosipe Mt Albert Edward PNG
Z314	*Paramelomys levipes*		AMSM14722	46692	SAM	30/09/2005	Namosado SHP
Z147	*Paramelomys lorentzii*		AMSM13654	42748	SAM	16/12/2003	Sol River PNG
Z306	*Paramelomys lorentzii*			8391	SAM	30/09/2005	Kaironk MAP PNG
Z41	*Paramelomys platyops*	Vo4	AMSM16361	46657	SAM	17/03/1994	Namosado
Z313	*Paramelomys platyops*		AMSM24961	46655	SAM	30/09/2005	Namosado SHP
Z71	*Paramelomys rubex*	Q05	AMSM16179	46290	SAM	30/03/1993	Namosado
Z309	*Paramelomys rubex*		AMSM16172	45204	SAM	30/09/2005	Bobole SHP

Seq No.	Genus and species name	Tube Name	Voucher No.	ABTC No.	Source of DNA	Extraction date	Location
Z196	*Paruromys dominator*			65763	SAM	24/02/2005	Mt Nokilalaki
Z154	*Pogonomelomys mayeri*		AMSM16747	16747	SAM	16/12/2003	Telefomin PNG
Z113	*Pogonomelomys mayeri**	FA441	AMSM16747	47403	SAM	30/05/1994	Telefomin PNG
Z149	*Pogonomys macrourus*		AMSM15142	43144	SAM	16/12/2003	Yuro PNG
Z136	*Pseudomys albocinereus*			08165	SAM	20/10/2003	no data
Z45	*Pseudomys apodemoides*	ap1	SAMAM19345	37703	Fred Ford		Wanneroo WA
Z30	*Pseudomys australis*	NP5377	SAMAM17980	34769	SAM		Billakalina Stn
Z208	*Pseudomys bolami*		SAMAM12491	21998	SAM	24/02/2005	Coralbignie Outstn Gawler Ra SA
Z47	*Pseudomys calabyi*	c1			Fred Ford		
Z48	*Pseudomys chapmani*	40577	SAMAM40577		Fred Ford		
Z67/Z68	*Pseudomys delicatulus*	d26	NTMU4712	30596	Fred Ford		Wickham River NT
Z120b	*Pseudomys desertor*			08143	SAM	26/06/2003	Alice Springs NT
Z123	*Pseudomys fieldi*			08146	SAM	26/06/2003	Bernier Island
Z50	*Pseudomys fumeus*	f5	SAMAM13689	08046	Fred Ford		Inlet Coastal East Gippsland
Z135	*Pseudomys gracilicaudatus*			08031	SAM	20/10/2003	Townsville
Z65	*Pseudomys hermannsburgensis*	h18	WAMM29451	62062	Fred Ford		Woodstock WA
Z128	*Pseudomys higginsi*			08139	SAM	20/10/2003	captive breed
Z69	*Pseudomys johnsoni*	J1			Fred Ford		Helen Springs NT
Z117	*Pseudomys laborifex*	8172		08172	SAM		Mitchell Plateau
Z121	*Pseudomys nanus*		WAMM21751	08119	SAM	26/06/2003	
Z52	*Pseudomys novaehollandiae*	no5			Fred Ford		Loch Sport, Vic
Z122	*Pseudomys occidentalis*			08144	SAM	26/06/2003	17K NE Berdering
Z53	*Pseudomys oralis*	O2			Fred Ford		Billilimbra Qld
Z118	*Pseudomys patrius*	32211	QMJM11271	32211	SAM		Kilkivan Qld
Z127	*Pseudomys pilligaenis*			18120	SAM	20/10/2003	Pilliga Scrub
Z70	*Pseudomys shortridgei*	S1	M45	08145	Fred Ford		Nolokaton
Z327	*Rattus colletti*		M131	8530	SAM	30/09/2005	
Z262	*Rattus exulans*		AMSM12652	42509	SAM	07/04/2005	Kosipe Mission CP PNG
Z162	*Rattus fuscipes*		SAMAM11156	18144	SAM	26/06/2003	N Neptune I. SA.
Z185	*Rattus leucopus*				SAM		
Z183	*Rattus lutreolus*	P50		51763	SAM		Derwent Valley, TAS
Z212	*Rattus mordax*		AMSM19055	48962	SAM	24/02/2005	Mokopo PNG
Z214	*Rattus niobe*		AMSM12872	42489	SAM	24/02/2005	Neon Basin, Mt Albert Edward PNG
Z211	*Rattus praetor*		AMSM17461	47272	SAM	24/02/2005	Munbil PNG
Z260	*Rattus sordidus*		NTMU1503	41160	SAM	07/04/2005	Sir Edward Pellew Island
Z210	*Rattus steini*		AMSM18609	49258	SAM	24/02/2005	Bundi PNG
Z259	*Rattus tunneyi*		NTMU1446	41159	SAM	07/04/2005	Sir Edward Pellew Island
Z213	*Rattus verecundus*		AMSM16326	49292	SAM	24/02/2005	Bundi PNG
Z258	*Rattus villosissimus*		NTMU1488	41151	SAM	07/04/2005	Sir Edward Pellew Island
Z290	*Rhabdomys pumilio*		D1001	65831	SAM	24/06/2005	40K N Pretoria South Africa
Z87	*Solomys salebrosus*	LA70			SAM	10/05/1993	
Z318	*Solomys salebrosus*		AMSM22280	50537	SAM	30/09/2005	Levaleva Choiseul Is., Soloman Islands
Z91	*Uromys anak*	FZ62	AMSM15854	44256	SAM	01/06/1993	Nong River, PNG
Z63	*Uromys caudimaculatus*	Uc3			Fred Ford		Paluma Dam Qld
Z120a	*Xeromys myoides*	7458		07458	SAM		Goyder River Arnhem Land
Z62	*Zyzomys argurus*	Za22	NTMU4168	29460	Fred Ford		Litchfield NP, NT
Z140	*Zyzomys maini*			08030	SAM	20/10/2003	Nourlangie Rock
Z209	*Zyzomys palatilis*			30744	SAM	24/02/2005	Moonlight Gorge, Woologorang Stn NT
Z158	*Zyzomys pedunculatus*	no voucher		65820	SAM	16/12/2003	W Macdonnell Ranges
Z139	*Zyzomys woodwardi*		WAMM21644	07938	SAM	20/10/2003	Mitchell Plateau

2.1.3 Nomenclature

In relation to the nomenclature of ZP genes, the National Center for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov) has recently determined that the numerical system of naming ZP genes should be used (Conner *et al.* 2005) and therefore, to reduce confusion and introduce consistency, this nomenclature of ZP genes will be followed in this thesis.

In relation to species nomenclature, due to the large number of sequences that have been produced in this study, it was necessary to allocate species into divisions. The placing of murine species into divisions is a recent taxonomy (Musser & Carleton 2005) and as that authority is also used by NCBI GenBank, it has been followed in this study. However, at times it has been useful to refer to groups of species or divisions by a common name. Hence, the New Guinean divisions of Pogonomys and Lorentzimys have been grouped together as the New Guinean clade or New Guinean Old Endemics, and the divisions of Hydromys, Xeromys, Uromys and Pseudomys as the Australasian clade or Australasian Old Endemics. Tables 2.2 and 2.3 summarise the various groupings used to classify these species and the nomenclature adopted in this study.

Table 2.2. Proposed nomenclature for murine taxonomy used in this study (right column).

Watts & Aslin 1981	Watts & Baverstock 1995	Steppan *et al.* 2005	Musser & Carleton 2005 Divisions	*Genus*	Proposed Nomenclature
New Guinea Old Endemics	New Guinea clade	Australo-Papuan (clade B)	Pogonomys	*Macruromys* *Pogonomys* *Anisomys* *Chiruromys* *Mallomys* *Coccymys* *Hyomys*	New Guinean clade or New Guinean Old Endemics
			Lorentzimys	*Lorentzimys*	
Australasian Old Endemics Or Hydromyinae	Australasian clade	Australo-Papuan (clade B)	Hydromys	*Parahydromys* *Hydromys* *Crossomys*	Australasian clade or Australasian Old Endemics
			Xeromys	*Xeromys* *Leptomys* *Pseudohydromys*	
			Uromys	*Uromys* *Solomys* *Melomys*	
			Pseudomys	*Leggadina* *Pseudomys* *Notomys* *Conilurus* *Mesembriomys* *Leporillus* *Zyzomys* *Mastacomys*	Australian clade or species or Australian Old Endemics
	South-east Asian clade	South-east Asia plus (clade A)	Rattus (see table 2.4)	*Rattus* *Bandicota* *Paruromys* *Bunomys*	South-east Asian species, clade or divisions
			Dacnomys	*Leopoldamys* *Niviventer*	
			Maxomys	*Maxomys*	
	African clade	African (clade C)	Aethomys	*Aethomys* *Micaelamys*	African species, clade or divisions
			Arvicanthis	*Rhabdomys* *Lemniscomys*	
		(clade G)	Stenocephalemys	*Mastomys* *Hylomyscus*	
	Dasymys clade		Dasymys	*Dasymys*	African clade

Watts & Aslin 1981	Watts & Baverstock 1995	Steppan *et al.* 2005	Musser & Carleton 2005 Divisions	*Genus*	Proposed Nomenclature
	Mus clade	Eurasia (clade F)	Mus	*Mus*	Mus
	Apodemus clade	Palearctic (clade E)	Apodemus	*Apodemus*	Apodemus

Table 2.3. Nomenclature for the *Rattus* species used in this study (right column), based on Musser and Carleton 2005.

Watts & Aslin 1981	Watts & Baverstock 1995	Steppan *et al.* 2005	Musser and Carleton 2005	Species names	Occurring	Proposed Nomenclature
	South-east Asian clade	South-east Asian clade	R. norvegicus species group	*R. norvegicus*	Eurasia	RattusN species group
	South-east Asian clade	South-east Asian clade	R. exulans species group	*R. exulans*	Pacific	RattusE species group
	South-east Asian clade	South-east Asian clade	R. leucopus species group	*R. leucopus* *R. mordax* *R. niobe* *R. praetor* *R. steini* *R. verecundus*	New Guinea (*R.leucopus* also in Australia)	RattusL species group
Australian *Rattus* (Murinae)	South-east Asian clade New Endemics (?)	South-east Asian clade	R. fuscipes species group	*R. colletti* *R. fuscipes* *R. lutreolus* *R. sordidus* *R. tunneyi* *R. villosissimus*	Australia	RattusF species group

2.1.4. Method of extraction of DNA

DNA was extracted from frozen tissue or alcohol–stored tissue from the ABTC using a PureGene DNA Extraction Kit (Gentra, MN) following manufacturers instructions with modification, as follows:

1. Frozen tissue was placed in a 1.5 ml eppendorf tube (microcentifuge tube), and 300 µl Cell Lysis Solution was added. The tissue in solution was then homogenized thoroughly using a microcentrifuge tube pestle.

2. When the tissue was thoroughly homogenized, 1.5 µl Proteinase K Solution was added to the lysate, which was mixed by inverting 25 times.

3. The solution was then incubated overnight at 55^{o} C in a water bath.

4. The following day, 1.5 µl RNAse Solution (or 0.75 µl of 10 mg/ml) was added to the cell lysate, mixed by inverting 25 times and incubated at 37^{o}C for 30 minutes in a heating block. The sample was then allowed to cool to room temperature. This step was not performed for all DNA extractions, and did not adversely affect the outcome of the extraction.

5. 100 µl of Protein Precipitation Solution was added to the lysate and vortexed at high speed for 20 seconds. The solution was placed on ice for 5 minutes then centrifuged at around 12,000 rpm for 3 minutes. A tight protein pellet should then have been visible. If not, the solution was vortexed again then re-centrifuged.

6. The supernatant was then poured into a clean 1.5 ml tube containing 300 µl of 100% isopropanol. This was mixed by gently inverting approximately 50 times and placed in a –20^{o}C freezer overnight. At this stage, some solutions were left for a longer period not exceeding three days.

7. After the solution had been frozen for more than 12 hours, it was centrifuged at approx 12,000 rpm for 15 minutes. The supernatant was removed by pipetting, avoiding the pellet. Once all

supernatant had been removed, 300 µl of 70% ethanol was added and the tube was inverted several times in order to wash the DNA pellet. This solution was then centrifuged at 12,000 rpm for 5 minutes.

8. The supernatant was micropipetted carefully, avoiding the pellet of DNA which may have been loose. The pellet was then allowed to air dry for 3 to 4 hours on the bench top. Once air dry, 50 µl of DNA Hydration Solution was added, and rehydrated overnight.

If the tissue had been stored in alcohol, then prior to DNA extraction, the following steps were taken:

9. The tissue, within the alcohol, was chopped into smaller pieces, if necessary. The alcohol was then removed and the tissue was washed in 0.5 ml of 10mM TRIS. The tube was then spun for a few minutes.

10. The TRIS was removed and a further 0.5ml of TRIS was added. This was then spun again. The washing and spinning was repeated another two times.

11. The TRIS was removed after the final spinning, and DNA was extracted from the tissue following the same protocol as that for frozen tissue.

The concentration of DNA was not quantified. Prior to PCR, the rehydrated DNA was initially diluted 1/100 with distilled, DNase free water, and then used immediately. If no DNA was detected after PCR then the DNA was diluted 1/50, 1/10 and 1/5 in subsequent reactions.

2.1.5. Design of primers

Primers were designed based on the *Zp*3 cDNA sequence of *Notomys alexis* (accession number AY078054) and *Rattus norvegicus* (accession number D78482). As the sequence for intron 5 and 7 were not available for *Notomys alexis*, the 5' end of exon 6 and the 3' end of exon 7 were selected as sites for primer design. Figure 2.1 provides a diagrammatic representation of the location of the primer sequences relative to the *Zp3* gene.

For all New Guinean and Australasian Old Endemic species, the following primers were used:

- Forward: 5' ACC TGC CAT CTC AAA GTC GC 3' (G510)
- Reverse: 5' TGC GGT TTC GAG AGG TTA GC 3' (G511)

For South-east Asian (including *Rattus*) and African species, these primers did not work. As the forward primer sequence was identical to that of *R. norvegicus*, the reverse primer was designed based solely on the *R. norvegicus* exon 7 sequence (accession number D78482).

- Reverse: 5' TGC GGT TTC GAG AAA CTA GC 3' (G693)

All primers were purchased through GeneWorks (Adelaide) and were delivered in concentrated form.

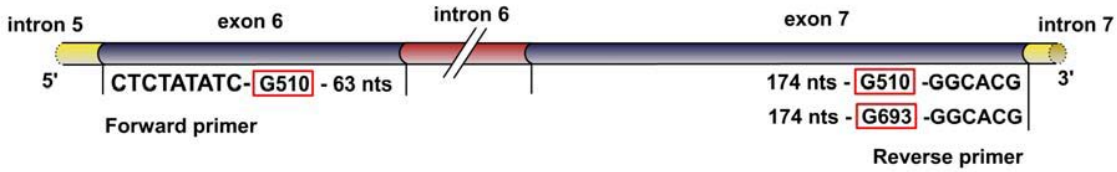


Fig. 2.1. Diagrammatic representation of the position of the primers relative to the *Zp3* gene. The red boxes contain the identification number of the primer (SAM).

2.2 PCR amplification

PCR was conducted in standard 25 µl reactions as set out in Table 2.4:

Table 2.4. Reagents and concentrations used in each PCR reaction.

Reagent	*Concentration*	*Aliquot*	*final concentration*
MgCl	25 mM	4 µl	4 mM
TGold Buffer	10x	2.5 µl	1 x
dNTP	10 mM	2 µl	0.8 mM
dH2O		11.9 µl	
AmpliTaq Gold		0.1 µl	0.6 U
Primers (x2)	5 µM each	2 µl	5 pmole each
DNA		2.5 µl	
Total		25 µl	

All PCR reagents (AmpliTaq Gold, buffer, MgCl) were purchased from Applied Biosystems, California, USA.

PCR amplifications were carried out on either a Hybaid Omni Thermocycler (Hybaid, Middlesex, UK), or an Eppendorf Mastercycler (Eppendorf, New York). Each reaction was carried out at:

- 95^{o} C for 9 minutes.
- 34 cycles of:
  - 94^{o} C for 45 seconds.
  - 55^{o} C for 45 seconds.
  - 72^{o} C for 45 seconds.
- Final elongation cycle of 72^{o}C for 6 minutes.

Each reaction contained a negative control (no DNA) in order to test for contaminated reagents.

Gel electrophoresis was performed, using agarose gel of 1.5% w/v concentration, at 100 milliAmps for 45 minutes. All gels were stained with ethidium bromide and visualised under UV light. Those reactions that resulted in a clear band on the gel, at ~ 300 bp in size, were selected for PCR sequencing.

2.2.1. Clean up of PCR products

PCR products were purified using the UltraClean PCR Clean Up Kit (MoBio Laboratories, Solana Beach, California) following the manufacturer's instructions.

2.3. Cycle-sequencing reactions

The product of each PCR, after clean up, was added to two sequencing reactions using the forward and reverse primers.

PCR sequencing reactions were conducted using Big Dye Terminator Sequencing reagents (Applied Biosystems, California, USA). Big Dye versions changed during the course of this study and Table 2.5 details the varying amounts of reagents.

Table 2.5. Amounts of reagents used in Cycle-sequencing reactions.

Reactions	*Reagents*	*Amounts*
ZP1/2 to Z124	Big Dye version 3.0	3 µl
	dH2O	3 µl
	PCR products	3 µl
	Primer (either forward or reverse)	1 µl
	Total	10 µl
Z127 to Z158	Big Dye version 3.1	2 µl
	Big Dye Buffer	1 µl
	dH2O	11 µl
	PCR products	5 µl
	Primer (either forward or reverse)	1 µl
	Total	20 µl
Z160 to Z327	Big Dye Version 3.1	2 µl
	Big Dye Buffer	6 µl
	dH2O	6 µl
	PCR product	5 µl
	Primer (either forward or reverse)	1 µl
	Total	20 µl

Cycle-sequencing reactions were carried out on either a Hybaid Omni Thermocycler (Hybaid, Middlesex, UK), or an Eppendorf Mastercycler (Eppendorf, New York). Each reaction was carried out at:

- 36 cycles at:
  - 96^{o} C for 30 seconds.
  - 50^{o} C for 15 seconds.
  - 60^{o} C for 4 minutes.

2.3.1. PCR sequencing clean up

An isopropanol precipitation method was employed to precipitate the sequencing products prior to sequencing. The following is the protocol that was used for a 20 µl reaction:

1. 80 µl of 75% isopropanol was added to each PCR tube.
2. The contents were mixed gently, then left at room temperature for 15 minutes at least.
3. The tubes were then centrifuged for 20 min at 11000 rpm.
4. All the supernatant was immediately aspirated, avoiding the pellet.
5. 200 µl of 75% isopropanol was then added.
6. The tubes were centrifuged for 5 minutes at 11000 rpm.
7. All the supernatant was carefully, and completely, aspirated.
8. The PCR sequencing samples were then left to dry for 1 minute (no longer) on a heating block at 90^{o} C or air dried on bench protected from the light.
9. The samples could then be stored at –20^{o} C prior to sequencing.

The sequence fragments were separated by electrophoresis on an ABI 3700 Capillary DNA Sequencer situated at the Institute of Medical and Veterinary Science (IMVS) Adelaide.

2.4. DNA sequencing

The IMVS produced sequencing files in the form of chromatograms that could then be read using Chromas (shareware available from Technelysium). Chromatogram printouts were visually scanned to detect multiple peaks, then the sequence was imported to GeneDoc software (Nicholas *et al.* 1997), and the forward and reverse sequences were then synchronized.

In situations where the sequencing was unreadable, the PCR was repeated, cleaned and re-sequenced. If this did not resolve the problem, then DNA was extracted from a different specimen and the process repeated.

2.5. Sequence alignment

Multiple sequences were imported into GeneDoc (Nicholas *et al*. 1997) and manually aligned (by eye). Exon boundaries were determined by comparing the sequences to that of the mouse and rat and by the presence of conserved intron/exon splicing signals.

2.6 Polymorphisms and insertions/deletions

Polymorphic sites (where two different nucleotides occur at a particular loci) were identified when the chromatogram for both the forward and reverse sequence clearly and unambiguously showed a double signal at the same nucleotide site. This suggested that the particular species was heterozygous at that specific loci. As sampling of more than one animal for each species was not routinely performed, the frequency at which each allele is present in a specific population is not known. For the purposes of the Phylogenetic Analysis using Maximum Likelihood (PAML) software (see 2.10.2), ambiguous sites are ignored by the program and therefore, rather than assigning the polymorphism as ambiguous, the nucleotide most common within the particular genus, group or division (depending on the particular species) was chosen. Where it was not possible to do this, the nucleotide polymorphic site was designated N.

2.7. Estimation of nucleotide sequence divergence

For the purposes of estimating nucleotide sequence divergence rates, the *Zp3* sequence from the mouse (*Mus musculus*) and rat (*Rattus norvegicus*) were used for comparison. To measure the evolutionary distances between and within divisions the Kimura 2-parameter method was used (Kimura 1980). This model allows for multiple hits at each base. It also takes into account transitional/transversional bias (that is, it allows a different purine to purine substitution rate compared to

the purine to pyrimidine substitution rate) while making the assumption that the four nucleotide frequencies are the same and the substitution rates do not vary among sites. A simple model of molecular evolution, that of Kimura 2-parameter, was preferred for estimating evolutionary distances, as the majority of distances (*d*) between sequences were very low. If *d* is less than 0.2 there is little difference between alternative methods used to estimate the rate of divergence (Nei & Kumar 2000). Furthermore, phylogentic trees based on ZP data are poorly resolved, and hence the use of software such as ModelTest (Posada & Crandall 1998) to carry out likelihood ratio tests of the most appropriate model for distance calculations could not be used reliably.

As there are 96 species within this study, it was expedient to calculate the mean evolutionary distance (*d*) within and between divisions. Pairwise distances between species were calculated using MEGA software version 3.1 (Kumar *et al.* 2004), then the mean *d* was calculated for within and between divisions. Estimations of divergence were made between and within divisions for the whole of the sequence obtained (exon 6, intron 6 and exon 7). Where insertions or deletions occurred, the alignment gap was treated as a single pairwise deletion in that the gap sequence was ignored in only those pairwise comparisons where it was present.

2.8. Estimations of amino acid sequence divergence rates

For the purposes of estimating amino acid sequence divergence rates, the *Zp3* sequence from the mouse (*Mus musculus*) and rat (*Rattus norvegicus*) were used for comparation. To measure the level of amino acid divergence, a simple *P* distance was calculated for both within and between divisions. This method calculates the number of amino acid changes between two pairs of sequences then divides this number by the total number of amino acid sites compared. The mean *P* distance was then calculated for within and between divisions. Only the coding regions of exon 6 and exon 7 were used, and were estimated by MEGA software version 3.1 (Kumar *et al.* 2004).

2.9. Isoelectric points and hydropathic profiles

To assess the possible effects sequence change may have on a protein, three methods have been used in the present study: potential glycosylation sites, isoelectric points and hydropathy profiles.

2.9.1 Glycosylation and isoelectric points (IEP)

O-linked glycosylation occurs where oligosaccharides are linked to either a serine or threonine amino acid present in the polypeptide. While it is a complicated and expensive exercise to investigate the actual *O*-linked glycosylation sites using methods such as mass spectrometry, an estimation of the comparative potential glycosylation sites can be obtained by calculating the relative percentage of serine and threonine residues in each sequence.

The isoelectric point of a protein is the pH at which the overall charge of a molecular is neutral. This value usually reflects the differences in the pKs (pK: the pH at which there is 50% dissociation of the amino acid ions) of the amino acid side chains. Calculating the overall charge of a molecule allows a comparison of the effect changing amino acid sequence has on the charge of the protein.

To calculate both the relative serine/threonine composition and the isoelectric points of sequences, the software WinPep (Hennig 1999: version 3.01) was used.

2.9.2 Hydropathy

One method of assessing the impact sequence change has on the functionality of a protein is to determine the sequence hydropathy profile. Kyte and Doolittle (1982) devised a method to evaluate hydrophilicity (water liking) and hydrophobicity (water avoiding) along a protein's amino acid sequence. The resultant hydropathy scale took into account the hydrophilic and hydrophobic properties of each of the amino acid's side chains (Table 2.6). A program was developed (not used in the present study) using a moving segment approach that continuously determines the average hydropathy within the segment. While this method was primarily developed to establish where membrane spanning regions of

a protein are likely to be, it still has its use in the present study, as it can highlight regions that are likely to be on the surface of the protein. Those regions that are identified as being likely to be buried inside the folded protein may not be available for molecular interaction with the fertilizing sperm. While this method cannot show what three dimensional shape ZP3 takes or how it interacts with ZP1 and ZP2 (or possibly ZP4), it is nonetheless an informative guide as to how molecular changes may affect the folding of the protein.

The hydropathy index of Kyte and Doolittle (1982) (Table 2.6) was used to determine the hydropathic attributes or profiles of the common amino acid sequences of the exon 6 and 7 coding regions of *Zp*3 within murines.

Table 2.6. Hydropathy scale, from highly hydrophobic (positive values) to highly hydrophilic (negative values).
Table taken from Kyte and Doolittle (1982).

NOTE: This table is included on page 64 of the print copy of the thesis held in the University of Adelaide Library.

The software program WinPep (Henning 1999: version 3.01) was used to calculate the hydropathic profile of each sequence and this was then imported into Microsoft Excel in order to graphically display the results. The hydropathic profile of each group of species was graphically displayed in a three dimensional format in order to highlight both the high conservation of sequence and hydropathy as well as changes.

2.10. Analyses to test for evidence of positive selection

To test for evidence of positive selection occurring during the evolution of the exon 6 and 7 coding region within murine rodents two methods were used. The first method estimated the nonsynonymous (amino acid replacing) substitutions per site (d_N) and synonymous (silent) substitutions per site (d_S), and in the second method the d_N/d_S ratio (ω) was calculated using a maximum likelihood method of codon substitution which detects positive selection occurring on individual codons, rather than averaging across the protein. If d_N is less than d_S (i.e. $d_N/d_S = \omega = < 1$), the gene is considered to be evolving under purifying selection, whereas when the value of d_N is the same as d_S (i.e. $\omega = 1$) the gene is evolving under neutral evolution. A higher d_N than d_S value ($\omega > 1$) is indicative of position selection occurring, resulting from nonsynonymous substitutions providing a fitness advantage and being fixed within the populations at a higher rate than silent mutations.

2.10.1 Estimations of nonsynonymous and synonymous substitution rates

The first method tested for evidence of positive selection by estimating the nonsynonymous (replacement) substitutions per site (d_N) and synonymous (silent) substitutions per site (d_S) within the exon 6 and exon 7 regions of *Zp3*. Estimations of d_N and d_S were calculated using the Yang & Nielsen (2000) approximation method which allows for transition and base/codon frequency biases. Pairwise comparisons were conducted using the yn00 program of the PAML (Phylogenetic Analysis by Maximum Likelihood) software package (version 3.15; Yang 1997) and mean estimations were calculated over all species, within and between the divisions categorised by Musser and Carlton (2005). The ω ratio of

d_N/d_S calculated by the yn00 program was not used, as the high percentage of d_S = 0.000 produced undefined d_N/d_S ratios (given as -1 or 99 in the PAML software, producing a warning "method is inapplicable").

2.10.2 Maximum likelihood methods for detecting evidence of positive selection

As the approximation method averages the d_N and d_S over all codons, it is not a robust method for detecting evidence of positive selection, particularly when it only occurs at a few codon positions, with the majority of sites under stabilising selection. To detect positive selection at only a few codon positions, the maximum likelihood approach developed by Nielsen & Yang (1998) was used. This method uses likelihood ratio tests (LRTs) to compare a null model that does not allow for $\omega > 1$ and an alternative model which does.

Three LRTs were applied to the data, using the 'site' models. The first pair of models compared was the nearly neutral model (M1a) and the positive selection model (M2a) (Wong *et al.* 2004). The M1a model assumes two codon site classes in proportions p_0 and $p_1 = 1 - p_0$ with $0 < \omega < 1$ and $\omega_1 = 1$. The M2a model adds a proportion (p_2) of sites with $\omega_2 > 1$ estimated from the data. The next pair of models compared was the null M7 model with the alternative model of M8. The M7 model assumes a beta distribution for ω (within the interval $0 < \omega < 1$) and the M8 model adds an extra class of sites under positive selection ($\omega_s > 1$). The third LRT compared a null model of M8 (M8a) with ω_s constrained to 1 ($\omega_s = 1$) with the M8 model with $\omega_s > 1$ (Swanson *et al*. 2003).

For each LRT, the null model is nested within the alternative (positive selection) model so that a test statistic $-2\Delta\ln L$ (where $\Delta\ln L$ is the difference in log likelihood values of the two models) follows a χ^2 distribution, with degrees of freedom (df) equal to the difference in the number of parameters between the null and alternative models. Two df for the M1a/M2a and M7/M8 LRTs and one df for the M8a/M8 LRT were used in the present study. The correct distribution of the latter test statistic is unknown and,

hence, the more conservative one df, rather than using a χ^2 distribution with a 50:50 mix of a χ_1^2 and χ^2 with point mass of zero, a suggested alternative (Wong *et al.* 2004), was chosen. All tests are considered to be conservative (Yang *et al.* 2005).

If the LRT is significant ($P < 0.05$) then positive selection can be inferred. The Bayes Empirical Bayes (BEB) method (Yang *et al.* 2005) was then used to calculate the posterior probability (PP) that each codon site is from a particular class of ω. Sites with $\omega > 1$ and with high posterior probabilities (PP > 95%) are considered to be under positive selection.

Chapter 3

Evolution of exon 6 and 7 of *Zp*3 within New Guinean and Australasian Old Endemic murine rodents.

Image on reverse: Australian Old Endemic rodent, *Notomys alexis*.
Image modified from the private collection of Assoc. Prof. Bill Breed

Chapter 3

Evolution of exon 6 and 7 of Zp3 within New Guinean and Australasian Old Endemic murine rodents.

3.1 Introduction

Mouse ZP3 (mZP3), one of three glycoproteins that assemble into the mouse zona pellucida, appears to be important for structure, primary binding of sperm to oligosaccharides on the surface of the ZP matrix and induction of the acrosome reaction (Bleil & Wassarman 1983). As stated in Chapter 1.2.4.2, within the region encoded by exon 7 of *Zp*3 is the putative combining-site for sperm containing a short stretch of serine residues (Ser-329, Ser-331 to Ser-334) purported to be essential in primary sperm-ZP binding (Kinloch *et al.* 1995). This region shows a lack of amino acid conservation relative to that of the remainder of ZP3 (Wassarman & Litscher 1995), and this observation has led to the proposal that the exon 7 coding region contributes to species specificity of sperm-ZP (Chen *et al.* 1998; Wassarman *et al.* 1999; Wassarman & Litscher 2001; Wassarman 2002; Williams *et al.* 2003; Wassarman *et al.* 2004b; Williams *et al.* 2006).

If this hypothesis is correct, it would be expected that the region encoded by exon 7 would evolve at a higher rate during speciation or have rapidly diverged in recently evolved species, where a mechanism for pre-mating isolation may be a selective advantage. However, preliminary findings early in this present study (Swann *et al.* 2002) showed an identical amino acid sequence within the region encoded by exon 7 between two closely related species of Australian Old Endemic rodents (*Notomys alexis* and *Pseudomys australis*). These early results suggested that the region played little, if any, role in species-specific sperm-ZP binding in these Australian rodents. However, it is also possible that the two species shared an ancestral sequence and may not represent the level of diversity in the exon 7 region that is

typical for members of this group of Australasian native rodent species. Recently, divergence within the exon 7 coding region was determined for members of the cricetid group of North American rodents within the genus *Peromyscus* and it was found that there was considerable sequence divergence both within and between species within this group (Turner & Hoekstra 2006).

In this chapter, the hypothesis is proposed that there is a high level of sequence divergence within the region encoded by exon 7 of *Zp*3 which contributes to potential species specificity of sperm-ZP binding. In order to test this hypothesis, the nucleotide and predicted amino acid sequence from a broad range of New Guinean and Australasian Old Endemic murine species (see Chapter 2.1.3 for nomenclature of species groups) was determined. DNA extraction and PCR amplification and sequencing was performed on a total of 68 species from six divisions (taxonomy according to Musser & Carleton 2005), and the level of nucleotide and amino acid divergence was ascertained. Current hypothetical phylogenetic relationships were used to probe the evolution of the exon 7 coding region of the *Zp*3 gene. The significance of the sequence differences in relation to the above hypothesis was then discussed.

3.2 Materials and Methods

For materials and methods see Chapter 2.

For nomenclature of species groups used in this chapter see Chapter 2.1.3. Thus the New Guinean divisions of Pogonomys and Lorentzimys are grouped together as the New Guinean clade or New Guinean Old Endemics, and the divisions of Hydromys, Xeromys, Uromys and Pseudomys are referred to collectively as Australasian clade or Australasian Old Endemics. The New Guinean and Australasian Old Endemics do not contain the endemic *Rattus* species (the New Australasian Endemics) which are placed within the South-east Asian clade from where they recently arose.

Primers used to amplify nucleotide sequences for exons 6 to 7 are listed in Chapter 2.1.5.

The exon 6, intron 6 and exon 7 *Zp*3 sequence from the laboratory mouse (*Mus musculus*, referred to from here as 'the mouse') and the laboratory rat (*Rattus norvegicus*, referred to from here as 'the rat) have been used as reference sequences.

For the method used to estimate evolutionary distances see Chapter 2.7.

Phylogenetic trees used to plot amino acid changes are those hypothesized by Watts and Baverstock (1994b) and Ford (2006) (see Figs. 3.2 and 3.3). Ancestral reconstructions for plotting amino acid changes (Yang *et al.* 1995) were computed using the codeml program of the PAML (Phylogenetic Analysis using Maximum Likelihood) software version 3.15 (Yang 1997) and displayed in Figs. 3.2 and 3.3. Also displayed are those amino acid sites that show evidence of parallel evolution on different lineages.

3.3 Results

3.3.1 PCR sequencing

Sequencing was performed only when the PCR products, after gel electrophoresis, ethidium bromide staining and visualization under UV light, showed a clear band of approximately 300 base pairs (bp).

PCR amplification products of DNA from several species produced a secondary faint band on the agarose after gel electrophoresis and, when these fragments were sequenced (after repeat PCR failed to remove the faint double band), a double signal appeared on the chromatogram (see Chapter 2.4). The 5' end of the forward sequence was clear, with double sequence starting from approximately a TTTTGT sequence within intron 6. The reverse sequence showed a clean signal from this loci to the 3' end. It was therefore possible to obtain the full sequence by combining both sequences. The double signal appeared to be as a result of the species being heterozygous for either an insertion or deletion within intron 6. Table 3.1 provides a list of those species where this occurred.

Table 3.1 . List of those sequences where the species was heterozygous for an indel within intron 6, resulting in a double signal on the chromatogram.

Species name	PCR seq No.	ABTC No	Result
Melomys burtoni	Z144	08239	PCR repeated twice. 1 nt insertion in intron 6
Mesembriomys macrurus	Z72 Z137	18112 07451	Two extractions, only Z72 had a 2 nt insertion within intron 6
Pseudohydromys ellermani	Z151	43919	1 nt insertion within intron 6
Mallomys aroensis	Z161 Z307	45750 45196	Two extractions, both heterozygous for 5 nt deletion within intron 6

To ensure accuracy of DNA extraction and PCR sequencing, DNA was extracted from two members of several species (see Table 3.2 for identification of those species).

Two extractions of DNA were performed for *Paramelomys lorentzi* (ABTC 42748 [Z147], and ABTC 08391 [Z306]), and the PCR sequencing produced a sequence very similar (99%) to that of *Mammelomys lanosus*. These two sequences shared only 92% identity with that of *Paramelomys rubex*.

It is possible that the ABTC specimens of *P. lorentzi* have been misidentified (Flannery 1990) and therefore the sequencing results from this species have been excluded from the present study.

All sequences have been lodged with GenBank (accession numbers provided in Table 3.2) at the National Center for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov). For the full nucleotide sequence data see Appendix 1.

Table 3.2. List of New Guinean and Australasian species investigated in this Chapter, showing division, genus, species and common names, and including the ABTC and GenBank accession numbers. The asterisk in the accession number column represents those specimens where repeat sequences were determined but not lodged with GenBank.

Division	Genus	Species	Common Name	ABTC Number	Accession Number
Lorentzimys					
	Lorentzimys	*nouhuysi*	Long-footed Tree Mouse	42732	EF364459
	Lorentzimys	*nouhuysi*	Long-footed Tree Mouse	45251	*
Pogonomys					
	Anisomys	*imitator*	Uneven-toothed Rat	13770	EF364448
	Chiruromys	*vates*	Lesser Chiruromys	43096	EF364449
				45611	*
	Coccymys	*ruemmleri*	Rummler's Coccymys	47187	EF364450
	Hyomys	*goliath*	Eastern Hyomys	42697	EF364454
				45613	*
	Macruromys	*major*	Greater Macruromys	43909	EF364460
	Mallomys	*aroaensis*	De Vis's Mallomys	45750	EF364461
				45196	*
	Mallomys	*rothschildi*	Rothschild's Mallomys	47402	EF364462
				45199	*
	Mammelomys	*lanosus*	Highland Mammelomys	64875	EF364463
				47208	*
	Mammelomys	*rattoides*	Lowland Mammelomys	44170	EF364464
				47282	*
	Pogonomelomys	*mayeri*	Shaw Mayer's Pogonomelomys	47403	EF364483
	Pogonomys	*macrourus*	Chestnut Pogonomys	43144	EF364484
Hydromys					
	Crossomys	*moncktoni*	Earless New Guinea Water Rat	46614	EF364452
	Hydromys	*chrysogaster*	Common Water Rat	07432	EF364453
	Parahydromys	*asper*	New Guinea Waterside Rat	45382	EF364479
Xeromys					
	Leptomys	*elegans*	Elegant Leptomys	n/a	EF364458
	Pseudohydromys	*ellermani*	Shaw Mayer's Shrew Mouse	43919	EF364466
	Xeromys	*myoides*	False Water Rat	07458	EF364510
Uromys					
	Melomys	*burtoni*	Grassland Melomys	08239	EF364467
	Melomys	*capensis*	Cape York Melomys	08349	EF364468
	Melomys	*cervinipes*	Fawn-footed Melomys	08336	EF364469
	Melomys	*rubicola*	Bramble Cay Melomys	08416	EF364470
	Melomys	*rufescens*	Black-tailed Melomys	n/a	EF364471
	Paramelomys	*levipes*	Papuan Lowland Paramelomys	42506	EF364480
				46692	*
	Paramelomys	*platyops*	Common Lowland Paramelomys	46655	EF364481
				46657	*
	Paramelomys	*rubex*	Mountain Paramelomys	46290	EF364482
				45204	*
	Solomys	*salebrosus*	Bougainville Island Solomys	50537	EF364507
	Uromys	*anak*	Black-tailed Uromys	44256	EF364508
	Uromys	*caudimaculatus*	Giant White-tailed Uromys	n/a	EF364509
Pseudomys					
	Conilurus	*penicillatus*	Brush-tailed Conilurus	07431	EF364451

Division	Genus	Species	Common Name	ABTC Number	Accession Number
	Leggadina	*forresti*	Forrest's Leggadina	41838	EF364455
	Leggadina	*lakedownensis*	Lakeland Downs Leggadina	07441	EF364456
	Leporillus	*conditor*	Greater Stick-nest Rat	13331	EF364457
	Mesembriomys	*gouldii*	Black-footed Mesembriomys	07449	EF364472
	Mesembriomys	*macrurus*	Golden-backed Mesembriomys	07451	EF364473
				18112	*
	Mastacomys	*fuscus*	Broad-toothed Mastacomys	07354	EF364465
	Notomys	*alexis*	Spinifex Hopping Mouse	n/a	EF364474
	Notomys	*aquilo*	Northern Hopping Mouse	18253	EF364475
	Notomys	*cervinus*	Fawn Hopping Mouse	27132	EF364476
	Notomys	*fuscus*	Dusky Hopping Mouse	34069	EF364477
	Notomys	*mitchellii*	Mitchell's Hopping Mouse	27066	EF364478
	Pseudomys	*albocinereus*	Ash-grey Pseudomys	08165	EF364485
	Pseudomys	*apodemoides*	Silky Pseudomys	37703	EF364486
	Pseudomys	*australis*	Plains Pseudomys	34769	EF364487
	Pseudomys	*bolami*	Bolam's Pseudomys	21998	EF364488
	Pseudomys	*calabyi*	Kakadu Pebble-mound Pseudomys	n/a	EF364489
	Pseudomys	*chapmani*	Western Pebble-mound Pseudomys	n/a	EF364490
	Pseudomys	*delicatulus*	Delicate Pseudomys	30596	EF364491
	Pseudomys	*desertor*	Desert Pseudomys	08143	EF364492
	Pseudomys	*fieldi*	Shark Bay Pseudomys	08146	EF364493
	Pseudomys	*fumeus*	Smoky Pseudomys	08046	EF364494
	Pseudomys	*gracilicaudatus*	Eastern Chestnut Pseudomys	08031	EF364495
	Pseudomys	*hermannsburgensis*	Sandy Inland Pseudomys	62062	EF364496
	Pseudomys	*higginsi*	Long-tailed Pseudomys	08139	EF364497
	Pseudomys	*johnsoni*	Central Pebble-mound Pseudomys	n/a	EF364498
	Pseudomys	*laborifex*	Kimberley Pseudomys	08172	EF364499
	Pseudomys	*nanus*	Western Chestnut Pseudomys	08119	EF364500
	Pseudomys	*novaehollandiae*	New Holland Pseudomys	n/a	EF364501
	Pseudomys	*occidentalis*	Western Pseudomys	08144	EF364502
	Pseudomys	*oralis*	Hastings River Pseudomys	n/a	EF364503
	Pseudomys	*patrius*	Eastern Pebble-mound Pseudomys	32211	EF364504
	Pseudomys	*pilligaensis*	Pilliga Pseudomys	18120	EF364505
	Pseudomys	*shortridgei*	Heath Pseudomys	08145	EF364506
	Zyzomys	*argurus*	Common Australasian Rock Rat	29460	EF364511
	Zyzomys	*maini*	Arnhem Land Rock Rat	08030	EF364512
	Zyzomys	*palatilis*	Carpentarian Rock Rat	30744	EF364513
	Zyzomys	*pedunculatus*	Central Australasian Rock Rat	65820	EF364514
	Zyzomys	*woodwardi*	Kimberley Rock Rat	07938	EF364515

3.3.2 Nucleotide sequences

The complete nucleotide sequence for each taxon is provided in Appendix 1.

Sequences varied in length from 273 nucleotides (*Pseudomys desertor* and *P. shortridgei*) to 305 nucleotides (*P. occidentalis*) which is attributable to the presence of various insertions and deletions (indels) within intron 6. The partial sequences of exon 6 and exon 7 (171 nucleotides in length) were identified by their similarity to the corresponding cDNA sequence of the mouse and rat *Zp3* gene and by the presence of conserved intron/exon splicing signals. Both exon 6 and exon 7 did not contain any indels. Exon 6 is 62 nucleotides in length and exon 7 is 109 nucleotides in length for all species.

3.3.2.1 Polymorphisms

Polymorphic sites were recognized when the chromatogram for both the forward and reverse sequence showed a double signal at the same nucleotide site, thus suggesting heterozygosity at that specific loci (see Chapter 2.6.1). Unambiguously identified polymorphisms are listed in Table 3.3.

Table 3.3 Nucleotide sequences where a double signal appeared on the chromatogram, suggesting the presence of a polymorphism. Nucleotide numbers correspond to positions of bases on the nucleotide alignment of exon 6, intron 6 and exon 7 (Appendix 1).

Species name	Nucleotide No.	Polymorphism	Affect
Exon 6			
Hydromys chrysogaster	42	G/A	TCG/TCA (S) silent
Notomys aquilo	30	C/T	AAT/AAC (N) silent
Pseudomys bolami	19	G/A	GAT (D), AAT (N) amino acid replacing
Pseudomys laborifex	45	A/C	TTA (L), TTC (F) amino acid replacing
Pseudomys laborifex	59	A/T	CAG (Q), CTG (L) amino acid replacing
Paramelomys rubex	42	A/G	TCA/TCG (S) silent
Intron 6			
Crossomys moncktoni	111	A/C	none
Coccymys ruemmleri	89	G/A	none
Hydromys chrysogaster	65	G/A	none
Hydromys chrysogaster	100	G/A	none
Leporillus conditor	179	A/G	none
Leggadina lakedownensis	173	T/G	none
Melomys capensis	100	G/C	none
Notomys aquilo	182	C/T	none
Notomys cervinus	78	C/T	none
Notomys fuscus	156	C/T	none
Pseudomys apodemoides	200	C/T	none
Pseudomys bolami	83	G/A	none
Pseudomys bolami	156	T/C	none
Pseudomys bolami	163	T/C	none
Pseudomys bolami	166	A/T	none
Pseudomys delicatulus	99	C/T	none
Pseudomys johnsoni	172	C/T	none
Pseudomys oralis	75	G/A	none
Pseudomys pilligaensis	114	C/A	none
Paramelomys rubex	171	G/T	none
Uromys anak	136	C/T	none
Zyzomys pedunculatus	147	A/G	none
Exon 7			
Melomys capensis	247	T/G	ATC (I), AGC (S) amino acid replacing
Melomys cervinipes	179	T/C	TGG (W), CGG (R) amino acid replacing
Parahydromys asper	209	A/G	TCA/TCG (S) silent
Pseudomys australis	178	A/G	TCA/TCG (S) silent
Paramelomys levipes	209	G/A	TCA/TCG (S) silent

Single letter amino acid codes: D = aspartic acid, F = phenylalanine, I = isoleucine, L = leucine, N = asparagine, Q = glutamine, R =arginine, S = serine, W = tryptophan.

Within exon 6 there were three amino acid replacing and three silent polymorphisms. *Pseudomys laborifex* had two polymorphisms within 15 nucleotides of each other. In respect of the amino acid changing polymorphisms, each of the alternative codons would have resulted in a change of the amino acid to a residue not shared by any other species investigated. There were considerably more polymorphisms within intron 6 than in either exon 6 or 7. Within exon 7 there were two amino acid

replacing polymorphisms and three silent polymorphisms. In all three silent polymorphisms, the codon involved was a serine residue. The two amino acid changing polymorphisms were at residue 325, a potential change from a serine residue (shared by the majority of species) to an isoleucine residue (shared by a range of species), and at residue 335, a potential change from a tryptophan (shared by all New Guinean and Australasian Old Endemic species) to an arginine (uncommon).

Silent substitutions do not change the amino acid residue. Within exon 6 half of the six polymorphisms were silent of which two were transitional substitutions (purine to purine), occurring in position 42. This position is quite variable among species but this is not surprising as the substitution occurs in the 3rd base of the serine codons. The third silent polymorphism also involved a transition, C/T (cytosine to thymine), and is also at the 3rd base for the codons encoding asparagine (N). Of the potential amino acid changing polymorphisms within exon 6, *Pseudomys bolami* had a G/A transition at position 19. All species investigated had a G in this position. *Pseudomys laborifex* had an A/C transversion (purine to pyrimidine) at position 45 and another transversion (A/T) at position 59. In both instances, these sites having a C in position 45 and an A in position 59, did not vary between all other species investigated.

Within exon 7, three of the five polymorphisms were silent, and all involved transitions, at the 3rd codon base encoding for the residue serine. Of the two amino acid replacing polymorphisms, *Melomys capensis* had a T/G transversion at position 247, which is a variable site with other species having either a G (other Melomys species), C or T in this position. *Melomys cervinipes* had a T/C transition at position 279 while all other species, with the exception of the mouse and the rat, have a T in this position.

3.3.3. Estimation of nucleotide sequence divergence

Estimations of nucleotide sequence divergence were calculated using the Kimura 2-parameter method (see Chapter 2.7 for method) and computed with MEGA version 3.1 software (Kumar *et al.* 2004), using the mouse and rat sequences for comparison (Tables 3.4, 3.5 and 3.7). As the mouse and rat are the single representatives of the divisions Mus and Rattus in this chapter, when comparing divisions, only Mus and Rattus are referred to.

3.3.3.1 Exon 6 of Zp3

Table 3.4. Estimated pairwise comparisons of the evolutionary distances (*d* value) within and between divisions of New Guinean and Australasian Old Endemic murine species of the partial exon 6 sequence of *Zp3.*

Overall: 0.0584

	Within	Between							
Divisions		Mus	Rattus	Lorentzimys	Pogonomys	Hydromys	Xeromys	Uromys	Pseudomys
Mus	-								
Rattus	-	0.0163							
Lorentzimys	-	0.0000	0.0163						
Pogonomys	0.0145	0.0075	0.0224	0.0075					
Hydromys	0.0220	0.0109	0.0219	0.0109	0.0175				
Xeromys	0.0109	0.0109	0.0163	0.0109	0.0165	0.0147			
Uromys	0.0176	0.0134	0.0255	0.0134	0.0206	0.0230	0.0215		
Pseudomys	0.0122	0.0068	0.0193	0.0068	0.0136	0.0157	0.0135	0.0193	

Mus musculus (Mus division) and *Rattus norvegicus* (Rattus division) diverged approximately 12 million years ago (Jaeger *et al.* 1986). For the exon 6 sequence, the estimated evolutionary divergence (*d*) rate between the rat and mouse is 0.016. Between Mus/Rattus and the other divisions, mean *d* value ranged from 0.000 between Mus and Lorentzimys to 0.0255 between Rattus and Uromys. All pairwise comparisons between Mus and the New Guinean/Australasian divisions have a *d* value below 0.02, whereas those between Rattus and the other divisions were generally higher.

All within division *d* values of the New Guinean and Australasian Old Endemic murines were less than 0.022. Between these divisions, all *d* values were below 0.023. The lowest *d* value of 0.0068 was between Lorentzimys and Pseudomys, the most distantly related divisions (Watts & Baverstock 1995), and the highest *d* value was between Hydromys and Uromys (0.0230).

3.3.3.2 Intron 6 of Zp3

Table 3.5. Estimated pairwise comparisons of the evolutionary distances (*d* value) within and between divisions of New Guinean and Australasian Old Endemic murine species of intron 6 sequence of *Zp3*.

Overall: 0.0584

	Within	Between							
Divisions		Mus	Rattus	Lorentzimys	Pogonomys	Hydromys	Xeromys	Uromys	Pseudomys
Mus	-								
Rattus	-	0.2034							
Lorentzimys	-	0.1134	0.1236						
Pogonomys	0.0457	0.1177	0.1338	0.0397					
Hydromys	0.0213	0.1610	0.1740	0.0719	0.0772				
Xeromys	0.0271	0.1561	0.1689	0.0726	0.0792	0.0250			
Uromys	0.0303	0.1239	0.1285	0.0406	0.0545	0.0690	0.0680		
Pseudomys	0.0443	0.1388	0.1402	0.0531	0.0613	0.0713	0.0714	0.0519	

The *d* values for both within and between divisions were higher than for exon 6, as would be expected for a non-coding region not subject to selective constraint. The *d* value between Mus and Rattus was 0.2034. Between Mus/Rattus and the other divisions, the *d* value ranged from 0.1388 (Mus and Pseudomys) to 0.1740 (Rattus and Hydromys). All mean *d* values were higher between Rattus and the other divisions than between Mus and those divisions, a result that was similar to that observed in exon 6. Within the New Guinean and Australasian divisions, the mean *d* values were below 0.05, and between divisions the mean *d* value ranged from 0.0250 between Hydromys and Xeromys to 0.0792 between Pogonomys and Xeromys.

These estimated evolutionary distances are conservative due, in part, to the approach these methods have to alignment gaps. Intron 6 contains numerous insertions and deletions and hence alignment gaps appear in multiple sequence alignments. In estimating pairwise evolutionary distances, the Kimura 2-parameter method among others, regards alignment gaps in two ways. Complete deletion removes the gap from all sequences in the data set and pairwise deletion removes the gap only where it occurs between each pairwise comparison. In effect, the method ignores alignment gaps and are not factored into the results. An example of this can be seen between two sibling taxa, *Pseudomys albocinereus* and *P. apodemoides*. The *d* value for the pairwise comparison between these two species is 0.0000, using

both the complete and the pairwise deletion methods. However, within intron 6, *P. albocinereus* has 2 groups of deletions, 8 and 2 nucleotides respectively, that are not shared with *P. apodemoides*. The alignment gap is ignored.

There are other deletions and insertions within intron 6 that were also ignored in estimating evolutionary distances. Table 3.6 provides a list of those species which have either an insertion and/or deletion.

Table 3.6. Species that have either a deletion or an insertion, or both, within intron 6 of *Zp3*, together with the number of nucleotides (nts) involved are listed. Position numbers correspondence to the nucleotide sequence in Appendix 1.

Species	Deletion (nts)	Insertion (nts)	Position
All species within Hydromys and Xeromys divisions & *Pseudomys fumeus*	2		84-85
Anisomy s initator	14		130-144
Hyomys goliath	4		82-85
Leptomys elegans & *Pseudomys albocinereus*	5		165-169
Mallomys aroaensis	5 (heterozygous)		169-173
Mammelomys rattoides & *Uromys caudimaculatus*		1	69
Paramelomys levipes & *P. platyops*		1	188
Paramelomys spp	3, 2		67-70, 89-90
Pogonomys macrourus	7		172-178
Pseudomys albocinereus	8		84-91
Pseudomys calabyi		1	178
Pseudomys desertor & *P. shortridgei*	28		127-158
Pseudomys hermannsburgensis	11		114-125
Pseudomys occidentalis		4	148-151
Pseudomys patrius	9		169-178
Uromys caudimaculatus	9		164-173
Zyzomys argurus	8		118-126
Zyzomys argurus & *Z. palatilis*		1	165

3.3.3.3. Exon 7 of Zp3

Table 3.7. Estimated pairwise comparisons of the evolutionary distances (*d* value) within and between divisions of New Guinean and Australasian Old Endemic murine species of the partial exon 7 sequence of *Zp3*..

Overall; 0.0724

	Within	Between							
Divisions		Mus	Rattus	Lorentzimys	Pogonomys	Hydromys	Xeromys	Uromys	Pseudomys
Mus	-								
Rattus	-	0.0875							
Lorentzimys	-	0.0677	0.1085						
Pogonomys	0.0336	0.0950	0.1259	0.0433					
Hydromys	0.0124	0.1063	0.1349	0.0545	0.0438				
Xeromys	0.0000	0.0989	0.1311	0.0477	0.0372	0.0062			
Uromys	0.0304	0.1159	0.1477	0.0561	0.0519	0.0222	0.0171		
Pseudomys	0.0208	0.1086	0.1406	0.0563	0.0524	0.0216	0.0155	0.0323	

Between Mus and Rattus the estimated evolutionary distance was 0.0875. Between Mus and the New Guinean/Australasian divisions the *d* value ranged from 0.0677 (Mus versus Lorentzimys) to 0.1159 (Mus versus Uromys). All *d* values were higher between Rattus and the other divisions, ranging from 0.1085 (versus Lorentzimys) to 0.1477 (versus Uromys).

In respect of the New Guinean/Australasian Old Endemic divisions, all mean within divisions pairwise comparisons were below 0.0336 (Pogonomys). The lowest *d* value was within the Xeromys division. Between divisions, the *d* value was below 0.0563 (Lorentzimys versus Pseudomys), and the lowest *d* value was between Hydromys and Xeromys (0.0062).

Summary of results:

All evolutionary distances between divisions were considerably lower for exon 6 than for exon 7, although intron 6 had the higher *d* values. The sequences within the New Guinean and Australasian species are more similar to that of the mouse sequence than to that of the rat. Of all the divisions, the Lorentzimys sequence is closest in evolutionary distance to the mouse.

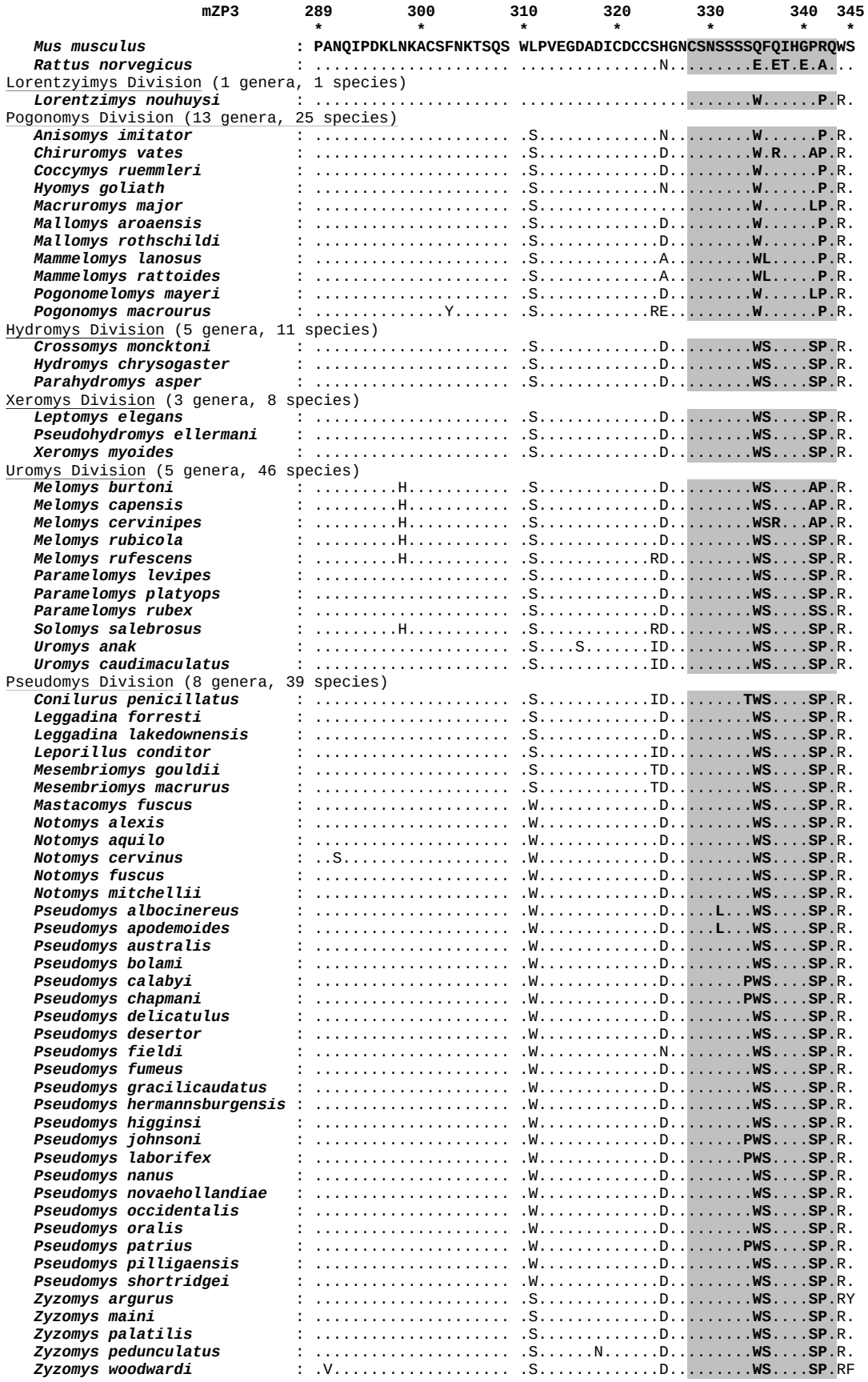
3.3.4. Estimated amino acid sequence divergence rates and sequence comparison

While the nucleotide sequence provides data on the evolutionary distances between species, it is the effect these nucleotide substitutions have on the possible function and structure of ZP3 that is most relevant to potential sperm-ZP binding. Selection, positive or otherwise, acts on the protein itself, rather than on the nucleotide sequence. To estimate the rate of divergence of the amino acid sequence, a simple *P* distance has been calculated. This method calculates the number of amino acid changes between two pairs of sequences then divides this number by the total number of amino acid sites compared. The mean P distance is then calculated for within and between divisions.

The predicted amino acid sequences for all 68 species, as well as for the mouse and rat, are shown in Fig. 3.1. Amino acid residue position numbers are the same as for mZP3. The amino acid changes have been plotted against two hypothetical phylogenetic trees (Figs. 3.2 and 3.3).

Codon 21 is shared between exon 6 and exon 7, as the exon boundary is at nucleotide positions 62-63. For analytical ease, the first nucleotide of exon 7 has been included in the nucleotides for exon 6, providing for 21 codons for exon 6 and 36 codons for exon 7.

Fig 3.1 (Opposite page) Alignment of the predicted amino acid sequence for the region encoded by the partial exon 6 and exon 7 of *Zp3* for New Guinean and Australasian Old Endemic rodent species. In parenthesis, next to each division, are the number of genera per division, and the number of extant species (according to Musser & Carleton 2005). Amino acid residues are numbered according to the corresponding mouse ZP3. Single dots (.) represent conservation of amino acid residues at any given position. The exon 6/exon 7 boundary is indicated by the space between residues 309 and 310. The putative combining-site for sperm, identified by Wassarman & Litscher (1995), is highlighted in grey with bold text. Each amino acid is represented by its single letter code.


mZP3 289 300 310 320 330 340 345
Mus musculus : PANQIPDKLNKACSFNKTSQS WLPVEGDADICDCCSHGNCSNSSSSQFQIHGPRQWS
Rattus norvegicus : ....N.....E.ET.E.A...
Lorentzyimys Division (1 genera, 1 species)
Lorentzimys nouhuysi : ....W.....P.R.
Pogonomys Division (13 genera, 25 species)
Anisomys imitator : ..... .S.....N.....W.....P.R.
Chiruromys vates : ..... .S.....D.....W.R...AP.R.
Coccymys ruemmleri : ..... .S.....D.....W.....P.R.
Hyomys goliath : ..... .S.....N.....W.....P.R.
Macruromys major : ..... .S.....W.....LP.R.
Mallomys aroaensis : ..... .S.....D.....W.....P.R.
Mallomys rothschildi : ..... .S.....D.....W.....P.R.
Mammelomys lanosus : ..... .S.....A.....WL.....P.R.
Mammelomys rattoides : ..... .S.....A.....WL.....P.R.
Pogonomelomys mayeri : ..... .S.....D.....W.....LP.R.
Pogonomys macrourus : .....Y..... .S.....RE.....W.....P.R.
Hydromys Division (5 genera, 11 species)
Crossomys moncktoni : ..... .S.....D.....WS....SP.R.
Hydromys chrysogaster : ..... .S.....D.....WS....SP.R.
Parahydromys asper : ..... .S.....D.....WS....SP.R.
Xeromys Division (3 genera, 8 species)
Leptomys elegans : ..... .S.....D.....WS....SP.R.
Pseudohydromys ellermani : ..... .S.....D.....WS....SP.R.
Xeromys myoides : ..... .S.....D.....WS....SP.R.
Uromys Division (5 genera, 46 species)
Melomys burtoni : .....H..... .S.....D.....WS....AP.R.
Melomys capensis : .....H..... .S.....D.....WS....AP.R.
Melomys cervinipes : .....H..... .S.....D.....WSR...AP.R.
Melomys rubicola : .....H..... .S.....D.....WS....SP.R.
Melomys rufescens : .....H..... .S.....RD.....WS....SP.R.
Paramelomys levipes : ..... .S.....D.....WS....SP.R.
Paramelomys platyops : ..... .S.....D.....WS....SP.R.
Paramelomys rubex : ..... .S.....D.....WS....SS.R.
Solomys salebrosus : .....H..... .S.....RD.....WS....SP.R.
Uromys anak : ..... .S....S.....ID.....WS....SP.R.
Uromys caudimaculatus : ..... .S.....ID.....WS....SP.R.
Pseudomys Division (8 genera, 39 species)
Conilurus penicillatus : ..... .S.....ID.....TWS....SP.R.
Leggadina forresti : ..... .S.....D.....WS....SP.R.
Leggadina lakedownensis : ..... .S.....D.....WS....SP.R.
Leporillus conditor : ..... .S.....ID.....WS....SP.R.
Mesembriomys gouldii : ..... .S.....TD.....WS....SP.R.
Mesembriomys macrurus : ..... .S.....TD.....WS....SP.R.
Mastacomys fuscus : ..... .W.....D.....WS....SP.R.
Notomys alexis : ..... .W.....D.....WS....SP.R.
Notomys aquilo : ..... .W.....D.....WS....SP.R.
Notomys cervinus : ..S..... .W.....D.....WS....SP.R.
Notomys fuscus : ..... .W.....D.....WS....SP.R.
Notomys mitchellii : ..... .W.....D.....WS....SP.R.
Pseudomys albocinereus : ..... .W.....D.....L...WS....SP.R.
Pseudomys apodemoides : ..... .W.....D.....L...WS....SP.R.
Pseudomys australis : ..... .W.....D.....WS....SP.R.
Pseudomys bolami : ..... .W.....D.....WS....SP.R.
Pseudomys calabyi : ..... .W.....D.....PWS....SP.R.
Pseudomys chapmani : ..... .W.....D.....PWS....SP.R.
Pseudomys delicatulus : ..... .W.....D.....WS....SP.R.
Pseudomys desertor : ..... .W.....D.....WS....SP.R.
Pseudomys fieldi : ..... .W.....N.....WS....SP.R.
Pseudomys fumeus : ..... .W.....D.....WS....SP.R.
Pseudomys gracilicaudatus : ..... .W.....D.....WS....SP.R.
Pseudomys hermannsburgensis : ..... .W.....D.....WS....SP.R.
Pseudomys higginsi : ..... .W.....D.....WS....SP.R.
Pseudomys johnsoni : ..... .W.....D.....PWS....SP.R.
Pseudomys laborifex : ..... .W.....D.....PWS....SP.R.
Pseudomys nanus : ..... .W.....D.....WS....SP.R.
Pseudomys novaehollandiae : ..... .W.....D.....WS....SP.R.
Pseudomys occidentalis : ..... .W.....D.....WS....SP.R.
Pseudomys oralis : ..... .W.....D.....WS....SP.R.
Pseudomys patrius : ..... .W.....D.....PWS....SP.R.
Pseudomys pilligaensis : ..... .W.....D.....WS....SP.R.
Pseudomys shortridgei : ..... .W.....D.....WS....SP.R.
Zyzomys argurus : ..... .S.....D.....WS....SP.RY
Zyzomys maini : ..... .S.....D.....WS....SP.R.
Zyzomys palatilis : ..... .S.....D.....WS....SP.R.
Zyzomys pedunculatus : ..... .S.....N.....D.....WS....SP.R.
Zyzomys woodwardi : .V..... .S.....D.....WS....SP.RF

3.3.4.1 Region encoded by exon 6 of Zp3

Table 3.8. Estimated pairwise comparisons of evolutionary distances (*P* value %) distances within and between New Guinean and Australasian Old Endemic divisions of the amino acid sequence encoded by the partial exon 6 sequence of *Zp3*. All results are presented as a percentage.

Overall: 1.16	Within %	Between %							
Divisions		Mus	Rattus	Lorentzimys	Pogonomys	Hydromys	Xeromys	Uromys	Pseudomys
Mus	-								
Rattus	-	0							
Lorentzimys	-	0.00	0.00						
Pogonomys	0.87	0.43	0.43	0.43					
Hydromys	0.00	0.00	0.00	0.00	0.43				
Xeromys	0.00	0.00	0.00	0.00	0.43	0.00			
Uromys	2.60	2.60	2.60	2.60	3.03	2.60	2.60		
Pseudomys	0.49	0.24	0.24	0.24	0.68	0.24	0.24	2.84	

The region encoded by exon 6 has very low divergence rates, both within and between divisional groups. Some mean pairwise *P* scores are zero, including between Mus and Rattus. This lack of divergence is reflected in the sequence comparisons where 61 out of 70 sequences (including Mus and Rattus) share exactly the same amino acid sequence. The highest level of divergence is seen within the Uromys divison (2.7%) and between this division and other groups (2.7%).

All five *Melomys* species and *Solomys salebrosus* share an asparagine (N) to histidine (H) substitution at position 298. This asparagine residue in the mouse is not a potential *N*-linked glycosylation site as it is not in the conformation NXS/T, with X being any amino acid but proline. An asparagine residue in the mouse at position 291 (also not an *N*-linked glycosylation site) has been substituted to a serine (S) residue in *Notomys cervinus*. An *N*-linked glycosylation site at position 304 (Boja *et al.* 2003) has been conserved across all 70 species.

3.3.4.2 Region encoded by exon 7 of Zp3

Table 3.9. Estimated pairwise comparisons of evolutionary distances (*P* value %) distances within and between New Guinean and Australasian Old Endemic divisions of the amino acid sequence encoded by the partial exon 7 sequence of *Zp3*. All results are presented as a percentage.

Overall: 6.79

	Within	Between							
	%	%							
Divisions		Mus	Rattus	Lorentzimys	Pogonomys	Hydromys	Xeromys	Uromys	Pseudomys
Mus	-								
Rattus	-	16.67							
Lorentzimys	-	8.33	19.44						
Pogonomys	5.40	15.40	23.23	7.07					
Hydromys	0.00	19.44	27.78	11.11	7.58				
Xeromys	0.00	19.44	27.78	11.11	7.58	0.00			
Uromys	4.34	20.96	29.04	12.88	9.09	2.53	2.53		
Pseudomys	3.30	20.51	28.77	12.18	10.63	3.13	3.13	5.53	

The amino acid sequence of the region encoded by exon 7 in this study (36 codons) finished 6 codons downstream of the predicted C-terminus of the mature secreted glycoprotein in the mouse (Boja *et al.* 2003). Within the region there are four cysteine residues (potential sites of disulfide bonds) and two *N*-linked glycosylation sites (asparagine residues in the conformation of NXS/T where X is not proline). All four cysteine residues and both *N*-linked glycosylation sites have been conserved across all 70 species.

The divergence rates for the region encoded by exon 7 are considerably higher than for the region encoded by exon 6. Within group means vary from zero (Hydromys and Xeromys) to 5.4% (Pogonomys). Between group means are highest in group comparisons between Rattus and the Australasian/New Guinea divisions (20.96% between Rattus and Uromys divisions) and lowest between the various Australasian and New Guinea divisions (2.53% between Hydromys/Xeromys and Uromys).

Amino acid sequence comparisons, in the following sections, are discussed division by division.

Lorentzimys Division

Lorentzimys nouhuysi is the only species in this division. Three amino acid changes have occurred between the ZP3 sequence from *L. nouhuysi* and the mouse. *L*. *nouhuysi* has a tryptophan (W) in position 335 (mZP3: Gln-335), a proline (P) in position 342 (mZP3: Arg-342) and an arginine (R) in position 344 (mZP3: Trp-342). None of the amino acid differences that exist between the mouse and the rat are present in the exon 7 coding region of *Zp3* of *L*. *nouhuysi* (Fig. 3.1).

Pogonomys Division

The Pogonomys division is comprised of 23 extant species (13 genera) of which the predicted amino acid sequence for 11 species from 9 of the genera has been determined. These species differ from the mouse in having Trp-335, Pro-342 and Arg-344. In addition, all species have a serine (S) in position 311 (mZP3: Leu-311). The two species of *Mammelomys* have an identical amino acid sequence to each other and two species of *Mallomys* have an identical amino acid sequence to that of *Coccymys ruemmleri*, albeit different from *Mammelomys*. *Anisomys imitator* and *Hyomys goliath* also share an identical amino acid sequence. Within the ten amino acid stretch from residue 335 to 344, 54% (6/11) of species in this division share an identical sequence with each other as well as to that of *L*. *nouhuysi,* although they differ at these sites from the mouse.

Hydromys and Xeromys Divisions

Of the 11 extant species in the Hydromys division and 8 species in the Xeromys division, the predicted amino acid sequence for three species from three genera in both of the two divisions was determined. In all six species an identical sequence was found. They all share Ser-311, Trp-335, Pro-342 and Arg-344 with species in the Pogonomys division. In addition, all six species have an aspartic acid (D) in position 325 (mZP3: His-325) and serine (S) residues in positions 336 and 341 (mZP3: Phe-336 and Pro-341).

Uromys Division

The Uromys division consists of 46 extant species from 5 genera and the predicted amino acid sequence from 11 of these species from 4 genera has been determined. All species in this division have, in common with the Hydromys and Xeromys division species, Ser-311, Asp-325, Trp-335, Ser-336, Pro-342 (with one exception being *Melomys burtoni*) and Arg-344. In addition, in common with species in the Hydromys and Xeromys divisions, all but three *Melomys* species have a serine (S) residue in position 341. Of the 23 extant *Melomys* species the amino acid sequence in three Australian species (*M. burtoni*, *M. capensis*, *M. cervinipes*), one species from New Guinea (*M. rufescens*) and the fifth, *M. rubicola*, from Bramble Cay, a Torres Strait island, have been determined. The three Australian *Melomys* species all have an alanine (A) residue in position 341, which is not present in the other two species of this genus where a serine occurs.

The genus *Paramelomys* consists of nine species, all endemic to New Guinea, of which the amino acid sequence for three species has been determined. *Paramelomys rubex* has a serine (S) in position 342, with all other species having a proline (P) in this position. *Solomys salebrosus,* which is one out of five extant *Solomys* species endemic to the Solomon Islands, has an identical amino acid sequence to *Melomys rufescens* with the two species sharing an arginine (R) in position 324, albeit that the majority of murines from the New Guinean and Australasian clades have a serine (S) in this position. Of the eight extant *Uromys* species, the amino acid sequence from two have been determined, one of which, *Uromys anak* occurs only in New Guinea and the other, *U. caudimaculatus*, occurs in both north-east Australia and New Guinea. *U. anak,* uniquely, has a serine (S) in position 316, while other species, including the mouse and rat, have a glutamic acid (E) in this position. Both *Uromys* species share the common residues with the other members of this division in addition to having an isoleucine (I) residue at position 324, rather than a serine. Within the ten amino acid stretch, from residue 335 to 344, 64% of species from the Uromys division (7/11) shared an identical amino acid sequence.

Pseudomys Division

The Pseudomys division comprises 39 extant species, at present allocated to 8 genera and the amino acid sequence from all these species has been determined. All species have retained the common Trp-335, Pro-342 and Arg-344, which are also shared with species in the other divisions, but not with mouse and rat. In addition, all species have the two additional serine (S) residues at positions 336 and 341 in common with all six species from the Xeromys and Hydromys divisions, and most species from the Uromys division. Furthermore, all but *Pseudomys fieldi* have an aspartic acid (D) in position 325, in common with all species from Xeromys, Hydromys and Uromys divisions. All species investigated in the large genus of *Pseudomys*, together with five species of *Notomys* and *Mastacomys fuscus,* have a tryptophan (W) in position 311. Other species in this division (from the genera of *Conilurus*, *Leggadina*, *Leporillus*, *Mesembriomys* and *Zyzomys*), together with species from the Lorentzimys, Pogonomys, Hydromys, Xeromys and Uromys divisions, have a serine in position 311.

At residue position 324, *Conilurus penicillatus* and *Leporillus conditor* have an isoleucine (I) in common with two *Uromys* species, whereas the two *Mesembriomys* species have a threonine (T) at this position (Fig. 3.1). The majority of species also retained the five serine (S) residues (Ser-329, Ser-331 to Ser-334) found in the mouse and rat amino acid sequence. Of the 22 species of *Pseudomys* investigated two sibling taxa, *P. albocinereus* and *P. apodemoides*, have a leucine (L) in position 331 and five species collectively known as pebble-mound *Pseudomys* (*P. calabyi*, *P. chapmani*, *P. johnsoni*, *P. laborifex* and *P. patrius*) have a proline (P) in position 334. In addition, *Conilurus penicillatus* has a threonine (T) in this position. Hence, the five potential *O*-linked glycosylation sites (serine/threonine residues) conserved in the Mus, Rattus, Lorentzimys, Pogonomys, Hydromys, Xeromys and Uromys divisions have been conserved in all but six of the Pseudomys division species (Fig. 3.1).

The five *Notomys* species all share an identical amino acid sequence within the exon 7 coding region, in common with the single *Mastacomys* species and 64% of species from the *Pseudomys* genus. Within the ten amino acid stretch from residue 335 to 344, all 39 species share the same identical sequence.

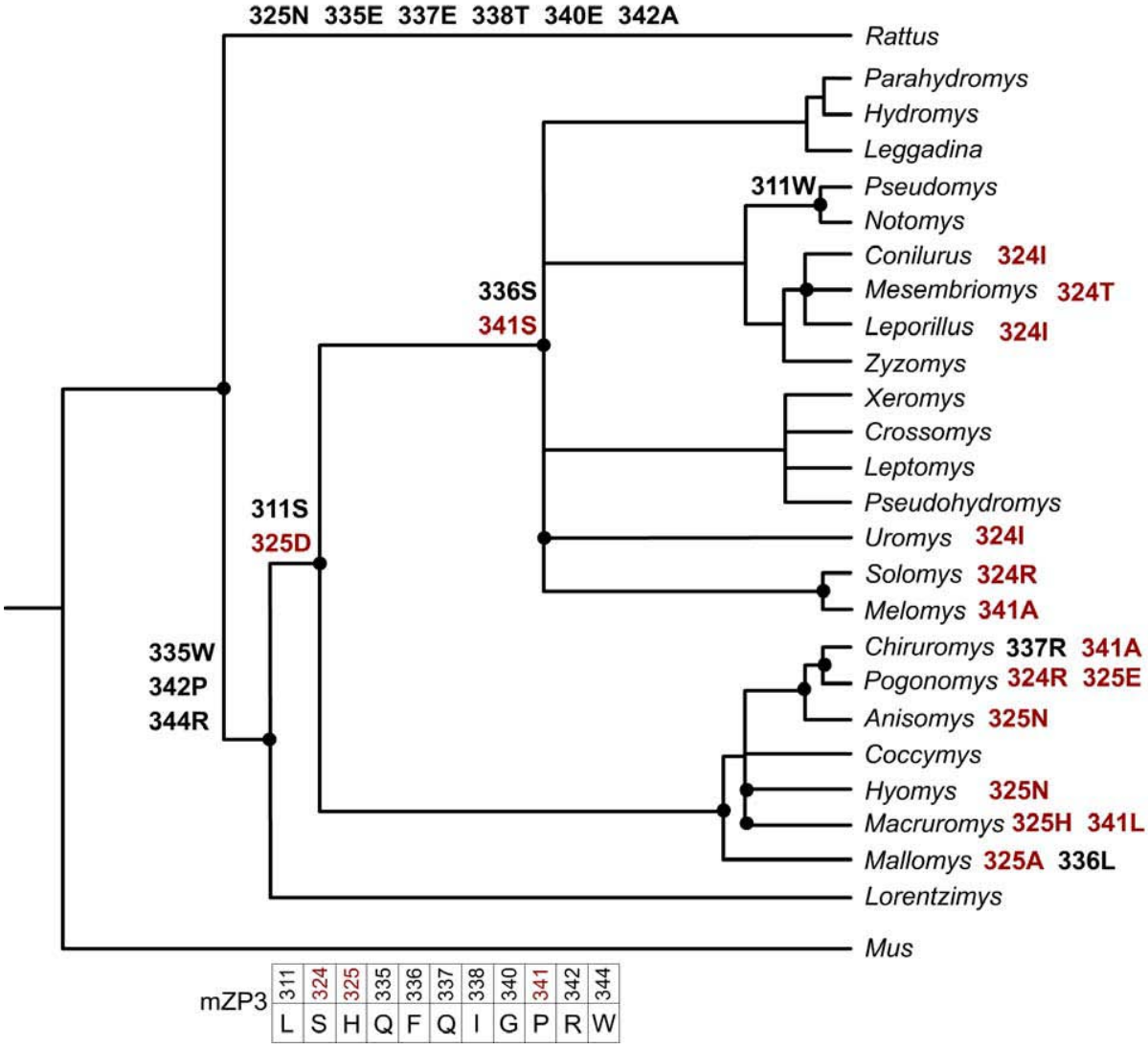


Fig. 3.2 Ancestral reconstruction of the region encoded by exon 7 of *Zp3* from the New Guinean and Australasian Old Endemic murines. Proposed phylogeny of the New Guinean and Australasian murines, based on microcomplement fixation of albumin data (Watts & Baverstock, 1994b, 1995). Amino acid substitutions (single letter code) for selected residue positions for the exon 7 coding region have been plotted against lineages, with ZP3 from *Mus musculus* used as reference. Black dots indicate those nodes where amino acid substitutions have taken place. Note the many unresolved branches, and not all genera are represented. Amino acid residues that show evidence of parallel evolution on different lineages are highlighted in red.

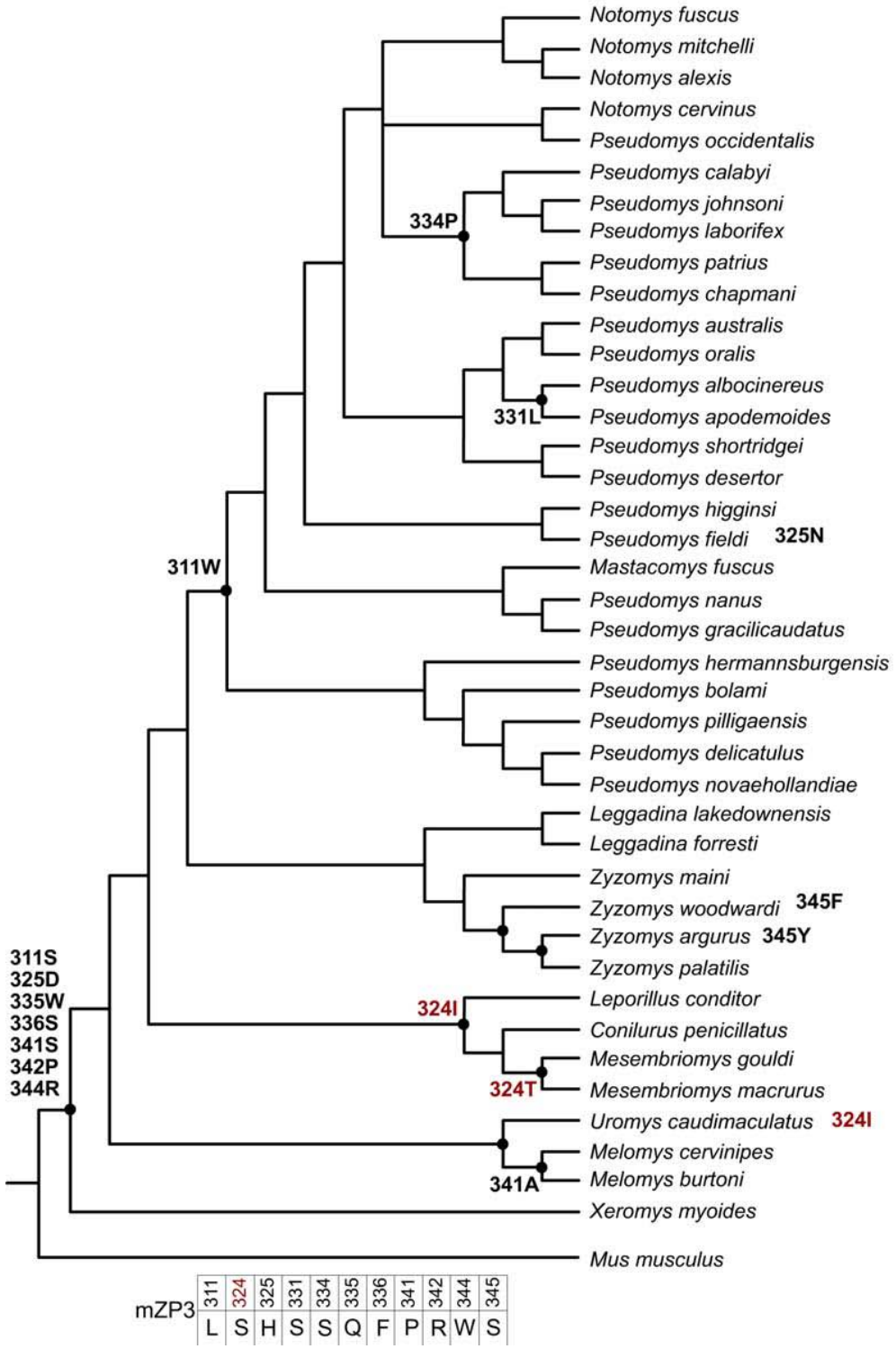


Fig. 3.3. Ancestral reconstruction of the region encoded by exon 7 of *Zp3* from the Australian murines. Proposed phylogeny of the Australian murine rodents, based on nucleotide sequence data from the mitochondrial control region (Ford 2006). Amino acid substitutions (single letter code) for selected residue positions for the exon 7 coding region have been plotted against lineages, with ZP3 from *Mus musculus* used as reference only. Black dots indicate those nodes where amino acid substitutions have taken place. Amino acid residues that show evidence of parallel evolution on different lineages are highlighted in red.

3.3.5 Possible effects of amino acid change

3.3.5.1 Isoelectric points and hydropathic profiles

To assess the possible effects amino acid sequence change may have on a protein, three methods have been used in the present study: relative serine/threonine composition (potential *O*-linked glycosylation sites); isoelectric point (change in overall charge) and hydropathy profiles (changes in hydrophobicity and hydrophilicity). All methods were computed using the software program WinPep (version 3.01, Hennig 1999), then imported into Microsoft Excel. For the purposes of these analyses, species were selected to represent the common amino acid sequences, and therefore data from all species are not presented. As the carboxy terminal of exon 7 is the region not only hypothesized as being involved in sperm-ZP binding, but is also the region of most variability, only the eighteen residues between 328 and 345 have been analysed.

Table 3.10. Isoelectric point, relative serine/threonine percentages and the average hydropathy index for selected species of New Guinean old endemic murines in respect of the region from residues 328 to 345. A negative hydropathy index suggests that the region is hydrophilic and may be present on the surface of the protein.

	Isoelectric point	Relative serine/threonine composition (%)		Average hydropathy index
		Serine	Threonine	
Mus musculus	8.4	33	0	-1.09
Anisomys imitator	8.4	33	0	-0.98
Chiruromys vates	10.5	33	0	-0.85
Coccymys ruemmleri	8.4	33	0	-0.98
Hyomys goliath	8.4	33	0	-0.98
Lorentzimys nouhuysi	8.4	33	0	-0.98
Macruromys major	8.4	33	0	-0.68
Mallomys rothschildi	8.4	33	0	-0.98
Mammelomys lanosus	8.4	33	0	-0.93
Pogonomys macrourus	8.4	33	0	-0.98
Pogonomelomys mayeri	8.4	33	0	-0.68

The isoelectric point ranges from 8.4 in the majority of species to 10.5 in *Chiruromys vates*. The charge of a region of a peptide sequence is thought to influence glycosylation patterns as well as folding of the molecule (Nehrke *et al.* 1996). The serine residue percentage of this region of ZP3 is 33% for all New Guinean species. Therefore, the charge and serine composition remain unchanged in all but one

species. However, although the average hydropathy index of the species indicates that the region is hydrophilic, there is marked variation in the mean hydropathy index, ranging from an average of -0.68 in two species (*Pogonomelomys mayeri* and *Macruromys major*) to -0.98 in most others. A more detailed analyses of the pattern of hydropathy is conducted in section 3.3.5.2.

Table 3.11. Isoelectric point, relative serine/threonine percentages and the average hydropathy index for selected species of Australasian old endemic murines. A negative hydropathy index suggests that the region is hydrophilic and may be present on the surface of the protein.

	Isoelectric point	Relative serine/threonine composition (%)		Average hydropathy
		Serine	Threonine	
Mus musculus	8.4	33	0	-1.09
Hydromys chyrsogaster	8.4	44	0	-1.14
Leggadina forresti	8.4	44	0	-1.14
Leporillus conditor	8.4	44	0	-1.14
Mastacomys fuscus	8.4	44	0	-1.14
Melomys burtoni	8.4	39	0	-0.99
Melomys cervinipes	10.5	39	0	-1.05
Melomys rufescens	8.4	44	0	-1.14
Mesembriomys gouldi	8.4	44	0	-1.14
Notomys alexis	8.4	44	0	-1.14
Notomys cervinus	8.4	44	0	-1.14
Paramelomys platyops	8.4	44	0	-1.14
Paramelomys rubex	8.4	50	0	-1.09
Pseudomys albocinereus	8.4	39	0	-0.88
Pseudomys australis	8.4	44	0	-1.14
Pseudomys johnsoni	8.4	39	0	-1.18
Solomys salebrosus	8.4	44	0	-1.14
Uromys anak	8.4	44	0	-1.14
Zyzomys argurus	8.3	39	0	-1.17
Zyzomys maini	8.4	44	0	-1.14
Zyzomys palatilis	8.4	44	0	-1.14
Zyzomys woodwardi	8.4	39	2	-0.94

The isoelectric points for the Australasian Old Endemic murines are almost identical to the New Guinean Old Endemic murines, ranging from as low as 8.4 in most species to 10.5 in *Hydromys chrysogaster*. The serine composition of this region of ZP3 varied considerably between species, ranging from 39% in several species to 50% in *Paramelomys rubex*. The mean hydropathy index was also more hydrophilic than the New Guinean species with most species having a mean index of less than -1.

3.3.5.2 Hydropathy profiles

Each amino acid has its own hydropathy index based on how much the amino acid side chains seek to avoid contact with any particular aqueous substance. Hydrophobic side chains tend to be buried inside the protein, while hydrophilic side chains tend to be on the surface, in contact with its aqueous surroundings. ZP3 is a globular protein and therefore it is reasonable to assume that its hydrophobic side chains may be buried inside the protein and those amino acid regions that are hydrophilic may lie on the outside. Thus, determining the hydropathy profile of the exon 6 and 7 coding region of ZP3 may provide a method of visualizing the impact amino acid changes may have on a protein.

Fig.3.4 shows the hydropathy profile of the exon 6 and 7 coding region of ZP3 from representative New Guinean Old Endemic murine species (see Chapter 2.9.2 for method). At residue position 311, the majority of New Guinean old endemic species have a serine which makes the surrounding region hydrophilic and therefore possibly on the surface of the glycoprotein. *Mammelomys lanosus*, at position 325, has an alanine while the remainder of the New Guinean rodents have all hydrophilic residues: asparagines (N), aspartic acid (D), histidine (H) or glutamic acid (E). Within the region 335 to 340 the region appears to be hydrophobic due to the presence of either a leucine (L) or phenylalanine (F) in position 345. The residue in position 341 is the highly hydrophobic residue leucine (+3.8) in *Pogonomelomys mayeri* and *Macruromys major*, the mildly hydrophobic residue alanine (+1.8) in *Chiruromys vates* whereas the other New Guinean Old Endemic murines have a proline (-1.6) in this position.

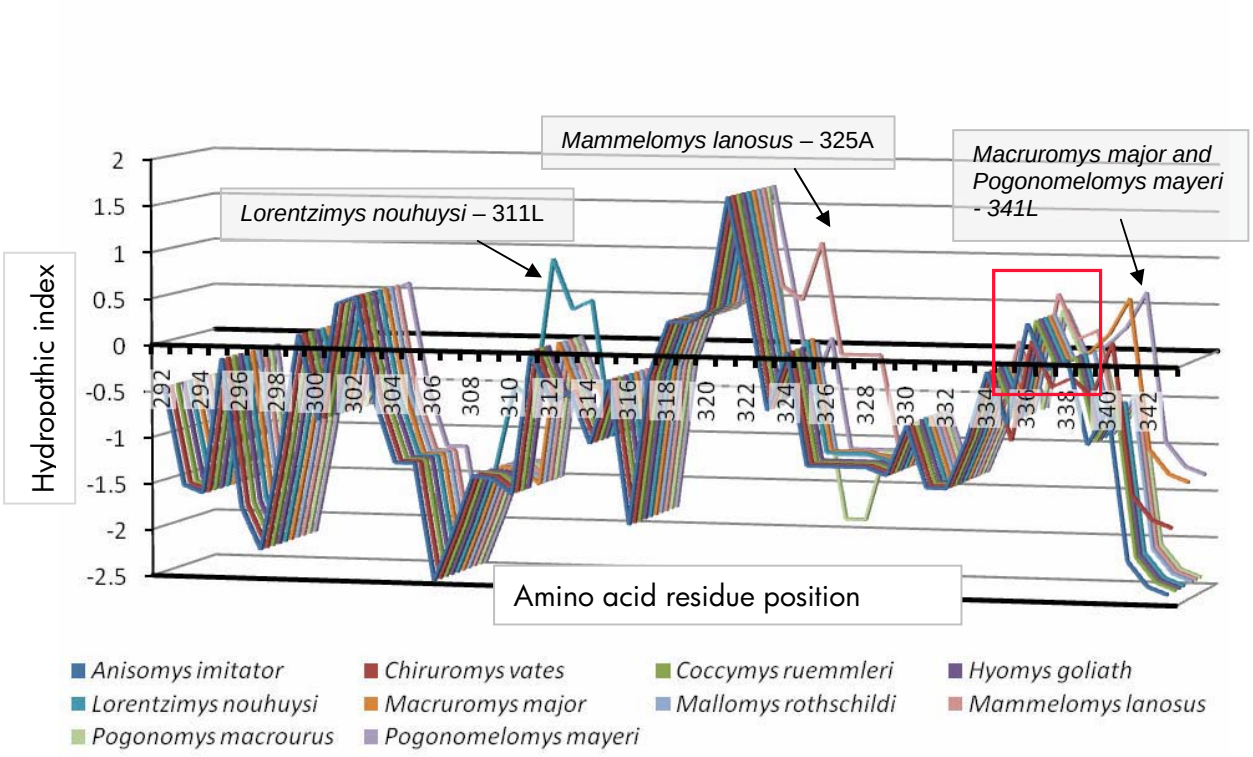


Fig. 3.4. Three dimensional graphical representation of the hydropathic profile of the exon 6 and 7 coding region of *Zp3* from New Guinean Old Endemic murines, showing the similarities and differences of hydropathy between species . The hydropathic index was determined using the average of a five residue sliding window. Below the line is that part of the protein that is hydrophilic and therefore likely to be on the surface of the protein. Conversely, above the line is that part of the protein that is hydrophobic and likely to be buried inside the protein. The text above the graph shows the species and the amino acid change that increased the hydrophobicity of the region proposed to be involved in sperm-zona pellucida binding. The boxed area, in red, shows the increased hydropathy of the region 335 to 345 due to the amino acid in position 336 being either a leucine (+3.8) or a phenylalanine (+2.8).

Fig. 3.5 shows the hydropathy profile of the Australasian Old Endemic murine species. This profile shows that the hydropathy of the sequence is conserved until residue 324. *Leporillus conditor* and *Uromys anak* have the highly hydrophobic isoleucine residue (+4.5) in this position, while the majority have a serine (-0.8), threonine (-0.7) or arginine (-4.5) which are all hydrophilic. At position 331, the majority of species have a serine in this position with the exception of *Pseudomys albocinereus* (as well as *P. apodemoides*: data not shown) which has a leucine (+3.8) in this position.

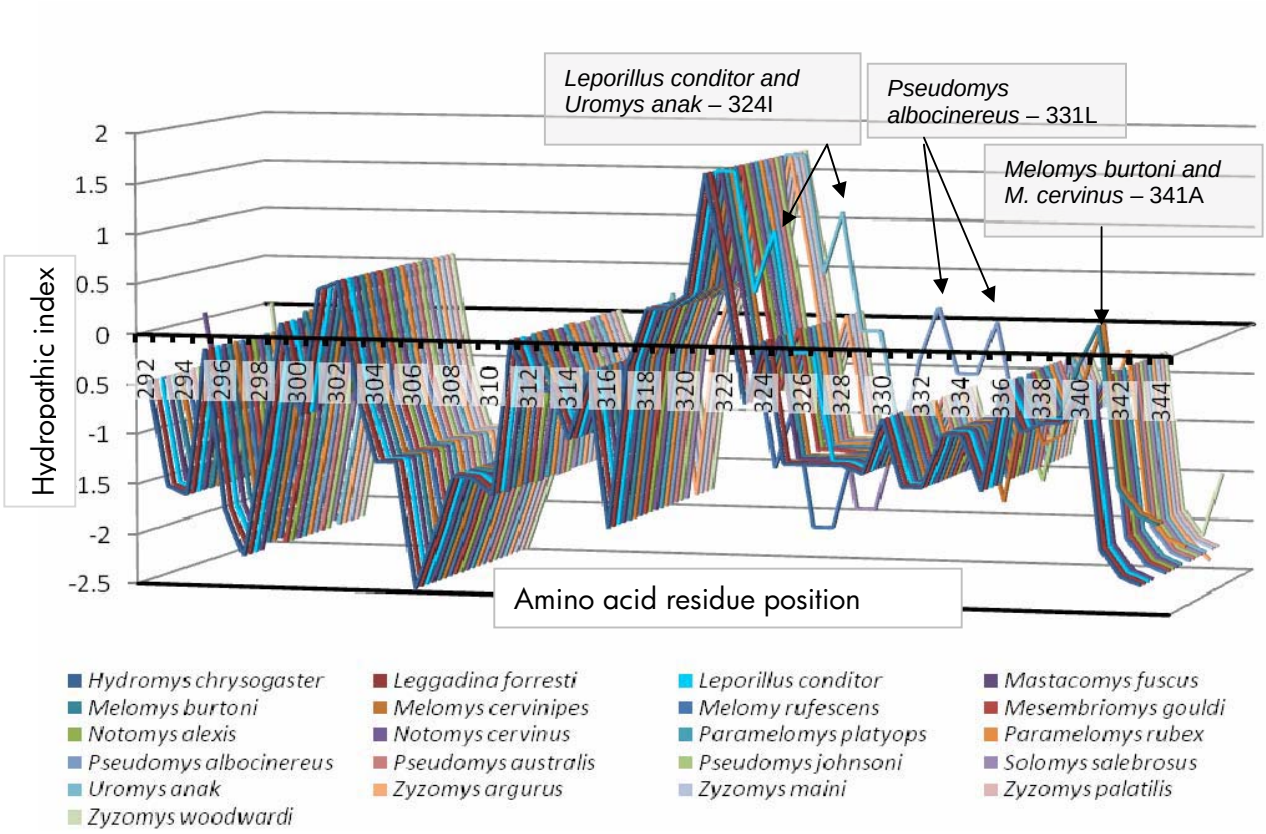


Fig. 3.5. Three dimensional graphical representation of the hydropathic profile of the exon 6 and 7 coding region of *Zp3* from Australasian old endemic murines, showing the similarities and differences of hydropathy between species. The hydropathic index was determined using the average of a five residue sliding window. Below the line is that part of the protein that is hydrophilic and therefore likely to be on the surface of the protein. Conversely, above the line is that part of the protein that is hydrophobic and likely to be buried inside the protein. The box above the graph shows the species and the amino acid change that increased the hydrophobicity of the region proposed to be involved in sperm-zona pellucida binding.

Between residues 328 to 345, the only other change to the hydrophilicity of this region was two *Melomys* species (*M. burtoni and M. cervinipes*) with an alanine (+1.8) in position 341. The hydropathy profile of the Australasian Old Endemic rodents differs from that of the New Guinean rodents in this region as the sequence of the majority of Australasian species is hydrophilic and therefore likely to be on the surface of the glycoprotein, while the sequence in the New Guinean species is hydrophobic and likely to be, at least, partially buried within the protein.

3.4 Discussion

The hypothesis tested in this Chapter was that there is a high level of sequence divergence of the exon 7 coding region of *Zp3* between closely related species which has the potential to contribute to species specificity of sperm-ZP binding. In order to test this hypothesis, the amino acid sequence of the proposed region involved in primary sperm-ZP binding (in the mouse) in over 50% of species (68 species) and 80% of genera of Old Endemic murine rodents from New Guinea and Australasia has been determined. The results show that within each division, the sequence divergence of the exon 7 coding region is low, or even nonexistent, although it was found to be higher than in exon 6. For instance, in all six species within two divisions, Hydromys and Xeromys, an identical amino acid sequence of the exon 7 coding region was found to be present. Furthermore, it was also identical between all five species of *Notomys* and 63% (14/22) of *Pseudomys* species. These results do not support the hypothesis that there is a higher level of sequence divergence of the exon 7 coding region of *Zp3* between closely related species.

Within all divisions (as defined by Musser & Carleton 2005), the rates of nucleotide sequence divergence of the exon 7 coding region were low (<0.026), although it was greater between divisions that included the more distantly related species. For example, between the Hydromys and Xeromys divisions, the rate of divergence within exon 7 was less than 0.01, but this increased to 0.0383 between the Hydromys and Lorentzimys divisions. The amino acid divergence rates (expressed as a percentage) showed there was no divergence between Hydromys and Xeromys, but it is 11.11% between these two divisions and Lorentzimys. An identical amino acid sequence shared between the Hydromys and Xeromys divisions is also present in all five species of *Notomys* and 63% (14/22) of *Pseudomys*.

The degree of amino acid sequence divergence between the New Guinean Old Endemic divisions (Lorentzimys versus Pogonomys: 7.07%) was found to be higher than that which occurred within the large Pseudomys division (3.3%). However, between sibling taxa, such as two species of the

Mammelomys genus, and two species of the *Paramelomys* genus, it was identical. Within the large *Pseudomys* genus, divergence occurs but is limited to monophyletic clades of species. For example, *P. albocinereus* and *P. apodemoides*, long considered to be sister taxa (Watts & Baverstock 1992), share a leucine in position 331, and the five species of *Pseudomys* known as pebble-mound Pseudomys (*P. calabyi*, *P. chapmani*, *P. johnsoni*, *P. laborifex* and *P. patrius*) share a proline in position 334. Within a ten amino acid stretch from 335 to 344, identified in the rat as having a high percentage of substitutions relative to the mouse (Scobie *et al.* 1999) and contained within the putative combining-site for sperm, 100% of species from the Hydromys, Xeromys and Pseudomys divisions all have an identical sequence.

Generally, within New Guinean and Australasian Old Endemic rodents, divergence in the exon 7 coding region occurs only between distantly related species, such as between those in the Pogonomys and Pseudomys divisions. Nevertheless, the amino acid differences, many of which change the charge of the residue and/or involve a loss/gain of glycosylation sites, indicate that these changes took place in a common ancestor prior to divergence into sibling, or closely related species. In the data set, three detected amino acid changes relative to the mouse (Trp-335, Pro-342 and Arg-344) are present in all species studied and thus were probably fixed prior to the radiation of the New Guinean (Lorentzimys and Pogonomys divisions) and Australasian (Hydromys, Xeromys, Uromys and Pseudomys divisions) murines. Likewise, two serine residues, in positions 336 and 341, are shared by 95% of species from the Australasian divisions, but do not occur in the New Guinean species and hence probably became fixed in the common ancestor of the Australasian Old Endemic murines after its divergence from the New Guinean clade. Other changes are lineage specific and affect only a few species, for example, Ala-341 which occurs in all three Australasian species of *Melomys* but not in those from New Guinea. These results strongly suggest that divergence in this region has not played any role in reproductive isolation of these Australasian murine lineages.

Amino acid substitutions can have a number of effects on the secondary and tertiary structure of a peptide. Ionic interactions, resulting from a change in charge of a protein as well as varying glycosylation patterns, can change not only the three dimensional appearance of a heavily glycosylated protein but also influence how other molecules interact with the glycoprotein. Differing hydropathy profiles of a sequence can also affect how a peptide folds and interacts with other proteins. The region from 328 to 345, identified by Wassarman and colleagues as the combining-site for sperm in the mouse ZP3, shows variation in hydropathy and charge between genera, but very little between closely related species from within each genus. There are differences in hydropathy within this region between the Old Endemic New Guinean and Australasian divisions and yet the overall charge of the region is unchanged. The number of serine residues also varies which suggests that the changes to the primary structure of the region 328 to 345 may have only a small effect on glycosylation as a result of the loss/gain of glycosylation site and change in charge. However, the higher hydrophobicity of the region within the New Guinean murines suggests that this region of ZP3 may fold differently from that of the Australasian murines and which may alter the three dimensional structure of the zona matrix. It is therefore feasible that these changes may alter the way sperm recognise the ZP matrix and render it refractory to inter-species fertilization, in particular between members of different clades.

The low amino acid divergence of the region encoded by exon 7 of *Zp*3 within a large diverse group of murine species, coupled with the retention of *N*-linked glycosidic and disulfide bonding sites, suggest that there may be functional constraints operating to ensure stability of this region of the glycoprotein and hence the occurrence of purifying selection is likely to have occurred. It may alternatively be that speciation has occurred over a relatively short period of time which is insufficient to allow the generation and fixation of new mutations; a conclusion that is not supported by other data (Godthelp 1999; Watts & Baverstock 1995; Watts *et al.* 1992). Evidence of positive Darwinian selection, where mutations that result in a change of amino acid residue are more common than silent mutations, has been found to

occur between distantly related species such as the mouse, rat, human and cat (Swanson *et al.* 2001) as well as between several species of *Mus* (Jansa *et al.* 2003). Within the Australasian murines, the low amino acid sequence divergence occurring within the exon 7 coding region of ZP3 suggests an absence of positive selection within this part of the ZP3 glycoprotein, although it may have occurred within the New Guinean lineages at two sites, 325 and 341. For investigation into whether this region of the gene is evolving under positive selection see Chapter 5.

Conclusion

In conclusion, these data show that closely related species of New Guinean and Australasian Old Endemic murine rodents, in general, share an identical amino acid sequence in the region purported to be involved in sperm-ZP binding in the mouse. Nevertheless, this region of ZP3 appears to have evolved faster than that of the exon 6 coding region, although most of the amino acid changes appear to have become fixed prior to the radiation of New Guinean and Australasian murine divisions. While changes to the primary structure of ZP3 could alter binding sites for the sperm, the low rate of sequence divergence within each of the six murine divisions suggests the divergence in this region did not result in potential species specificity of sperm-ZP binding. The results detailed in this chapter do not support the hypothesis that sequence divergence of the region encoded by exon 7 of *Zp3* contributes to potential species specificity of sperm-ZP binding.

Chapter 4

Evolution of exon 6 and 7 of *Zp3* within African, Eurasian and South-east Asian murine rodents

Image on reverse: New Guinean Old Endemic rodent, *Mallomys rothschildi*.
Modified image from Flannery, 1990

Chapter 4

Evolution of exon 6 and 7 of Zp3 within African, Eurasian and South-east Asian murine rodents.

4.1 Introduction

In Chapter 3, the amino acid sequence of the region encoded by exon 7 of *Zp*3 from a wide range of New Guinean and Australasian Old Endemic murine rodents was determined. While the sequence varied from that of the laboratory mouse and rat, there was a high level of sequence conservation between species within divisions and a high percentage of species shared an identical amino acid sequence. It was concluded that this region did not contribute towards species specificity of sperm-zona pellucida binding, if it occurs, at least within the Australasian Old Endemic murine species.

South-east Asia contains a large number of murine rodents, including *Rattus* and *Rattus*-like species, that are not closely related to species within the New Guinean and Australasian Old Endemic clades. The sperm head morphology of the South-east Asian murines is also generally distinct from those of New Guinean and Australasian species (Breed 2004). During the species radiation of South-east Asian murines into Indochina, Central Asia, Philippines and Sulawesi, migration of early rodents was also occurring into Africa, where a large number of species now occur (Steppan *et al.* 2005). As stated in Chapter 1.2.8.1, the present genera of Africa have traditionally been placed in one clade (Watts & Baverstock 1995) but recently it has been suggested that these species belong to two clades, separated by long intervening branches (Steppan *et al.* 2005; Lecompte *et al.* 2005).

It is possible that the low level of sequence divergence of the exon 7 coding region of *Zp3* seen with the New Guinean and Australasian Old Endemics is not reflected within either the South-east Asian (including Australasian endemic *Rattus* species) or African clades. The aim of this Chapter is to

determine whether two other major clades of murines show similar evolution of the exon 7 coding region of *Zp3* to that of New Guinean and Australasian Old Endemics. In order to investigate the level of sequence divergence of this region within these species, DNA extraction, PCR amplification and sequencing was performed on a total of 28 species from 8 divisions, and the level of nucleotide and amino acid divergence was determined. As for the New Guinean and Australasian Old Endemic murines, current hypothetic phylogenetic relationships were used to detect the evolution of the exon 7 coding region of the *Zp3* gene and data was analysed in a similar way to that of the Old Endemic murines detailed in Chapter 3.

4.2 Material and Methods

For material and methods, see Chapter 2.

The species investigated in this chapter originate from either Africa, Eurasia or South-east Asia. They are divided into four divisions from Africa (Dasymys, Aethomys, Arvicanthis and Stenocephalemys), one division only from Eurasia (Apodemus from China) and three divisions from South-east Asia (Dacnomys, Maxomys and Rattus). Included in the Rattus division are *Rattus* species endemic to both New Guinea and Australia.

For primer design, see Chapter 2.1.5. For the method used to estimate evolutionary distances see Chapter 2.7.

For comparative purposes only, *Mus musculus* has been used as a reference sequence for both the nucleotide and amino acid sequences. This has been done for two reasons. Firstly, the species *Mus musculus* has been the model for comparative studies of murine ZP3. Secondly, *Mus musculus* was used as an outgroup for species investigated in Chapter 3, and therefore nucleotide and amino acids differences can be compared between species investigated in this chapter to those in the previous chapter.

Within the Rattus division, the *Rattus* genus, of which there are 66 extant species, has been divided into several species groups (Musser & Carleton 2005) including *Rattus norvegicus* group (RattusN), *Rattus exulans* (RattusE), *Rattus leucopus* (RattusL: containing mostly New Guinean species), and *Rattus fuscipes* (RattusF: containing Australian species). For comparative purposes the non-Rattus species (*Bunomys*, *Bandicota* and *Paruromys*) have been grouped together as non-Rattus species.

Ancestral reconstruction plotting amino acid changes (Yang *et al.* 1995) was computed using the codeml program of the PAML (Phylogenetic Analysis using Maximum Likelihood) software version 3.15 (Yang 1997) and displayed in Figs. 4.3 and 4.4.

4.3 Results

4.3.1 PCR Sequencing

Sequencing was performed only when PCR products, after gel electrophoresis, ethidium bromide staining and visualization under UV light, showed a clear band of approximately 300 base pairs. No double signals were observed and sequencing produced clean results in the majority of species.

Some problems in PCR amplification occurred. Three 10µl solutions of DNA for four species of *Aethomys* (*A*. *chrysophilus*, *A*. *ineptus and Micaelamys* (ex-*Aethomys*) *namaquensis* and *M*. *granti*) were obtained from the University of Pretoria, South Africa in April 2004. Only the third aliquot of *A. chrysophilus* produced PCR products, using primers G510/G693, that could be visualized under UV light (Z321). The first aliquot of *A. ineptus* and *M*. *namaquensis* produced clear bands after gel electrophoresis and clean sequence (Z198 and Z199). PCR was also performed on the second aliquots of *A*. *ineptus* and *M*. *namaquensis*. However, only *A*. *ineptus* produced a clear band after gel electrophoresis and PCR sequencing was performed (Z293). Unfortunately, this produced a sequence that was identical to *M*. *namaquensis* (Z199). To rule out mislabeling prior to PCR, a repeat PCR was performed and after sequencing, the same result was obtained (Z322). A specimen from the Australian Biological Tissue Collection for *A*. *ineptus* was obtained (ABTC 65829) and DNA was extracted. After PCR, and gel electrophoresis, a clean band indicated amplified product and PCR sequencing was performed (Z326). The sequence was identical to the first sequence obtained for *A*. *ineptus* (Z198). It was therefore concluded that the second aliquot of DNA from South Africa (An^2) had been mislabeled and was disregarded. No attempt was made to produce the DNA sequence for *A. granti*.

DNA was extracted from an animal identified in the ABTC listing as *Praomys natalensis* (ABTC 65842, caught 40 km from Pretoria, South Africa), although this species is now recognized as *Mastomys natalensis* (Musser & Carleton 2005). Jansa *et al.* (2003) published the nucleotide sequence for selected exons/introns of *Zp3* for *Mastomys hildebrandtii* (GenBank accession number AY057790). The

taxonomy of these two species has been revised since 2003 (Musser & Carleton 2005) and it now appears that *M. hildebrandti* is a synonym of *M. natalensis*. The nucleotide sequence obtained from *M. natalensis* in the present study is similar to the published sequence of *Mastomys hildebrandti* (Jansa *et al.* 2003), differing at only one nucleotide site within exon 7 (Fig 4.1), and therefore, it is not possible to determine if they are two different species. As the ABTC specimen of *M. natalensis* was collected from South Africa, its type locality (Musser & Carleton 2005), only the sequence data from this species is included in the present study.

```
                          *   10    *   20    *   30    *   40    *   50    *   60    *   70
Mastomys natalensis   : CCAGCTAACCAGATCCCTGACAAACTTAACAAAGCCTGTTCATTCAACAAGACTTCGCAGAGGTGAGGAG
Mastomys hildebrandti : CCAGCTAACCAGATCCCTGACAAACTTAACAAAGCCTGTTCATTCAACAAGACTTCGCAGAGGTGAGGAG

                          *   80    *   90    *  100    *  110    *  120    *  130    *  140
Mastomys natalensis   : ACCAGGCTTTGTGTGTGTG----GGCACCCGGAGGCTATTCACATCGATTTCTCTTCGATTACACAATGG
Mastomys hildebrandti : ACCAGGCTTTGTGTGTGTGTGTGGGCACCCGGAGGCTATTCACATCGATTTCTCTTCGATTACACAATGG

                          *  150    *  160    *  170    *  180    *  190    *  200    *  210
Mastomys natalensis   : CAAACTTCTGTCCTTTCTGAGCTAAGTAAGCTTTTTTGTCTTGTTACTCAGTTGGTTACCAGTAGAGGGC
Mastomys hildebrandti : CAAACTTCTGTCCTTTCTGAGCTAAGTAAGCTTTTTTGTCTTGTTACTCAGTTGGTTACCAGTAGAGGGC

                          *  220    *  230    *  240    *  250    *  260    *  270    *  280
Mastomys natalensis   : GATGCTGACATCTGTGATTGCTGCAGCCACGCCAACTGTAGTAATTCAAGCTCTTCACAGTTCCTGATCC
Mastomys hildebrandti : GATGCTGACATCTGTGATTGCTGCAGCCACGCCAACTGTAGTAATTCAAGCTCTTCACAGTTCCTGATCC

                          *  290    *  300    *
Mastomys natalensis   : ACGGACCCTACCAGTGGTCC
Mastomys hildebrandti : ACGGACCTTACCAGTGGTCC
```

Fig. 4.1. Nucleotide sequence alignment of exon 6 to exon 7 of *Zp3* highlighting the nucleotide differences between *Mastomys natalensis* and *M. hildebrandti*.

DNA was extracted from *Rattus colletti* (Rattus division) and PCR sequencing was performed (Z327). Only the reverse sequence produced a clean signal, and was missing the last 8 nucleotides. It was clear that the sequence was very similar to other Australian endemic *Rattus* species, such as *Rattus fuscipes*, but due to time constraints, a repeat PCR was not performed and the sequence for *R. colletti* was not used in any analysis.

All sequences have been lodged with GenBank at the NCBI. Table 4.1 lists divisions, genus, species and common names, ABTC and GenBank accession numbers.

Table 4.1. List of African, Eurasian and South-east Asian rodent species investigated in this study, showing division, genus, species and common names, ABTC and GenBank accession numbers (NCBI).

Division	Genus	Species	Common Name	ABTC number	GenBank Accession No.
African divisions					
Dasymys					
	Dasymys	*incomtus*	Common Dasymys	65735	EU004043
Aethomys					
	Aethomys	*chrysophilus*	Red Veld Aethomys	n/a	EU004037
	Aethomys	*ineptus*	Tete Veld Aethomys	65829	EU004038
	Micaelamys	*namaquensis*	Namaqua Micaelamys	n/a	EU004039
Arvicanthis					
	Lemniscomys	*griselda*	Griselda's Lemniscomys	65835	EU004044
	Rhabdomys	*pumilio*	Xeric Four-striped Grass Rat	65831	EU004064
Stenocephalemys					
	Mastomys	*natalensis*	Natal Mastomys	65842	EU004051
Eurasian division					
Apodemus					
	Apodemus	*chevrieri*	Chevrier's Field Mouse	13966	EU004040
South-east Asian divisions					
Dacnomys					
	Leopoldamys	*edwardsi*	Edward's Leopoldomys	67574	EU004045
	Leopoldamys	*sabanus*	Indomalayan Leopoldamys	67572	EU004046
	Niviventer	*fulvescens*	Indomalayan Niviventer	48010	EU004049
Maxomys					
	Maxomys	*bartelsii*	Bartel's Javan Maxomys	48059	EU004047
	Maxomys	*hellwaldii*	Hellwald's Sulawesi Maxomys	65760	EU004048
Rattus					
	Bunomys	*andrewsi*	Andrew's Bunomys	65755	EU004042
	Bandicota	*indica*	Greater Bandicoot Rat	69090	EU004041
	Paruromys	*dominator*	Giant Sulawesi Rat	65763	EU004050
Rattus exulans species group					
	Rattus	*exulans*	Pacific Rat	42509	EU004052
Rattus leucopus species group					
	Rattus	*leucopus*	Cape York Rat	n/a	EU004054
	Rattus	*mordax*	Eastern New Guinea Rat	48962	EU004056
	Rattus	*niobe*	Eastern New Guinea Mountain Rat	42489	EU004057
	Rattus	*praetor*	Large New Guinea Spiny Rat	47272	EU004058
	Rattus	*steini*	Stein's New Guinea Rat	49258	EU004060
	Rattus	*verecundus*	New Guinea Slender Rat	49292	EU004062
Rattus fuscipes species group					
	Rattus	*fuscipes*	Australasian Bush Rat	18144	EU004053
	Rattus	*lutreolus*	Australasian Swamp Rat	51763	EU004055
	Rattus	*sordidus*	Canefield Rat	41160	EU004059
	Rattus	*tunneyi*	Australasian Pale Field Rat	41159	EU004061
	Rattus	*villosissimus*	Australasian Long-haired Rat	41151	EU004063

4.3.2 Nucleotide Sequences

The complete nucleotide sequence data for each taxon is provided in Appendix 2.

Sequences varied in length from 297 bases in two species of *Leopoldamys* (Dacnomys division) to 308 bases in two *Aethomys* species (*A. chrysophilus* and *A. ineptus*). This variation in sequence length is partly attributable to insertions and deletions within intron 6 and partly due to some species having indels within exon 7. Relative to the mouse sequence, the exon 7 sequence for *Lemniscomys griselda* (Arvicanthis division) contained a six nucleotide deletion, *Maxomys hellwaldii* contained a three nucleotide insertion, and the sequences from five species of Australian *Rattus* contained a three nucleotide insertion.

4.3.2.1 Polymorphisms

Polymorphic sites were recognized when the chromatogram (see Chapter 2.4) for both the forward and reverse sequences indicated a double signal at the same nucleotide site. This suggested that the particular species was heterozygous at that specific site but sample sizes were insufficient to determine the frequency of the nucleotide site allelles. As carried out previously in chapter 3, the nucleotide most common within the division was selected for comparative purposes and where this was not possible, the nucleotide polymorphic site was designated with a N. However, for purely comparative purposes, the polymorphism was retained (Appendix 2).

Within exon 6, half of the polymorphisms were silent (Table 4.2). Both involved transitions at the third base of the codon encoding proline (P). At both sites (3 and 18), the nucleotide at these positions varied between species, although as there are four codons that specifically encode for proline, they are all silent substitutions. Of the two amino acid replacing polymorphisms, *Apodemus chevrieri* had a A/C transversion in position 20, indicating that this species is heterozygous at this loci. However, all other species, including the New Guinean and Australasian Old Endemic species investigated in Chapter 3,

have an adenine in this position. *Aethomys chrysophilus* had a transitional polymorphism (A/G) in position 4 whereas all other species, again including the New Guinean and Australasian Old Endemics, have a G in this position.

Table 4.2. Species where a double signal appeared on the chromatogram, suggesting the presence of a polymorphism within the African, Eurasian and South-east Asian species. Nucleotide position number is relative to the mouse sequence in Appendix 2.

	Species	Nucleotide position No.	Polymorphism	Affect
Exon 6				
	Apodemus chevrieri	20	A/C	GCC(A), GAC(D) amino acid replacing
	Aethomys chrysophilus	4	A/G	ACT(T), GCT(A) amino acid replacing
	Lemniscomys griselda	18	C/T	CCC/CCT (P) silent
	Rattus villosissimus	3	G/A	CCA/CCG (P) silent
Intron 6				
	Aethomys chrysophilus	109	C/G	None
	A. chrysophilus	128	A/G	None
	A. chrysophilus	148	A/G	None
	Lemniscomys griselda	149	C/T	None
	Leopoldamys edwardsi	124	T/C	None
	Leopoldamys sabanus	181	C/T	None
	Maxomys hellwaldii	68	A/G	None
	M. hellwaldii	126	C/T	None
	Rattus niobe	110	C/T	None
	Rhabdomys pumilio	152	G/T	None
Exon 7				
	Niviventer fulvescens	323	C/G	CGC(R), GGC(G) amino acid replacing
	Rattus tunneyi	225	C/T	TCG(S), TTG(L) amino acid replacing

Single letter amino acid codes: A = alanine, D = aspartic acid, F = phenylalanine, G = glycine, I = isoleucine, L = leucine, N = asparagine, P = proline, Q = glutamine, R =arginine, S = serine, T = threonine, W = tryptophan. Nucleotide numbers correspond to positions of bases on the nucleotide alignment of exon 6, intron 6 and exon 7 (Appendix 2).

Within exon 7, only two species had polymorphisms, both involving amino acid replacing substitutions. *Niviventer fulvescens* had a transversion (C/G) at position 323. This position is quite variable between species (A, C, G or T), although all New Guinean and Australasian Old Endemic species (nucleotide position 300) had a C in this position. *Rattus tunneyi* had a transitional polymorphism (C/T) at position 225. Other species had either a C or T in this position, changing the amino acid from a serine to a leucine.

4.3.3. Estimations of nucleotide sequence divergence

Estimations of nucleotide sequence divergence were calculated using the Kimura 2-parameter method and computed using the MEGA version 3 software (Kumar *et al.* 2004). In addition, the exon 6-7 nucleotide sequence for *Hylomyscus alleni*, available from GenBank (accession number AY057789), has been incorporated into the Stenocephalemys division for comparative and phylogenetic purposes.

The 28 species investigated in this study separate into two broad groups (with one exception): 21 South-east Asian species (including *Rattus* species from New Guinea and Australia as well as South-east Asia) and 8 African species (including *Hylomyscus alleni*). The exception was *Apodemus chevrieri* from China (Eurasia). The genus *Apodemus* is also the sole member of the division Apodemus as well as the single representative species in this study from Central Asia. Mean estimations of evolutionary distances (*d* values) using nucleotide sequence were first calculated within and between these three broad groups. These *d* values should provide some indication of how distantly related the three groups are. Secondly, mean estimations of evolutionary distances within and between divisions were computed. In addition, separate *d* values within and between the African divisions (8 species) and between species from the large Rattus division (15 species) are provided. To be consistent with reference divisions used in chapter 3, the Mus division (containing the single species of *Mus musculus*) and the Rattus division (containing the single species of *Rattus norvegicus*, the laboratory rat), here signified as RattusN to distinguish it from the large Rattus division (of which it is a member), have been used as reference only (see section 4.2).

4.3.3.1. Exon 6 of Zp3

The sequence alignment for exon 6 showed no insertions or deletions and each sequence was therefore 62 nucleotides in length in all species (Appendix 2).

Table 4.3. Estimated evolutionary distances (*d* value) within and between the African, Eurasian and South-east Asian groups for exon 6 of *Zp3*. The African group contains the divisions of Dasymys, Aethomys, Arvicanthis and Stenocephalemys. The South-east Asian group contains the divisions of Dacnomys, Maxomys and Rattus. *Apodemus chevrieri*, occurring in China, has not been included in the South-east Asian group. RattusN signifies *Rattus norvegicus*, the laboratory rat.

	Within	Between				
		Mus	RattusN	Apodemus	African	SE Asian
Mus division	n/a					
RattusN division	n/a	0.0163				
Apodemus division	n/a	0.0675	0.0854			
African group	0.0817	0.0425	0.0550	0.0947		
SE-Asian group	0.0110	0.0189	0.0322	0.0720	0.0575	

Within the African divisions the *d* value was 0.0817, whereas in South-east Asian divisions it was 0.0110. Between the African and South-east Asian groups of divisions, the *d* value was 0.0575. Note that the *d* value between Mus and RattusN division was 0.0163, a rate comparable to that between Mus and the South-east Asian group (of which RattusN is a member). The high *d* value between the Apodemus division and the South-east Asian group (0.0720) justifies its exclusion from the South-east Asian group of species in the present study.

Table 4.4. Estimated mean pairwise comparisons of evolutionary distances (*d* value), within and between divisions for exon 6 of *Zp3*. RattusN signifies *Rattus norvegicus*, the laboratory rat. 'Steno' is an abbreviation for the Stenocephalemys division.

	Within	Between								
Overall: 0.0382		Mus	RattusN	Dasymys	Aethomys	Arvicanthis	Steno	Apodemus	Dacnomys	Maxomys
Mus	n/a									
RattusN	n/a	0.0163								
Dasymys	n/a	0.0333	0.0502							
Aethomys	0.0450	0.0276	0.0445	0.0621						
Arvicanthis	0.0164	0.0247	0.0415	0.0592	0.0532					
Stenocephalemys	0.1059	0.0875	0.0864	0.1271	0.1150	0.1153				
Apodemus	n/a	0.0675	0.0854	0.1039	0.0978	0.0945	0.0855			
Dacnomys	0.0109	0.0219	0.0387	0.0566	0.0503	0.0473	0.0944	0.0735		
Maxomys	0.0330	0.0247	0.0164	0.0595	0.0532	0.0502	0.0780	0.0854	0.0304	
Rattus	0.0066	0.0176	0.0330	0.0521	0.0457	0.0428	0.0884	0.0699	0.0087	0.0254

All within division *d* values were quite low, with the exception of the Stenocephalemys division (0.1059), with estimations ranging from 0.0450 (Aethomys division) to 0.0066 (Rattus division). Between divisions, the *d* value varied considerably, ranging from as low as 0.0087 (Dacnomys versus Rattus) to 0.1271 (Dasymys versus Stenocephalemys). All mean pairwise comparisons involving the Stenocephalemys division were high. Apodemus had the second highest *d* value in pairwise comparisons. Mean pairwise comparisons between Mus and other divisions were generally lower than between Rattus and the other divisions. Surprisingly, the *d* value between Mus and the Rattus division (0.0176) was lower than that between RattusN and the Rattus divisions (0.0330).

Table 4.5. Estimated pairwise comparisons of evolutionary distances (*d* value) between eight species from the African group for exon 6 of *Zp3*.

			Between					
Overall mean: 0.0817	*D.incom*	*M.nama*	*A.chrys*	*A.inept*	*L.gris*	*R.pum*	*H.alleni*	*M.natal*
Dasymys incomtus								
Micaelamys namaquensis	0.0857							
Aethomys chyrsophilus	0.0502	0.0675						
Aethomys ineptus	0.0502	0.0675	0.0000					
Lemniscomys griselda	0.0502	0.0675	0.0331	0.0331				
Rhabdomys pumilio	0.0853	0.0853	0.0500	0.0500	0.0164			
Hylomyscus alleni	0.1076	0.1046	0.0866	0.0866	0.0866	0.1059		
Mastomys natalensis	0.1467	0.1639	0.1242	0.1242	0.1242	0.1446	0.1059	

Within the African group of divisions, the estimated pairwise comparisons were relatively high, compared to the rate of divergence between Mus and Rattus. Estimations ranged from 0.0000 (between the two *Aethomys* species) and 0.0164 (between *Lemniscomys griselda* and *Rhabdomys pumilio*, Arvicanthis division) to 0.1638 between *Micaelamys namaquensis* (Aethomys division) and *Mastomys natalensis* (Stenocephelamys division). The high rate of divergence between the two Stenocephalemys species (*Hylomyscus alleni* and *Mastomys natalensis*) is surprising given that they are contained within the same division.

Table 4.6. Estimated mean pairwise comparisons of evolutionary distances (*d* value) within and between species groups from the Rattus division for exon 6 of *Zp3*. RattusN signifies *Rattus norvegicus* species group. RattusE represents the *Rattus exulans* species group, RattusL represents the *Rattus leucopus* group (largely New Guinean species), and RattusF represents the *Rattus fuscipes* species group (Australasian species). The non-*Rattus* group, for the purposes of this study contain those species within the Rattus division that are not of the *Rattus* genus, namely *Bunomys, Bandicota* and *Paruromys*.

	Within	Between					
Overall: 0.0108		Mus	RattusN	Non-rattus	RattusE	RattusL	RattusF
Mus	n/a						
RattusN	n/a	0.0163					
Non-Rattus	0.0333	0.0222	0.0332				
RattusE	n/a	0.0164	0.0330	0.0164			
RattusL	0.0000	0.0164	0.0330	0.0164	0.0000		
RattusF	0.0000	0.0164	0.0330	0.0164	0.0000	0.0000	

Within the Rattus division species groups (see section 4.2), the highest mean pairwise comparison occurred within the non-Rattus group (0.0333). Between the RattusN group (containing the single species of *Rattus norvegicus*) and the other Rattus groups, the rate of divergence was 0.033 compared to 0.000 between the RattusE, RattusL and RattusF groups.

4.3.3.2 Intron 6 of Zp3

Table 4.7. Estimated evolutionary distances (*d* value) within and between the African and South-east Asian groups for intron 6 of *Zp3*. The African group contains the divisions of Dasymys, Aethomys, Arvicanthis and Stenocephalemys. The South-east Asian group contains the divisions of Dacnomys, Maxomys and Rattus. *Apodemus chevrieri*, occurring in China, has not been included in the South-east Asian group. RattusN signifies *Rattus norvegicus*, the laboratory rat.

	Within	Between				
		Mus	RattusN	Apodemus	African	SE Asian
Mus	n/a					
RattusN	n/a	0.1814				
Apodemus	n/a	0.1217	0.1413			
African	0.1164	0.1215	0.1574	0.1019		
SE Asian	0.0500	0.1923	0.0426	0.1480	0.1690	

Within the African and South-east Asian divisions, the mean pairwise distances were high relative to exon 6. All between group *d* values were greater than 0.1, with the exception of RattusN versus South-east Asian divisions. The highest *d* value was 0.1923 between Mus and the South-east Asian divisions. Between Mus and Apodemus and the African group, the *d* value was 0.121, suggesting that Mus is more closely related to Apodemus and African species than it is to South-east Asian species as previously found by others (Watts & Baverstock 1995; Steppan *et al.* 2005).

Table 4.8. Estimated mean pairwise comparisons of evolutionary distances (*d* value), within and between divisions for intron 6 of *Zp3*. RattusN signifies *Rattus norvegicus*, the laboratory rat. 'Steno' is an abbreviation for the Stenocephalemys division.

Overall: 0.1096	Within	Between								
		Mus	RattusN	Dasymys	Aethomys	Arvicant	Steno	Apode	Dacnomys	Maxomys
Mus	n/a									
RattusN	n/a	0.1814								
Dasymys	n/a	0.0956	0.1340							
Aethomys	0.0371	0.0712	0.1245	0.0456						
Arvicanthis	0.0959	0.1472	0.1865	0.0994	0.1015					
Stenocephalemys	0.1121	0.1840	0.1895	0.1482	0.1524	0.1962				
Apodemus	n/a	0.1217	0.1413	0.0734	0.0733	0.1136	0.1473			
Dacnomys	0.0685	0.2501	0.1051	0.1935	0.1792	0.2174	0.2674	0.1876		
Maxomys	0.1147	0.1727	0.0574	0.1268	0.1069	0.1802	0.2020	0.1250	0.1367	
Rattus	0.0169	0.1834	0.0281	0.1379	0.1280	0.1943	0.1952	0.1431	0.1034	0.0645

Within each division, the mean pairwise comparisons varied, ranging from 0.0169 within the Rattus division to 0.1147 within the Maxomys division. Mean *d* values between divisions ranged from 0.0281 (RattusN versus Rattus) to 0.2674 (Stenocephalemys versus Dacnomys). Apodemus showed lower *d* values in mean pairwise comparisons with the African divisions of Dasymys and Aethomys, than it did with the other divisions.

Table 4.9. Estimated pairwise comparisons of evolutionary distances (*d* value) between eight species from the African group for intron 6 of *Zp3*.

Overall mean: 0.1164	Between							
	D.incom	*M.nama*	*A.chrys*	*A.inept*	*L.gris*	*R.pum*	*H.alleni*	*M.natal*
Dasymys incomtus								
Micaelamys namaquensis	0.0562							
Aethomys chyrsophilus	0.0320	0.0399						
Aethomys ineptus	0.0485	0.0565	0.0150					
Lemniscomys griselda	0.0747	0.0829	0.0576	0.0749				
Rhabdomys pumilio	0.1242	0.1434	0.1154	0.1349	0.0959			
Hylomyscus alleni	0.1994	0.2310	0.1904	0.2114	0.2197	0.2878		
Mastomys natalensis	0.0970	0.1055	0.0786	0.0973	0.1055	0.1719	0.1121	

A closer look at the pairwise comparisons within the African divisions shows that *Hylomyscus alleni* has the greater *d* value in pairwise comparisons than the other species. Its lowest *d* value is in the pairwise comparison with *Mastomys natalensis* (0.1121). The highest *d* value occurred between *H. alleni* and *Rhabdomys pumilio* (0.2878). Between *Dasymys* and species from the *Aethomys* genus the *d* values are quite low, compared to comparisons with species from the Stenocephalemys division (*Hylomyscus alleni* and *Mastomys natalensis*).

Table 4.10. Estimated mean pairwise comparisons of evolutionary distances (*d* value) within and between species groups from the Rattus division for intron 6 of *Zp3*. RattusN signifies *Rattus norvegicus* species group. RattusE represents the *Rattus exulans* species group, RattusL represents the *Rattus leucopus* group (largely New Guinean species), and RattusF represents the *Rattus fuscipes* species group (Australasian species). The non-Rattus group contains those species within the Rattus division that are not of the *Rattus* genus, namely *Bunomys, Bandicota* and *Paruromys*.

Overall: 0.0406	Within	Between					
		Mus	RattusN	Non-rattus	RattusE	RattusL	RattusF
Mus	n/a						
RattusN	n/a	0.2153					
Non-Rattus	0.0187	0.2152	0.0241				
RattusE	n/a	0.1975	0.0498	0.0383			
RattusL	0.0078	0.2085	0.0283	0.0254	0.0455		
RattusF	0.0031	0.2043	0.0259	0.0230	0.0429	0.0054	

The large Rattus division shows relatively low mean *d* values in pairwise comparisons between species groups. Within species groups, the RattusL and RattusF groups have *d* values less than 0.01, while the non-Rattus group has 0.0187. Between the RattusL and RattusF species group, the *d* value is 0.0054. Other mean pairwise *d* values are less than 0.05, with the exception of all species groups versus Mus.

As stated previously, in computing estimates of evolutionary distances, alignment gaps caused by insertions and deletions are ignored by the method employed to estimate pairwise distances (see Chapter 3.3.3.2). Therefore, nucleotide estimations are conservative estimates only. The nucleotide sequence for intron 6 contained a number of insertions and deletions, which are listed according to species in Table 4.11.

Table 4.11. Species that have either a deletion or an insertion, or both, within intron 6 of *Zp3*, together with the number of nucleotides (nts) involved are listed. Position numbers correspond to the nucleotide sequence in Appendix 2.

Species	Deletion (nts)	Insertion (nts)	Position
Apodemus chevrieri & *Aethomys* spp.		1	157
A. chevrieri & *Rattus exulans*		2	86-87
A. chevrieri & species from the Stenocephalemys division	1		192
A. chevrieri, African divisions (exception *Mastomys natalensis*), *Mus musculus* & *Maxomys bartelsii*		4	163-166
Aethomys spp.	1	5	142, 176-180
All *Rattus* spp. (exceptions *R. exulans* and *R. norvegicus*)		3	144-146
Hylomyscus alleni		1	211
Leopoldamys spp.	7	4, 1	113-120, 101-104, 95
Mastomys natalensis		1	197
Maxomys bartelsii		3	174-176
Maxomys bartelsii & African divisions		1	113
Maxomys hellwaldii		1	187
Paruromys dominator		2	156-157
Niviventer fulvescens		1	146
Rattus exulans	2		135-136
Rattus leucopus		3	211-213
Rhabdomys pumilio	5		182-187
Species from the Arvicanthis & Stenocephalemys divisions	4		87-90
Species from the Dasymys, Aethomys & Arvicanthis divisions	1		125

4.3.3.3 Exon 7 of Zp3

The nucleotide sequence for exon 7 varied in length from 103 nucleotides in *Lemniscomys griselda* to 112 nucleotides in five species of *Rattus* and *Maxomys hellwaldii*. This variation is due to the presence of indels. Within the Arvicanthis division, *Lemniscomys griselda* has a unique six base pair (bp) deletion relative to the *Mus* sequence. *Maxomys hellwaldii* has a three nucleotide insertion, and five species of *Rattus* (all Australian occurring members of the RattusF species group) have a different three nucleotide insertion. The presence of these indels suggest that estimations of evolutionary distances will generally be conservative.

Table 4.12. Estimated evolutionary distances (*d* value) within and between the African and South-east Asian groups for exon 7 of *Zp3*. The African group contains the divisions of Dasymys, Aethomys, Arvicanthis and Stenocephalemys. The South-east Asian group contains the divisions of Dacnomys, Maxomys and Rattus. *Apodemus chevrieri*, occurring in China, has not been included in the South-east Asian group. RattusN signifies *Rattus norvegicus*, the laboratory rat.

	Within	Between				
		Mus	RattusN	Apodemus	African	SE Asian
Mus	n/a					
RattusN	n/a	0.0875				
Apodemus	n/a	0.0672	0.1085			
African	0.0756	0.0628	0.1056	0.0927		
SE Asian	0.0321	0.0698	0.0509	0.1024	0.0906	

The *d* value within the African divisions is 0.0756 while it is relatively low within the South-east Asian group of divisions (0.0321). Most mean pairwise comparisons between groups shows mostly higher *d* values than for exon 6, although two are similar (Mus versus Apodemus and African group versus Apodemus).

Table 4.13. Estimated mean pairwise comparisons of evolutionary distances (*d* value), within and between divisions for exon 7 of *Zp3*. RattusN signifies *Rattus norvegicus*, the laboratory rat. 'Steno' is an abbreviation for the Stenocephalemys division.

	Within	Between								
Overall: 0.0635		Mus	RattusN	Dasymys	Aethomys	Arvicanthis	Steno	Apodemus	Dacnomys	Maxomys
Mus	n/a									
RattusN	n/a	0.0875								
Dasymys	n/a	0.0475	0.0875							
Aethomys	0.0389	0.0380	0.0911	0.0412						
Arvicanthis	0.0614	0.0808	0.1063	0.0543	0.0672					
Stenocephalemys	0.0996	0.0897	0.1359	0.0892	0.1001	0.1063				
Apodemus	n/a	0.0672	0.1085	0.0673	0.0772	0.0956	0.1256			
Dacnomys	0.0188	0.0409	0.0706	0.0572	0.0562	0.0800	0.0992	0.0740		
Maxomys	0.0677	0.0723	0.0529	0.0724	0.0760	0.0990	0.1230	0.1031	0.0492	
Rattus	0.0181	0.0752	0.0468	0.0665	0.0758	0.0955	0.1327	0.1079	0.0488	0.0526

Within divisions, mean pairwise comparisons were generally higher than those for exon 6, although with some divisions the *d* value was lower. The Aethomys division had a within mean *d* value of 0.0389 compared to 0.0450 for exon 6. The Stenocephalemys division had a within pairwise comparison of 0.0996 for exon 7, while it was 0.1059 for exon 6. Between mean pairwise comparisons varied considerably and not all comparisons showed an increase in relation to that of exon 6. Mean *d* values ranged from 0.0380 (Mus versus Aethomys) to 0.1359 (RattusN versus Stenocephalemys).

Table 4.14. Estimated pairwise comparisons of evolutionary distances (*d* value) between eight species from the African group for exon 7 of *Zp3*.

Overall mean: 0.0756	Between							
	D.incom	*M.nama*	*A.chrys*	*A.inept*	*L.gris*	*R.pum*	*H.alleni*	*M.natal*
Dasymys incomtus								
Micaelamys namaquensis	0.0281							
Aethomys chrysophilus	0.0477	0.0583						
Aethomys ineptus	0.0477	0.0583	0.0000					
Lemniscomys griselda	0.0511	0.0730	0.0730	0.0730				
Rhabdomys pumilio	0.0575	0.0477	0.0681	0.0681	0.0614			
Hylomyscus alleni	0.1107	0.1231	0.1116	0.1116	0.1313	0.1221		
Mastomys natalensis	0.0677	0.0777	0.0882	0.0882	0.0943	0.0774	0.0996	

A closer inspection of the African species shows that the *Aethomys* species have a 0.0000 rate of divergence between them and 0.0583 between them and *Micaelamys namaquensis* (all three of members of the Aethomys division). *Hylomyscus alleni* has the highest pairwise comparisons with all other species (> 0.1), including between *H. alleni* and *Mastomys natalensis* (0.0996). All other pairwise comparisons (not involving *H. alleni*) are below 0.1, ranging from 0.0281 (*Dasymys* versus *Micaelamys*) to 0.0943 (*Lemniscomys griselda* versus *M. natalensis*).

Table 4.15. Estimated mean pairwise comparisons of evolutionary distances within and between species groups from the Rattus division for exon 7 of *Zp3*. RattusN signifies *Rattus norvegicus* species group. RattusE represents the *Rattus exulans* species group, RattusL represents the *Rattus leucopus* group (largely New Guinean species), and RattusF represents the *Rattus fuscipes* species group (Australasian species). The non-Rattus group, for the purposes of this study contain those species within the Rattus division that are not of the *Rattus* genus, namely *Bunomys*, *Bandicota* and *Paruromys*.

Overall: 0.0281	Within	Between					
		Mus	RattusN	Non-rattus	RattusE	RattusL	RattusF
Mus	n/a						
RattusN	n/a	0.0875					
Non-Rattus	0.0000	0.0672	0.0377				
RattusE	n/a	0.0876	0.0378	0.0477			
RattusL	0.0162	0.0773	0.0506	0.0186	0.0609		
RattusF	0.0036	0.0752	0.0493	0.0111	0.0555	0.0112	

Compared to exon 6, all pairwise comparisons were higher for exon 7, with the exception of the within mean pairwise comparisons of the non-Rattus species (0.0000). Within the species groups, *d* values were generally low, ranging from 0.0000 within the non-Rattus group to 0.0162 within the RattusL group. Between species groups, mean pairwise comparisons ranged from 0.0112 between RattusL and RattusF species groups, to 0.0609 between RattusE and the RattusF group. All mean pairwise

comparisons between Mus and these species groups were relatively high, ranging from between 0.0672 (versus non-Rattus) to 0.0876 (versus RattusE).

4.3.4 Estimated amino acid sequence divergence rates and sequence comparison.

To estimate the rate of divergence, or the evolutionary distance, of the amino acid sequence between species, a single *P* distance has been calculated. This method calculates the number of amino acid changes between two sequences then divides this number by the total number of amino acid sites compared. The mean *P* distance is then calculated for within and between divisions, species groups and species. *P* values are presented as a percentage.

The predicted amino acid sequence for the 28 species investigated in this Chapter is shown in Fig. 4.2. Included in the alignment are the amino acid sequences for *Mus musculus*, *Rattus norvegicus* and *Hylomyscus alleni*. The amino acid changes have been plotted against two hypothetical phylogenies, with a consensus sequence being used as a reference sequence (or outgroup sequence) (Figs. 4.3 and 4.4).

The coding sequences of the partial exon 6 and 7 provided a predicted amino acid sequence of between 55 residues (*Lemniscomys griselda*) and 58 residues in five species of Australian occurring *Rattus* species (RattusF group) and *Maxomys hellwaldii*. Exon 6 of m*Zp3* encodes 31 amino acids, of which the last 21 residues have been determined, and exon 7 of m*Zp3* encodes 46 amino acids, of which the first 33-37 residues have been determined in this project.

All five cysteine residues that are potential sites for disulfide bonding (Cys-301, Cys-320, Cys-322, Cys-323 and Cys-328) and all three potential N-linked glycosylation sites (Asn-304, Asn-327 and Asn-330) identified in the mouse (Boja *et al.* 2003) have been conserved in all species investigated.

Codon 21 is shared between exon 6 and 7, as the exon boundary is positioned between nucleotides 62 and 63 (exon nucleotides only). For analytical reasons, the first nucleotide for exon 7 has been included in the nucleotides of exon 6, providing for 21 codons for exon 6 coding region and up to 37 codons for the exon 7 coding region.

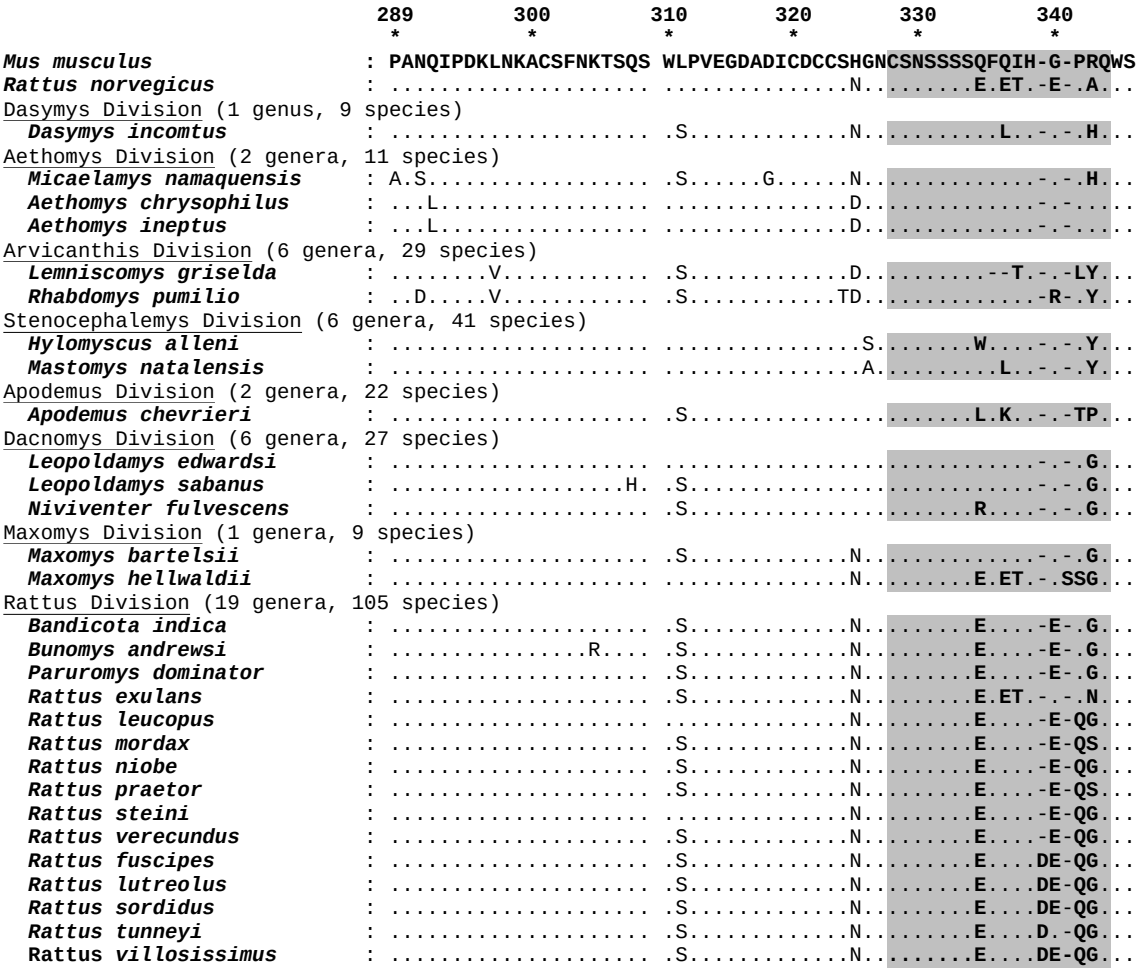


Fig 4.2. Alignment of the predicted amino acid sequence for the region encoded by the partial exon 6 and exon 7 of *Zp3* for a range of African, Eurasian and South-east Asian murine species. Species are grouped under the divisions defined by Musser & Carleton (2005). In parenthesis, next to the name of each division, are the number of genera and extant species per division (according to Musser & Carleton, 2005). Amino acid residues are numbered according to the corresponding mouse ZP3 (mZP3) position. Single dots (.) represent conservation of amino acid residues at any given position. The exon 6/7 boundary is indicated by the space between residues 309 to 310. The putative combining-site for sperm, identified by Wassarman & Litscher (1995) is highlighted in grey and text in bold. Each amino acid is represented by its single letter code.

4.3.4.1 Region encoded by exon 6 of Zp3

Table 4.16. Estimated evolutionary distances (*P* value %) within and between the African and South-east Asian groups for amino acid sequence encoded by exon 6 of *Zp3*. The African group contains the divisions of Dasymys, Aethomys, Arvicanthis and Stenocephalemys. The South-east Asian group contains the divisions of Dacnomys, Maxomys and Rattus. *Apodemus chevrieri*, occurring in China, has not been included in the South-east Asian group. RattusN signifies *Rattus norvegicus*, the laboratory rat.

Overall: 2.71

	Within	Between				
		Mus	RattusN	Apodemus	African	SE Asian
Mus	n/a					
RattusN	n/a	0.00				
Apodemus	n/a	0.00	0.00			
African	7.48	4.17	4.17	4.17		
SE Asian	0.95	0.48	0.48	0.48	4.64	

The mean *P* distances within groups of divisions showed considerable differences between the African and the South-east Asian divisions. Within the South-east Asian group there is a small *P* value, reflecting the single amino acid difference between all 15 species. The African group of divisions shows a higher rate of divergence.

Table 4.17. Estimated mean pairwise comparisons of evolutionary distances (*P* value %), within and between divisions for amino acid sequence encoded by exon 6 of *Zp3*. RattusN signifies *Rattus norvegicus*, the laboratory rat. 'Steno' is an abbreviation for the Stenocephalemys division.

Overall: 2.71

	Within	Between								
		Mus	RattusN	Dasymys	Aethomys	Arvicanthis	Steno	Apodemus	Dacnomys	Maxomys
Mus	n/a									
RattusN	n/a	0.00								
Dasymys	n/a	0.00	0.00							
Aethomys	9.52	6.35	6.35	6.35						
Arvicanthis	4.76	7.14	7.14	7.14	12.70					
Stenocephalemys	0.00	0.00	0.00	0.00	6.35	7.14				
Apodemus	n/a	0.00	0.00	0.00	6.35	7.14	0.00			
Dacnomys	3.17	1.59	1.59	1.59	7.94	8.73	1.59	1.59		
Maxomys	0.00	0.00	0.00	0.00	6.35	7.14	0.00	0.00	1.59	
Rattus	0.63	0.32	0.32	0.32	6.67	7.46	0.32	0.32	1.90	0.32

Within divisions, the *P* value was 0.00 for Stenocephalemys, and Maxomys. Within the Aethomys division, the evolutionary distance was high (9.52) while within Arvicanthis it was 4.76. The low *P* value within Stenocephalemys is in contrast to the nucleotide rate of divergence for exon 6 (10.59), suggesting that the amino acid differences within this division were silent. Between African divisions, *P* values

ranged from 6.35 (Aethomys versus Dasymys and Stenocephalemys) to 12.70 (Arvicanthis versus Aethomys). Between South-east Asian divisions, the evolutionary distance is quite low, less than 2.00. Between Apodemus and the South-east Asian divisions, the *P* value was also below 2.00.

Table 4.18. Estimated pairwise comparisons of evolutionary distances (*P* values %) between eight species from the African group for the amino acid sequence encoded by exon 6 of *Zp3*.

Overall mean: 7.48	Between							
	D.incom	*M.nama*	*A.chrys*	*A.inept*	*L.gris*	*R.pum*	*H.alleni*	*M.natal*
Dasymys incomtus								
Micaelamys namaquensis	9.52							
Aethomys chyrsophilus	4.76	14.29						
Aethomys ineptus	4.76	14.29	0.00					
Lemniscomys griselda	4.76	14.29	9.52	9.52				
Rhabdomys pumilio	9.52	14.29	14.29	14.29	4.76			
Hylomyscus alleni	0.00	9.42	4.76	4.76	4.76	9.52		
Mastomys natalensis	0.00	9.52	4.76	4.76	4.76	9.52	0.00	

A closer inspection of the African group shows that between the two Stenocephalemys division species (*Hylomyscus alleni* and *Mastomys natalensis*), the *P* value was 0.00. Between these two species and *Dasymys incomtus*, the *P* value was also 0.00. Evolutionary distances ranged from 4.76 between a number of difference species to 14.29 (*Micaelamys namaquensis* versus Aethomys and Arvicanthis division species).

Table 4.19. Estimated mean pairwise comparisons of evolutionary distances (*P* values %) within and between species groups from the Rattus division for the region encoded by exon 6 of *Zp3*. RattusN signifies *Rattus norvegicus* species group. RattusE represents the *Rattus exulans* species group, RattusL represents the *Rattus leucopus* group (largely New Guinean species), and RattusF represents the *Rattus fuscipes* species group (Australasian species). The non-Rattus group, for the purposes of this study contain those species within the Rattus division that are not of the Rattus genus, namely *Bunomys, Bandicota* and *Paruromys*.

Overall: 0.56	Within	Between					
		Mus	RattusN	Non-rattus	RattusE	RattusL	RattusF
Mus	n/a						
RattusN	n/a	0.00					
Non-Rattus	3.17	1.59	1.59				
RattusE	n/a	0.00	0.00	1.59			
RattusL	0.00	0.00	0.00	1.59	0.00		
RattusF	0.00	0.00	0.00	1.59	0.00	0.00	

All *P* values within the Rattus division were low (between 0.00 to 1.59). Within the non-Rattus (*Bandicota, Bunomys* and *Paruromys)* species, the distance was 3.17 and between this group and other species groups it was 1.59.

A comparison of the amino acid sequence reflects the low divergence rates. Relative to the mouse sequence, *Dasymys incomtus* and *Apodemus chevrieri* share the same sequence, as do the two Stenocephalemys division species. Within the Aethomys division, *Micaelamys namaquensis* has an alanine (A) in position 289, where other species have a proline (*P*). *M. namaquensis* also has a serine in position 291. Other species, with the exception of *Rhabdomys pumilio* (aspartic acid, D), have an asparagine (N) in this position. The two *Aethomys* species share a leucine in position 292, whereas other species have a glutamine (Q). Within the Arvicanthis division, the two species investigated share a valine in position 297, whereas other species have a leucine in this position. In addition, as mentioned before, *R. pumilio* has an aspartic acid in position 291.

Within the South-east Asian species, *Leopoldamys sabanus* has a histidine (H) in position 308. All other species have a glutamine (Q) in this position. *Bunomys andrewsi* has an arginine in position 305 whereas other species have a lysine (K). Apart from these two, all other South-east Asian species share an identical amino acid sequence with *Mus musculus* and *Rattus norvegicus*.

4.3.4.2 Region encoded by exon 7

The five serine residues (Ser-329, Ser-331 to Ser-334) present in the mouse ZP3 and implicated in providing the *O*-linked glycosylation sites (serine or threonine residues) thought to be involved in sperm-ZP binding are conserved in all 28 species investigated, including *Hylomyscus alleni*.

A noticeable difference, in the sequence alignment of the exon 7 coding region among this group of species, is the presence of three indels. Firstly, *Lemniscomys griselda* has a two amino acid deletion occurring at positions 336 and 337 relative to the mouse sequence. Secondly, *Maxomys hellwaldii* has a

single amino acid insertion (serine) occurring between residues 340 and 341. Thirdly, all five Australian *Rattus* species (all belonging to the RattusF species group), have an aspartic acid insertion between residues 339 and 340, relative to the mouse and rat. Therefore, in pairwise comparisons involving these seven species, the estimated evolutionary distances have been calculated without reference to these indels and are hence only conservative estimates.

Table 4.20. Estimated evolutionary distances (*P* value %) within and between the African and South-east Asian groups for amino acid sequence encoded by exon 7 of *Zp3*. The African group contains the divisions of Dasymys, Aethomys, Arvicanthis and Stenocephalemys. The South-east Asian group contains the divisions of Dacnomys, Maxomys and Rattus. *Apodemus chevrieri*, occurring in China, has not been included in the South-east Asian group. RattusN signifies *Rattus norvegicus*, the laboratory rat.

Overall: 11.76

	Within	Between				
		Mus	RattusN	Apodemus	African	SE Asian
Mus	n/a					
RattusN	n/a	16.67				
Apodemus	n/a	13.89	22.22			
African	12.19	9.13	18.53	16.42		
SE Asian	7.38	13.75	13.61	16.11	15.32	

The within group evolutionary distances were higher than for exon 6, with the African divisions showing the highest divergence (12.19).

Table 4.21. Estimated mean pairwise comparisons of evolutionary distances (*P* value %), within and between divisions for exon 7 of *Zp3*. RattusN signifies *Rattus norvegicus*, the laboratory rat. 'Steno' is an abbreviation for the Stenocephalemys division.

Overall: 11.76

	Within	Between								
		Mus	RattusN	Dasymys	Aethomys	Arvicanthis	Steno	Apodemus	Dacnomys	Maxomys
Mus	n/a									
RattusN	n/a	16.67								
Dasymys	n/a	11.11	16.67							
Aethomys	7.41	5.56	17.59	09.26						
Arvicanthis	11.76	14.30	19.93	12.83	12.39					
Stenocephalemys	8.33	8.33	19.44	13.89	12.96	16.42				
Apodemus	n/a	13.89	22.22	13.89	16.67	17.08	16.67			
Dacnomys	3.70	5.56	18.52	10.19	8.95	13.34	10.65	12.04		
Maxomys	13.89	12.50	12.50	11.11	12.50	15.65	16.67	16.67	10.19	
Rattus	4.28	15.56	12.78	13.33	14.88	17.06	19.63	16.85	11.98	10.93

Within divisions, *P* values were all higher than that for exon 6, ranging from 3.70 within the Dacnomys division to 13.89 within the Maxomys division. Between divisions, the *P* values were also quite high, ranging from 5.56 between Mus and Aethomys/Dacnomys to 22.22 between RattusN and Apodemus.

Table 4.22. Estimated pairwise comparisons of evolutionary distances (*P* values %) between eight species from the African group of divisions for the region encoded by exon 7 of *Zp3*.

Overall mean 12.19	Between							
	D.incom	*M.nama*	*A.chrys*	*A.inept*	*L.gris*	*R.pum*	*H.alleni*	*M.natal*
Dasymys incomtus								
Micaelamys namaquensis	5.56							
Aethomys chyrsophilus	11.11	11.11						
Aethomys ineptus	11.11	11.11	0.00					
Lemniscomys griselda	11.76	14.71	11.76	11.76				
Rhabdomys pumilio	13.89	13.89	11.11	11.11	11.76			
Hylomyscus alleni	16.67	16.67	11.11	11.11	17.65	16.67		
Mastomys natalensis	11.11	16.67	11.11	11.11	14.71	16.67	8.33	

A closer inspection of the species from the African divisions shows the pairwise comparisons of evolutionary distances were mostly above 1.00. Between *Dasymys incomtus* and *Micaelamys namaquensis* the *P* value was 5.56, between the two *Aethomys* species it was 0.00, and between the two Stenocephalemys division species, the *P* value was 8.33.

Table 4.23. Estimated mean pairwise comparisons of evolutionary distances (*P* values %) within and between species groups from the Rattus division for the region encoded by exon 7 of *Zp3*. RattusN signifies *Rattus norvegicus* species group. RattusE represents the *Rattus exulans* species group, RattusL represents the *Rattus leucopus* group (largely New Guinean species), and RattusF represents the *Rattus fuscipes* species group (Australian species). The non-*Rattus* group, for the purposes of this study contain those species within the Rattus division that are not of the Rattus genus, namely *Bunomys, Bandicota* and *Paruromys.*

Overall: 6.56	Within	Between					
		Mus	RattusN	Non-rattus	RattusE	RattusL	RattusF
Mus	n/a						
RattusN	n/a	16.67					
Non-Rattus	0.00	13.89	11.11				
RattusE	n/a	16.67	8.33	11.11			
RattusL	2.96	15.74	12.96	4.63	14.81		
RattusF	1.08	16.11	14.44	3.33	13.33	2.41	

Within the large Rattus division, the *P* values varied. Within the non-*Rattus* species group the *P* values was 0.00, and yet within the RattusL and RattusF species groups, the *P* value was 2.96 and 1.08 respectively. Between groups the *P* value was also relatively high, ranging from 2.41 between RattusE and RattusF, to 16.67 between Mus and RattusN/RattusF.

The partial exon 7 nucleotide sequence encodes for 36 amino acids. The first 25 residues show a small amount of divergence, while the remaining 11 (from residues 335 to 345) show relatively greater variation. The combining-site for sperm has been identified as commencing at residue 328 and ending at residue 343, although this site has varied from time to time (see section 1.2.4.2).

Within the first 25 residues, there are two sites that show variation among species. At position 311, there is either a leucine (L) or a serine (S). Both *Mus* and *Rattus norvegicus* have a leucine in this position, as do the two *Aethomys* species (Stenocephalemys division). *Leopoldamys edwardsi* also has a leucine in this position, as does *Maxomys hellwaldii*. Within the Rattus division, *R. steini* and *R. leucopus* are the only species that have a leucine in this position. All other species have a serine, as do most of the New Guinean and Australasian Old Endemic species.

At position 325, *Mus* has a histidine (H), which is shared by the two Stenocephalemys division species, Apodemus and all three Dacnomys division species. *Rattus norvegicus* has an asparagine (N) in this position, in common with *Dasymys incomtus*, *Micaelamys namaquensis*, the two *Maxomys* division species, and all the Rattus division species. The two *Aethomys* species and the two Arvicanthis division species have an aspartic acid (D) in this position, in common with the majority of the New Guinean and Australasian Old Endemic species.

In addition to the above mentioned sites, *Micaelamys namaquensis* has a glycine (G) in position 318, where all other species, including *Mus,* have an aspartic acid. *Rhabdomys pumilio* has a threonine (T) in position 324, while all other species have a serine. The two Stenocephalemys division species have a

serine (*H. alleni*) and an alanine (*M. natalensis*) in position 326, whereas all other species have a glycine.

It has been observed that between the mouse and the rat (*Rattus norvegicus*) there is a variable region within exon 7, a 10 amino acid stretch from 335 to 344, where five substitutions have occurred between the two species (Scobie *et al.* 1999). It is within this stretch that the most variation is seen in the African and South-east Asian species. The following section reports the differences within the region 335 to 344 within the various divisions.

Dasymys division

This division consists of only one genus with nine species (Musser & Carleton 2005) of which the sequence for only one species has been determined in this study, *Dasymys incomtus*. *D. incomtus* has a leucine (L) in position 337, in common with *Mastomys natalensis*, and a histidine (H) in position 342, in common with *Micaelamys namaquensis*. All other residues are the same as for Mus and Aethomys division species.

Aethomys division

This division consists of 2 genera, with a total of 11 extant species (Musser & Carleton 2005). The sequences for two *Aethomys* and one *Micaelamys* species have been determined. The two *Aethomys* species share an identical sequence within both the exon 6 and exon 7 coding regions. In the variable region of the exon 7 coding region, *M. namaquensis* has a histidine in position 342, in common with *Dasymys incomtus*. Within this region, the two *Aethomys* species share an identical sequence with the mouse.

Arvicanthis division

This division consists of 6 genera and 29 extant species (Musser & Carleton 2005), of which the amino acid sequence of one species each from two genera have been determined. *Lemniscomys griselda* and *Rhabdomys pumilio* share one amino acid difference from that of the mouse, at position 342 where they

have a tyrosine (Y), in common with the two species from the Stenocephalemys division. The mouse has an arginine in this position, and the South-east Asian species have a glycine. *L. griselda* has a unique two amino acid deletion at positions 336 and 337. *L. griselda* also has a threonine in position 338, in common with *R. norvegicus* and a leucine in position 341. Amino acids in position 341 vary within murine species, although none other than *L. griselda* have a leucine in this position. *R. pumilio* has an arginine (R) in position 340, not shared with other species.

Stenocephalemys division

This division consists of 6 genera with 41 extant species (Musser & Carleton 2005), of which the amino acid sequence has been determined in one representative species, *Mastomys natalensis*. As stated previously, available on GenBank is the nucleotide and predicted amino acid sequence for *Hylomyscus alleni* (Jansa *et al.* 2002), and for comparative purposes, this sequence has been included in analyses. The two species share a tyrosine (Y) with the Arvicanthis division species at position 342. At position 335, *H. alleni* has a tryptophan (W). This position tends to be variable within murine species, with the Rattus division having a glutamic acid (E), while Mus has a glutamine (Q). *M. natalensis* has a leucine (L) in position 337, in common with another African species, *Dasymys incomtus*.

Apodemus division

This division consists of only two genera with a total of 22 extant species (Musser & Carleton 2005). The amino acid sequence has been determined for only one of these species, *Apodemus chevrieri*. Its type locality, according to Musser and Carleton (2005) is western central China. For that reason, it has not been included in the South-east Asian groups of divisions. This species has four unique nucleotide differences from the mouse within the variable region. *A. chevrieri* has a leucine (L) at position 335, a lysine (K) at position 337, a threonine (T) at 341 has and a proline (P) at 342.

Dacnomys division

This division consists of 6 genera and 27 species (Musser & Carleton 2005), all from South-east Asia. The two *Leopoldamys* species share a glycine in position 342 with *Niviventer fulvescens* and other South-east Asian species investigated in this study. Other than that difference, they share the same sequence with that of the mouse. In addition to the Gly-342, *N. fulvescens* has an arginine (R) in position 335, a site that is variable among South-east Asian species.

Maxomys division

This division consists of only one genus, with 9 extant species (Musser & Carleton 2005). *M. bartelsii*, the type species from Java, has an amino acid sequence which is quite distinct from that of *M. hellwaldii* from Sulawesi. *M. bartelsii* shares the identical sequence, from residues 335 to 343, with the two *Leopoldamys* species. *M. hellwaldii* shares the same glycine in position 342, but also shares three amino acids with *Rattus norvegicus* and *R. exulans*, at positions 335 (E), 337 (E) and 338 (T). In addition, *M. hellwaldii* has a serine in position 341, while the *Rattus* species have a glutamine (Q). *M. hellwaldii* also has a serine insertion between residues 340 and 341.

Rattus division

This division consists of 19 genera and 105 extant species (Musser & Carleton 2005). These species occur throughout South-east Asia including the islands of Indonesia, the Malayan peninsula, Sulawesi, New Guinea and Australia. The amino acid sequence from 4 genera and 15 species has been determined, including 12 species of *Rattus*. This latter group has been divided into four species groups (see section 4.2). All 15 species, and *Rattus norvegicus*, have a glutamic acid (E) in position 335. *Rattus exulans* shares a glutamine (Q) in position 337 and a threonine (T) in position 338 with *Rattus norvegicus* and *Maxomys hellwaldii*, but does not share a glutamine (Q) in position 340 with either *R. norvegicus* or the other Rattus division species. The three non-*Rattus* species share a glutamine (Q) in position 337 and 340, and a glycine (G) in position 342, common to most Rattus division species. *Rattus* species that occur in New Guinea (RattusL group) and Australia (RattusF) share a glutamine (Q) in

position 341 and two species from the RattusL group (*R. mordax* and *R. praetor*) have a serine (S) in position 342 rather than the common glycine (G). The five species that occur solely in Australia (*Rattus fuscipes* group) all have an aspartic acid (D) insertion situated between positions 339 and 340.

Summary of amino acid differences

Some discernible patterns to the amino acid variations within these species emerge from these data. All Rattus division species, including *Rattus norvegicus*, have a glutamic acid in positions 335 and 340, in contrast to a glutamine and glycine in the mouse. The majority of South-east Asian species have a glycine in position 342, while the mouse has an arginine. The New Guinean and Australian *Rattus* species have a glutamine in position 341, while all other species have a proline (including the mouse). *R. norvegicus, R. exulans* and *Maxomys hellwaldii* have residues E and T in positions 337 and 338 respectively. Five species of Australian *Rattus* (RattusF species group) have an insertion (aspartic acid), not shared with *Rattus leucopus*, a species that occurs in both New Guinea and Australia. *Maxomys hellwaldii* shares similar amino acid differences with the Rattus division and on that basis, the sequence of Zp3 appears to be more similar to the sequence from that group than to that of *Maxomys bartelsii*.

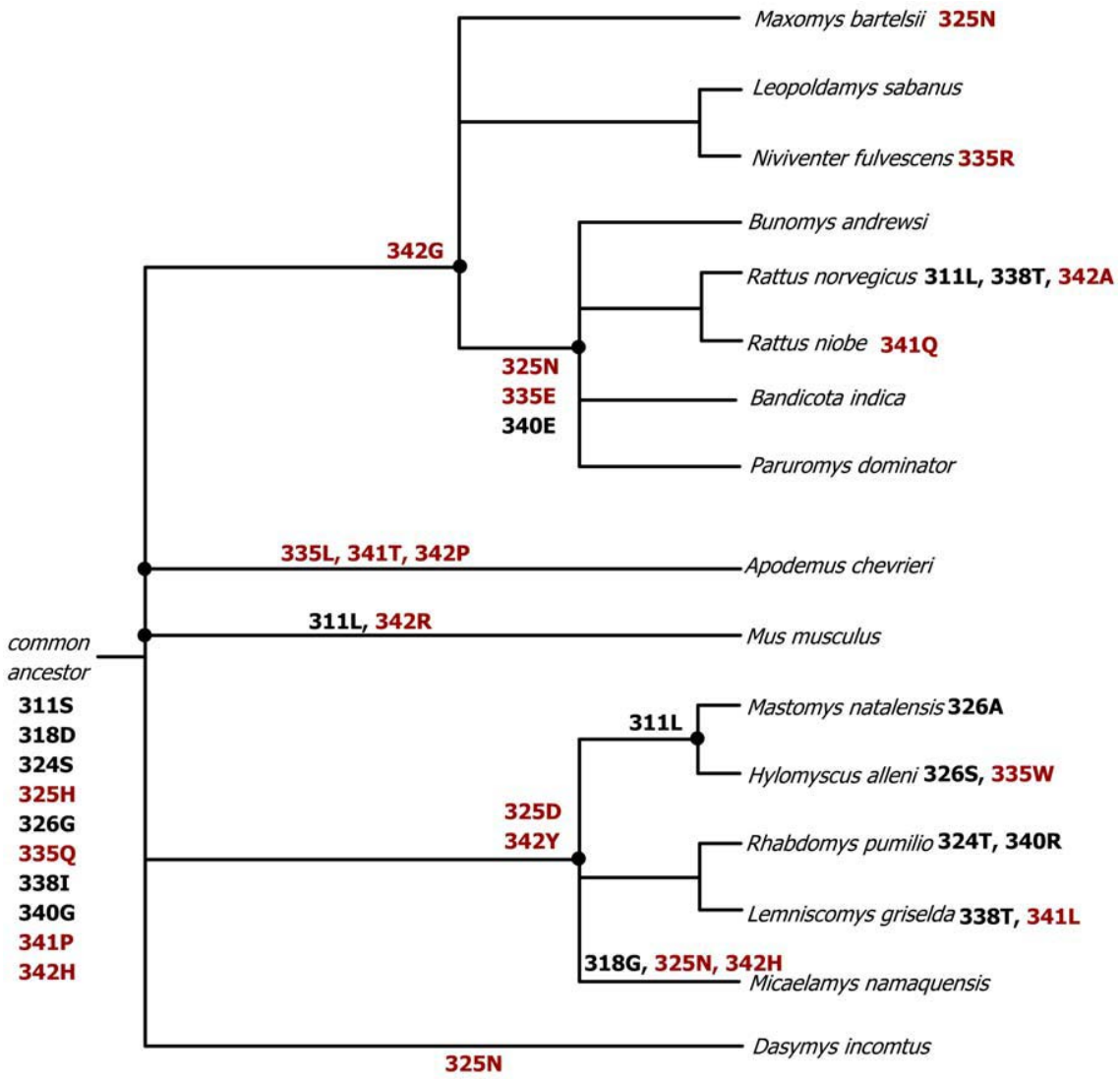


Fig. 4.3 Ancestral reconstruction of the region encoded by exon 7 of *Zp3* from African, Eurasian and South-east Asian murine species. The ancestral sequence of the common ancestor was inferred using the method of Yang *et al.* (1995) and computed using the PAML software (version 3.15: Yang 1997). The proposed phylogeny used is based on microcomplement fixation of albumin (Watts & Baverstock 1995). Amino acid residues that show evidence of parallel evolution on different lineages are highlighted in red.

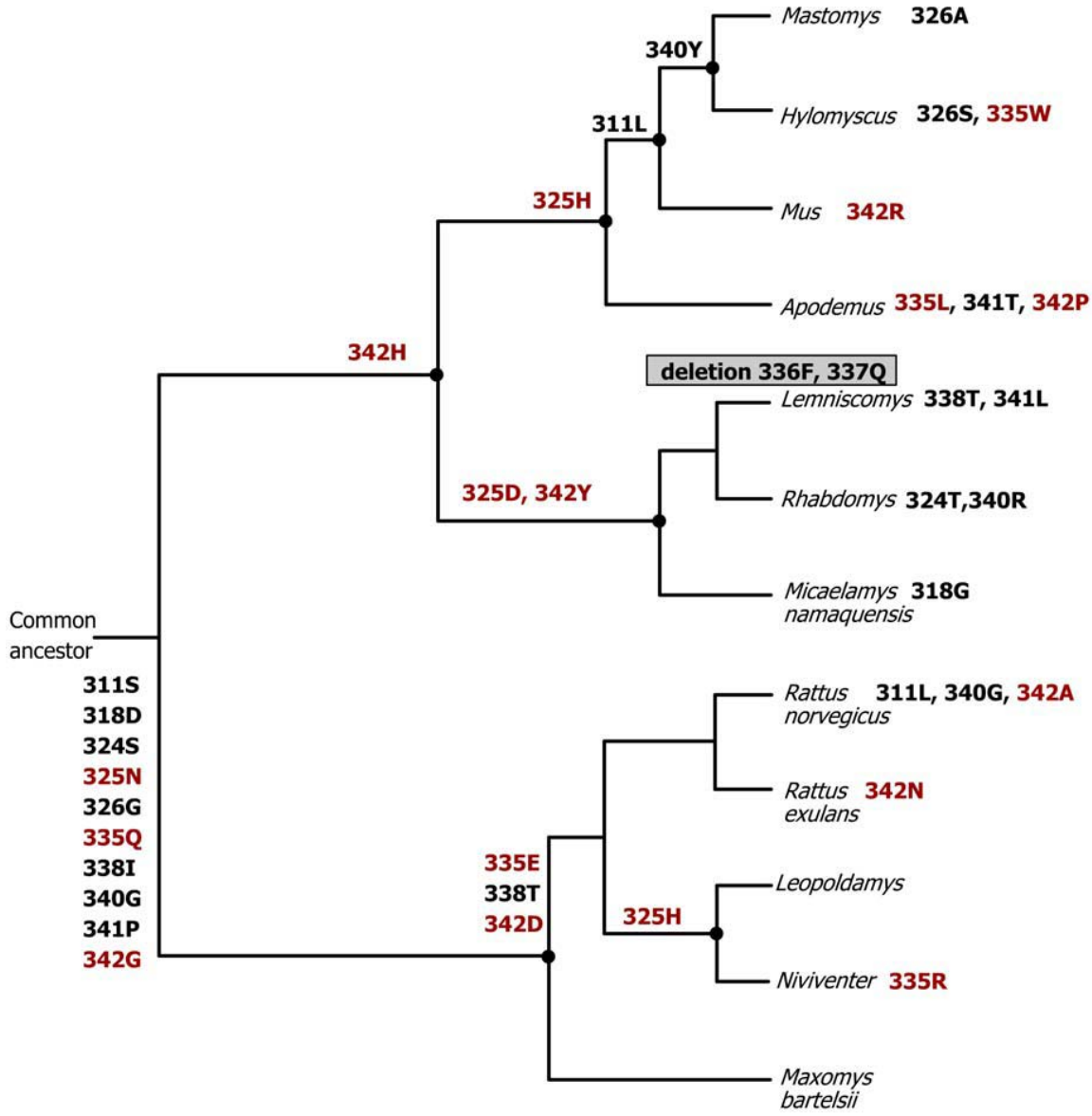


Fig. 4.4. Ancestral reconstruction of the region encoded by exon 7 of *Zp3* from African, Eurasian and South-east Asian murine rodent species. The ancestral sequence of the common ancestor was inferred using the method of Yang *et al.* (1995) and computed using the PAML software (version 3.15: Yang 1997). The proposed phylogeny used in based on nucleotide sequence data from nuclear and mitochondrial genes (Steppan *et al.* 2005). Amino acid residues that show evidence of parallel evolution on different lineages are highlighted in red.

4.3.5 Possible effects of amino acid change

4.3.5.1 Isoelectric points, potential glycosylation sites and mean hydropathic profiles

The isoelectric point, potential glycosylation sites and mean hydropathic profiles were calculated using the same method as was used in Chapter 3.4.1.

Table 4.24. Isoelectric point, relative serine/threonine percentages and the average hydropathy index for selected species of African murines in respect of the region from residues 328 to 345.

	Isoelectric point	Relative serine/threonine composition (%)		Average hydropathy index
		Serine	Threonine	
Mus musculus	8.4	33	0	-1.09
Dasymys incomtus	7.2	33	0	-0.61
Micaelamys namaquensis	7.2	33	0	-1.02
Aethomys chrysophilus	8.4	33	0	-1.09
Lemniscomys griselda	6.9	38	6	-0.97
Rhabdomys pumilio	8.3	33	0	-1.14
Hylomyscus alleni	6.9	33	0	-0.77
Mastomys natalensis	6.9	33	0	-0.51

The isoelectric points ranged from 6.9 in the two Stenocephelamys division species to 8.4 in *Aethomys chrysophilus*. The serine residue percentage is 33% for most African species, with the exception of *Lemniscomys griselda* which has a threonine instead of a serine, and hence has still retained a potential glycosylation site. However, although the average hydropathy index of the species indicates that the region is hydrophilic, there is a variation in the mean hydropathy index, ranging from an average of -0.61 in *Dasymys incomtus* to -1.14 in *Rhabdomys pumilio*. A more detailed analysis of the pattern of hydropathy within the African species is conducted in section 4.3.5.2.

Table 4.25. Isoelectric point, relative serine/threonine percentages and the average hydropathy index for selected species of Eurasian and South-east Asian murines.

	Isoelectric point	Relative serine/threonine composition (%)		Average hydropathy
		Serine	Threonine	
Mus musculus	8.4	33	0	-1.09
Apodemus chevrieri	8.4	33	6	-0.49
Leopoldomys sabanus	7.0	33	0	-0.86
Niviventer fulvescens	8.4	33	0	-0.92
Maxomys bartelsii	7.0	33	0	-0.86
Maxomys hellwaldii	4.6	42	6	-1.09
Bandicota indica	4.6	33	0	-1.03
Paruromys dominator	4.6	33	0	-1.03
Rattus norvegicus	4.3	33	6	-1.2
Rattus exulans	4.6	33	6	-1.32
Rattus steini	4.6	33	0	-1.14
Rattus mordax	4.6	39	0	-1.16
Rattus leucopus	4.6	33	0	-1.14
Rattus fuscipes	4.2	32	0	-1.26
Rattus tunneyi	4.4	32	0	-1.10

The isoelectric points for the Eurasian and South-east Asian murines shows considerably variation, ranging from 4.2 (*Rattus*) to 8.4 (*Apodemus chevrieri*). The species from the Rattus division have isoelectric points that are similar to each other, and to that of *Maxomys hellwaldii* (of the Maxomys division). The mean hydropathy, although all are hydrophilic, also shows variation with Rattus division species having a mean index below -1 while the other South-east Asian species have an index above -1. However, a more detailed analysis of the hydropathy profiles of the Eurasian and South-east Asian species is conducted in section 4.3.5.2.

The isoelectric points of the Australian occurring *Rattus* species (RattusF species group) is considerably lower than that for the Australasian Old Endemic murines investigated in Chapter 3. The Australian *Rattus* species have an isoelectric point of 4.4 while the Australasian Old Endemics is 8.4. However, the mean hydropathy index of the two groups of Australasian murines is similar.

4.3.5.2 Hydropathy profiles

The hydropathy profiles for the African, Eurasian and South-east Asian murines has been calculated using the same method as detailed in Chapters 2.9.2 and 3.4.2.

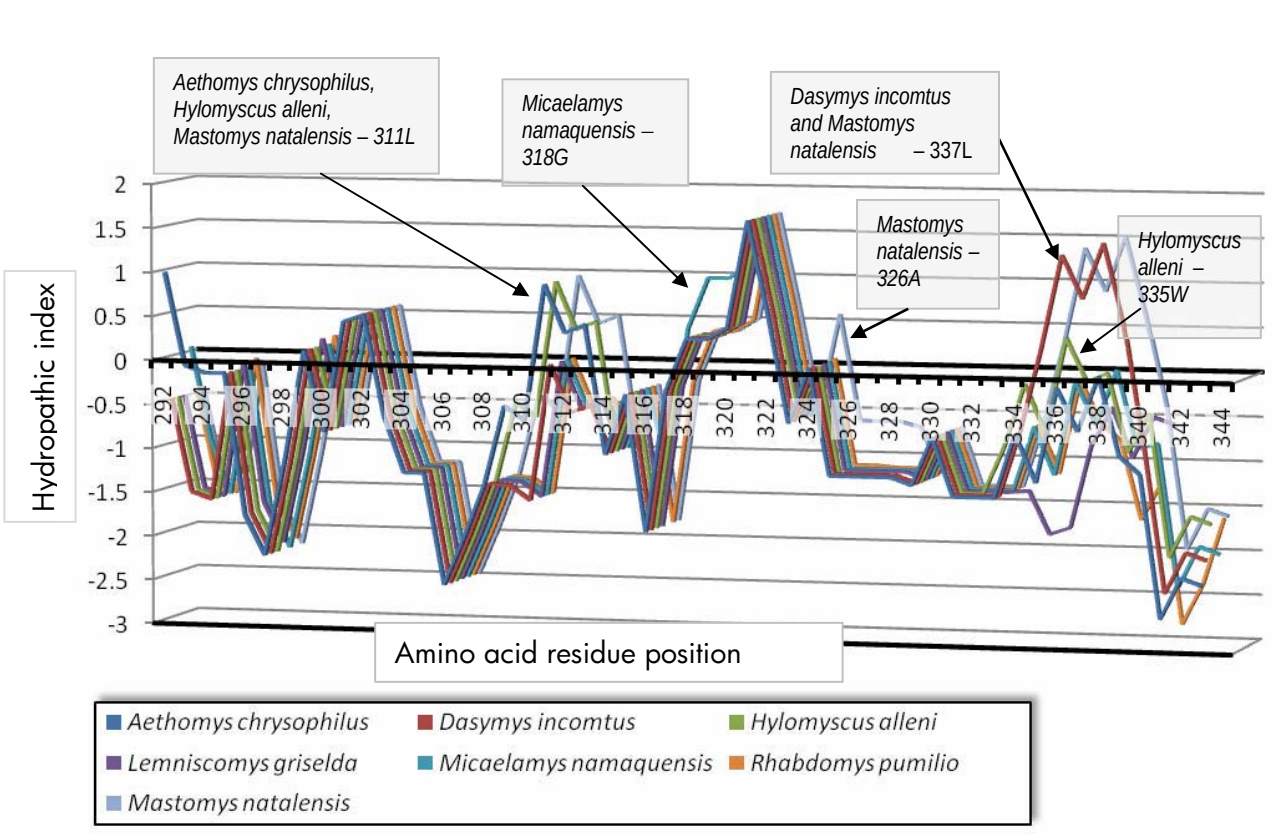


Fig. 4.5. 3D graphical representation of the hydropathic profile of the exon 6 and 7 coding region of *Zp3* from African murines, showing the similarities and differences of hydropathy between species . The hydropathic index was determined using the average of a 5 residue sliding window. Below the line is that part of the protein that is hydrophilic and therefore likely to be on the surface of the protein. Conversely, above the line is that part of the protein that is hydrophobic and likely to be buried inside the protein. The box above the graph shows the species and the amino acid change that increased the hydrophobicity of the region proposed to be involved in sperm-zona pellucida binding.

For the hydropathy profiles, species were selected to represent the common amino acid sequences, and therefore data from all species are not presented. Fig. 4.5 shows the hydropathy profile of African murine species

Within the African murine species, there was some variation in hydropathy around the residue in position 311, due to some species (*Aethomys chrysophilus*, *Hylomyscus alleni* and *Mastomys natalensis*) having the highly hydrophobic leucine in position 311, while other species have a serine (-0.8). Within the

region 328 to 345 the majority of African species had residues that made the region hydrophilic and therefore possibly on the surface of the protein. However for two species, *Dasymys incomtus* and *Mastomys natalensis*, the region between residues 335 and 340 was hydrophobic, due to residue 337 being a leucine (+3.8). The remainder of species had a glutamine (-3.5) in this position. *Hylomyscus alleni* also has a tryptophan (-0.9) in position 335 (majority have a glutamine (-3.5) residue in this position) which caused a variation around that site from hydrophilic to hydrophobic.

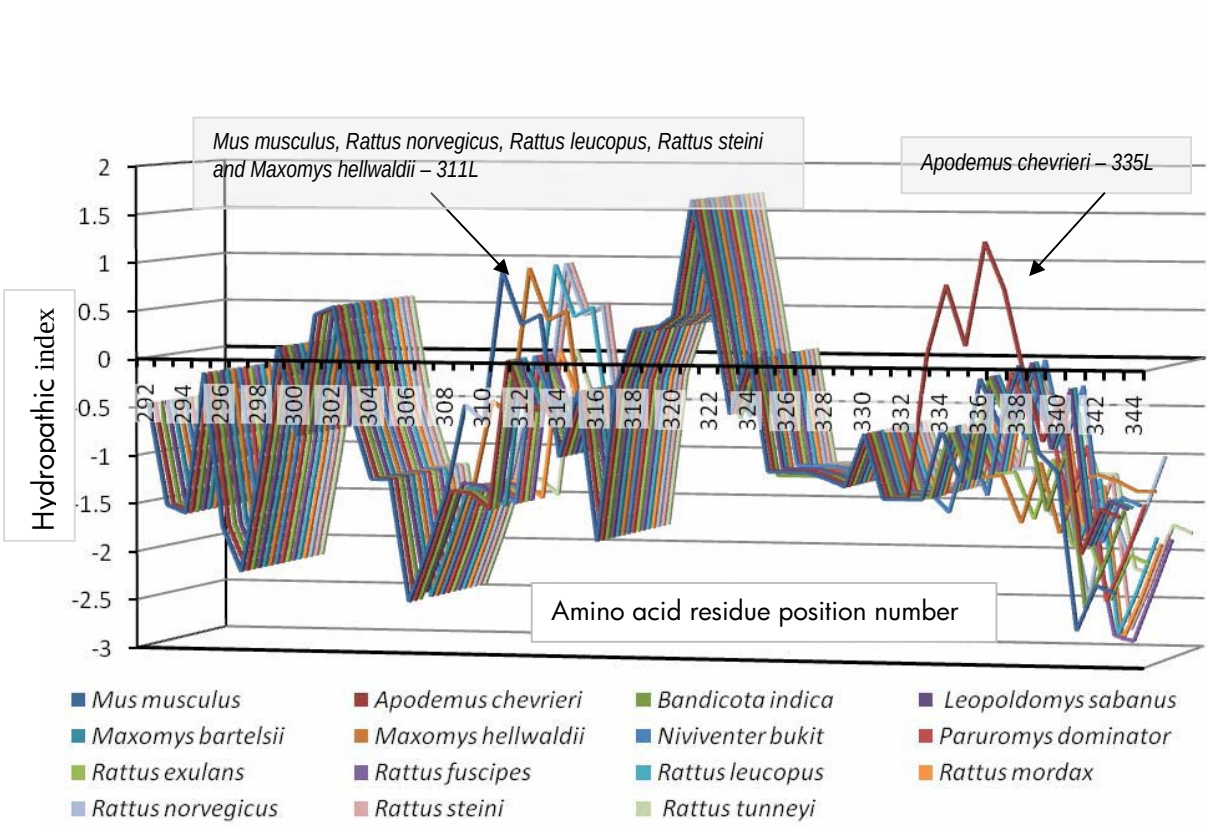


Fig. 4.6. 3D graphical representation of the hydropathic profile of the exon 6 and 7 coding region of *Zp3* from Eurasian and South-east Asian murines, showing the similarities and differences of hydropathy between species . The hydropathic index was determined using the average of a 5 residue sliding window. Below the line is that part of the protein that is hydrophilic and therefore likely to be on the surface of the protein. Conversely, above the line is that part of the protein that is hydrophobic and likely to be buried inside the protein. The box above the graph shows the species and the amino acid change that increased the hydrophobicity of the region proposed to be involved in sperm-zona pellucida binding.

The high level of conservation of hydropathy can be seen in the graph as well as the effect changes have on the hydropathy of the sequence. One region of variation occurs around residue 311. Within this group of species, position 311 contains either a serine (-0.8) or a leucine (+3.8). Five species (*Mus musculus*, *Rattus norvegicus*, *R. leucopus*, *R. steini* and *Maxomys hellwaldii*) have a leucine in this position, changing the area around that residue to hydrophobic. The lack of conservation of hydrophilicity within the region from position 328 to 345 is also highlighted in Fig. 4.6. Although the majority of the residues have remained hydrophilic and therefore possibly on the surface of ZP3, there is still considerable variation as to the degree. The exception is the sequence of *Apodemus chevrieri*, with an amino acid substitution at position 335, from either a glutamine (Q) or glutamic acid (E), both highly hydrophilic (-3.5), to the highly hydrophobic leucine (+3.8). This change may have caused subtle differences in the folding of the glycoprotein.

4.4 Discussion

The hypothesis tested in this chapter is an extension of that in Chapter 3, that there is a high level of sequence divergence within the exon 7 coding region between closely related species, which may contribute to potential species specificity of sperm-ZP binding. In Chapter 3, it was concluded that although exon 7 is evolving at a faster rate than exon 6, the similarity of amino acid sequence between closely related species suggests that this region does not play any role in species specificity of sperm-ZP binding. In the present chapter, the nucleotide and predicted amino acid sequence of exons 6 and 7 of *Zp3* was determined for 28 species from 14 genera of murine rodents from Africa, Eurasia and South-east Asia. More variation was seen within exon 6 in the African divisions than those from the South-east Asia. However, within exon 7 a relatively high level of variation was observed. The nucleotide evolutionary distances were quite high within the African compared to the South-east Asian divisions. Of particular note, was the high evolutionary distance value of both exon 6 and 7 within the Stenocephalemys division, and yet the amino acid sequence was identical between the two species within this division, suggesting strong purifying selection has been operating on *Zp3* within these lineages.

In Chapter 3 it was noted that in the New Guinean and Australasian Old Endemic murines, the more distantly related the species the more divergent the amino acid sequence, which would be expected if mutations accumulate in an approximately clock-like manner over time. A similar finding is evident amongst species investigated in the present chapter. The two species of *Aethomys* share an identical sequence which is not surprising as it was only relatively recently that these two species were recognized as being two distinct species (Chimimba *et al.* 1999). The two *Leopoldamys* species also share an identical sequence. If rapid evolution was occurring between closely related species then these observations could not be made.

Within the large South-east Asian genus of *Rattus*, species groups tend to share the same amino acid sequence, although some variation does occur. The pacific rat, *Rattus exulans*, shares Glu-337 and Thr-338 with *R. norvegicus* but not with the New Guinean and Australian occurring species. The *Rattus* species endemic to both New Guinea and Australia share the distinctive '*Rattus*' glutamic acid residues in positions 335 and 340. This suggests that they may be more closely related to other South-east Asian species such as *Bandicota* and *Bunomys* than they are to the New Guinean and Australasian Old endemic murines.

Five species of Australian occurring *Rattus* species (RattusF species group) all share a unique insertion within the putative combining-site for sperm. This insertion introduces a negatively charged, but hydrophilic, aspartic acid residue into the sequence but does not appear to change the overall charge or hydrophilicity of the region. *Rattus leucopus*, a species that is found in far north Queensland as well as in New Guinea, does not share this insertion, suggesting that it does not belong in the RattusF species group and probably evolved in New Guinea (Watts & Baverstock 1994b). The rate of divergence (d = 0.0036) is quite low within these species, lower than that for the Pseudomys group of Australasian Old Endemics (d = 0.0176). This may perhaps reflect their relatively recent adaptive radiation. Four of the species have an identical nucleotide sequence with each other, with *R. tunneyi* not sharing the common *Rattus* glutamic acid in position 340. Interestingly, fertile hybrids between several Australian *Rattus* species have been produced in the laboratory (Baverstock *et al.* 1983) and yet, the amino acid sequence of the exon 7 coding region is identical between species, suggesting that this region does not operate to prevent hybridization between species in this group of murines. Furthermore, hybrids were also produced between *R. leucopus* and *R. fuscipes* in the laboratory (Baverstock *et al.* 1983) although only one of these species has the single amino acid insertion. The successful hybridization between these two *Rattus* species suggests, again, that amino acid changes in this region have not resulted in a reproductive barrier based on sperm-ZP binding between these species.

Within the African, Eurasian and South-east Asian divisions, greater divergence of amino acid sequence occurred than observed within the New Guinean and Australasian Old Endemic divisions as would be expected given the higher levels of nucleotide evolutionary distances. Within the African division of Arvicanthis, the *P* value for exon 7 was 11.76 which did not take into account the two residue deletion in the *Lemniscomys griselda* sequence occurring at codon positions 336 and 337. Most species have the same two amino acids in this position as the mouse (FQ), including the majority of the *Rattus* species. Residue position 337 shows some variation, although it appears to be lineage specific. However, *Rattus norvegicus, R. exulans* and *Maxomys hellwaldii* all share a glutamic acid (E) in this position. Within the Maxomys divisions, the *P* value was also quite high (13.89) and this did not take into account the single amino acid insertion (S). Although *M. hellwaldii* is considered part of the *Maxomys* genus, the amino acid sequence within the exon 7 coding region is more similar to that of *Rattus norvegicus* than it is to *M. bartelsii.*

Of the African, Eurasian and South-east Asian divisions, the amino acid sequence of species within the Rattus division show the most divergence from the mouse. In particular, substitutions between *Rattus* species and the mouse tend to change the charge of the amino acid. This may have repercussions for protein folding due to differences in glycosylation although the hydropathy profile within the combining-site for sperm remains similar to that of the mouse.

Conclusion

The results presented in this chapter suggest that there is more variation within the exon 7 coding region within this diverse range of African, Eurasian and South-east Asian murines than there is seen in the New Guinean and Australasian Old Endemic rodents (see Chapter 3). There is little support for the hypothesis that closely related species show a high level of divergence within the exon 7 coding region, although within some divisions such as Stenocephalemys and Maxomys, the high divergence rates appear to provide limited support. Certainly, the greater evolutionary distance between the species, the

more the divergent the primary and tertiary structure of the exon 7 coding region is seen, as might be expected unless positive selection during speciation is leading to rapid evolution of the amino acid sequence. The question of whether positive selection has occurred within the exon 7 coding region of *Zp3* among African and South-east Asian murine species will be investigated in Chapter 5.

Chapter 5

Detection of positive selection occurring within the exon 7 coding region of *Zp*3

Image on reverse: Australian Old Endemic rodent, *Pseudomys fumeus*.
Image modified from the private collection of Assoc. Prof. Bill Breed

Chapter 5

Detection of positive selection occurring within the exon 7 coding region of Zp3

5.1 Introduction

If a region of a gene has been evolving rapidly, with higher rates of amino acid replacing nucleotide mutations over silent mutations, then the gene, or region of the gene, may have evolved under positive selection. In this situation, amino acid changes that provide a fitness advantage to an organism are fixed at a higher rate than neutral changes, and evidence of this adaptive evolution can be detected by comparing the gene's nucleotide sequences from different organisms.

A simple method of detecting adaptive evolution is to calculate the ratio of nonsynonymous (amino acid replacing) to synonymous (silent) substitutions between two sequences. A ratio greater than 1 suggests that positive selection has occurred and less than 1 implies that there has been purifying selection. However, this method averages the amino acid changes across the entire length of the gene, or the sequence, being investigated and does not take into account situations where the rate of evolution across a gene is variable (Graur & Li 2000). Certain regions of a gene may have evolved at a higher rate than others, and because most genes, or their protein products, are under functional and structural constraints, the signal of positive selection may not be detected due to predominant purifying selection (Nielsen & Yang 1998).

In order to detect the signal of positive selection occurring at only a few codons of a gene, a statistical based model-testing approach has been devised, using maximum likelihood and Bayesian statistics (Phylogenetical Analysis Using Maximum Likelihood: Yang 1997, see Chapter 2.10.2). This method, the codon substitution model, is applied to the data obtained in Chapters 3 and 4.

Reproductive proteins involved in sperm-egg interaction have been suggested as having evolved faster than other proteins, as well as having been of interest due to their potential role in speciation events. Several studies have used the codon substitution models approach to detect evidence of positive selection in the *Zp3* gene. For instance, Swanson *et al.* (2001) used the model to detect positive selection occurring in the mammalian ZP3 glycoprotein. They found a reasonably strong signal of positive selection within the exon 7 coding region of *Zp3* by applying the maximum likelihood method to the *Zp3* sequences from a diverse range of mammalian species that included such species as the mouse, rat, dog, cat, monkeys and humans. However, the codon substitution model has now become more rigorous and conservative since the publication of that paper, and, after reanalysis of the data using the updated model, Berlin and Smith (2005) found no convincing evidence of positive selection.

In 2003, Jansa *et al.* applied the codon substitution model to thirteen species in the genus *Mus*, and found that there was evidence of positive selection acting on a number of codons within exon 7. They applied the less rigorous version of the codon substitution model and included *Rattus norvegicus* as an outgroup. Turner and Hoekstra (2006) subsequently applied the newer version of the model, excluding the *R. norvegicus* sequence from the data, and found no evidence of positive selection. These latter authors also applied the newer model to the exon 6 and exon 7 coding sequences of *Zp*3 from the cricetid genus *Peromyscus* and obtained strong evidence of positive selection occurring within exon 7 between species.

Thus it would appear that within the *Peromyscus* genus positive selection has occurred within exon 7. One of the limitations of this approach is that it is possible that in distantly related species, the evolutionary distances are so great that the signal of positive selection is lost possibly due to saturation of the evolution of specific sites under positive selection (Anisimova *et al*. 2001). It is equally possible that species from the same genus, such as *Mus*, are so closely related that the signal cannot be detected due to the small evolutionary distance between species. It is also possible that different

genera, with different evolutionary pressures, evolve at different rates. The aim of the present Chapter is to investigate whether positive selection has occurred in lineages of a broad, but closely related, group of murine rodents, using sequence data obtained in both Chapters 3 and 4.

5.2 Materials and methods

In the present Chapter the method of detecting positive selection, by calculating the nonsynonymous and synonymous substitutions rates (see chapter 2.10.1), has been applied as well as the more statistically robust maximum likelihood method using codon substitution models (see chapter 2.10.2) to the exon 6 and 7 coding sequences from selected species of murine rodents. The first method was applied to each division, using all species (n = 96 species, excluding those species where sequences were available on GenBank) whereas only selected species were used for the codon substitution model due to the need for *a priori* phylogenetic trees. The phylogenetic trees proposed by Ford (2006) (Chapter 1, Fig. 1.13), Watts and Baverstock (1994, 1995) (Figs. 5.1, 5.2 and 5.3) and Steppan *et al.* (2005) (Figs. 5.4, 5.5 and 5.6) have been used.

The exon 6/intron 6 boundary divided codon 21 of the coding sequence between exon 6 and exon 7. For the purposes of this analysis, the first nucleotide of exon 7 was added to the exon 6 coding region in order to keep the sequence in the correct frame. Therefore, exon 6 contained 63 nucleotides and exon 7 108 nucleotides.

5.3 Results

5.3.1 Rates of nonsynonymous and synonymous substitution

5.3.1.1 New Guinean and Australasian old endemic murine species

Murine species endemic to New Guinea and Australia used in this section have been classified into the six divisions (Musser & Carleton 2005). All species investigated in Chapter 3 were used. These six divisions were Lorentzimys (1 species), Pogonomys (11 species), Hydromys (3 species), Xeromys (3 species), Uromys (11 species) and Pseudomys (39 species). Estimations of nonsynonymous substitutions (d_N) and synonymous substitutions (d_S) were calculated using the Yang & Nielsen (2000) approximation method and conducted using the yn00 program of the PAML (Phylogenetic Analysis by Maximum Likelihood) software package (version 3.15; Yang 1997). Mean estimations were calculated over all species, within and between divisions.

5.3.1.1.1 Exon 6 of Zp3

Table 5.1. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) within and between each of the New Guinean and Australasian Old Endemic murine divisions for the exon 6 coding region of *Zp3*.

Overall mean d_N = 0.006, d_S = 0.041

		Within	***Between***					
			Lorentzimys	**Pogonomys**	**Hydromys**	**Xeromys**	**Uromys**	**Pseudomys**
Lorentzimys	d_N							
	d_S							
Pogonomys	d_N	**0.004**	**0.002**					
	d_S	0.042	0.022					
Hydromys	d_N	**0.000**	**0.000**	**0.002**				
	d_S	0.076	0.038	0.057				
Xeromys	d_N	**0.000**	**0.000**	**0.002**	**0.000**			
	d_S	0.038	0.038	0.053	0.051			
Uromys	d_N	**0.012**	**0.012**	**0.014**	**0.012**	**0.012**		
	d_S	0.034	0.016	0.037	0.049	0.044		
Pseudomys	d_N	**0.002**	**0.001**	**0.003**	**0.001**	**0.001**	**0.013**	
	d_S	0.041	0.023	0.043	0.053	0.046	0.036	

Over all six divisions, the mean d_S for exon 6 was 0.041 and d_N was 0.006 (Table 5.1). Within each division, d_S was invariably higher than d_N, suggesting that, overall, exon 6 has been evolving under selective constraints. Between divisions, the mean d_S was considerably higher than the mean d_N value, with the highest d_N value of 0.014 between the Pogonomys and the Uromys divisions (Table 5.1).

5.3.1.1.2 Exon 7 of Zp3

Table 5.2. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) within and between each of the New Guinean and Australasian Old Endemic murine divisions for the exon 7 coding region of *Zp3*.

Overall mean d_N = 0.031, d_S = 0.037

		Within	***Between***					
			Lorentzimys	**Pogonomys**	**Hydromys**	**Xeromys**	**Uromys**	**Pseudomys**
Lorentzimys	d_N							
	d_S							
Pogonomys	d_N	**0.029**	**0.037**					
	d_S	0.049	0.062					
Hydromys	d_N	**0.000**	**0.055**	**0.042**				
	d_S	0.041	0.055	0.050				
Xeromys	d_N	**0.000**	**0.055**	**0.042**	**0.000**			
	d_S	0.000	0.033	0.028	0.020			
Uromys	d_N	**0.025**	**0.065**	**0.051**	**0.014**	**0.014**		
	d_S	0.043	0.037	0.055	0.043	0.025		
Pseudomys	d_N	**0.017**	**0.060**	**0.058**	**0.016**	**0.015**	**0.029**	
	d_S	0.030	0.049	0.041	0.036	0.015	0.041	

Over all divisions, the mean pairwise estimate of d_N for exon 7 was 0.031 (Table 5.2). The mean pairwise estimate of d_S was 0.037, a rate similar to that observed for the coding region of exon 6. Within all divisions, mean pairwise d_N estimates were lower than d_S estimates, as expected for a sequence that is generally under purifying selection (Table 5.2). However, comparisons between divisions showed a number of instances where d_N was higher than d_S (for example, Xeromys versus Lorentzimys/Pogonomys; Uromys versus Lorentzimys; Pseudomys versus Lorentzimys/Pogonomys). This suggests positive selection has been acting on the exon 7 region during the evolution of the gene in these lineages. This is particularly evident when mean pairwise comparisons of d_N are made between the New Guinean divisions of Lorentzimys and Pogonomys and the other four divisions.

5.3.1.2 South-east Asian and African murine species

For the purposes of this analysis, all species investigated in chapter 4, including *Hylomyscus alleni*, were used. In addition, *Mus musculus* and *Rattus norvegicus* were included since both these species are thought to have originated in southern or South-east Asia. New Guinean and Australian occurring *Rattus* species are included in this section as they are closer in evolutionary distance to the other South-east Asian species than to the New Guinean and Australasian Old Endemic murines.

5.3.1.2.1 Exon 6 of Zp3

Table 5.3. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) within and between each of the African, Eurasian and South-east Asian murine divisions for exon 6 coding region of *Zp3*.

Overall mean d_N = 0.013, d_S = 0.124

			Between								
		Within	Mus	RattusN	Dasymys	Aethomys	Arvican	Steno	Apod	Dacnomys	Maxomys
Mus	d_N	n/a									
	d_S	n/a									
RattusN	d_N	n/a	0.000								
	d_S	n/a	0.062								
Dasymys	d_N	n/a	0.044	0.044							
	d_S	n/a	0.116	0.190							
Aethomys	d_N	0.039	0.023	0.022	0.023						
	d_S	0.044	0.022	0.090	0.145						
Arvicanthis	d_N	0.033	0.022	0.022	0.023	0.052					
	d_S	0.000	0.000	0.063	0.118	0.022					
Stenoceph	d_N	0.023	0.000	0.000	0.000	0.030	0.034				
	d_S	0.546	0.389	0.406	0.609	0.432	0.399				
Apodemus	d_N	n/a	0.000	0.000	0.000	0.029	0.033	0.000			
	d_S	n/a	0.291	0.411	0.458	0.342	0.299	0.394			
Dacnomys	d_N	0.011	0.007	0.007	0.007	0.037	0.041	0.007	0.007		
	d_S	0.000	0.062	0.135	0.189	0.090	0.063	0.434	0.318		
Maxomys	d_N	0.000	0.000	0.000	0.000	0.029	0.033	0.000	0.000	0.007	
	d_S	0.129	0.096	0.063	0.230	0.125	0.097	0.353	0.410	0.097	
Rattus	d_N	0.003	0.002	0.002	0.002	0.031	0.035	0.002	0.001	0.009	0.002
	d_S	0.017	0.062	0.130	0.189	0.090	0.063	0.420	0.320	0.019	0.094

The mean pairwise estimation overall of d_N (0.013) was considerably lower than the mean d_S (0.124). Within divisions, there was a variation of d_N and d_S. Of the six divisions with more than one species, two had higher d_N than d_S values although d_N was low (<0.033). Within two divisions (Stenocephalemys and Maxomys) the d_S value was considerably higher than d_N. Within the Stenocephalemys division, d_S was 0.546 compared to d_N of 0.023. Within the Maxomys division, d_S was 0.129 while d_N was 0.000.

Between divisions, d_S was higher than d_N in most pairwise comparisons. The exceptions were Mus versus Aethomys (d_N 0.023, d_S 0.022) and Arvicanthis versus Aethomys (d_N 0.052, d_S 0.022) and involved low values. Pairwise comparisons involving Stenocephalemys and Apodemus produced high d_S values, ranging from 0.291 (Apodemus versus Mus: d_S was 0.000) to 0.609 (Stenocephalemys versus Dasymys: d_S was 0.000). A closer inspection of the two species within the Stenocephalemys (data not

shown) showed that the d_S values were high for both species although *Mastomys natalensis* showed the higher of the two, ranging from 0.3288 (versus Apodemus) to 0.7439 (versus Dasymys).

5.3.1.2.1 Exon 7 of Zp3

Table 5.4. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) within and between each of the African, Eurasian and South-east Asian murine divisions for exon 7 coding region of *Zp3*.

Overall mean d_N = 0.068, d_S = 0.061

			Between								
		Within	Mus	RattusN	Dasymys	Aethomys	Arvicant	Steno	Apod	Dacnomys	Maxomys
Mus	d_N	n/a									
	d_S	n/a									
RattusN	d_N	n/a	0.087								
	d_S	n/a	0.088								
Dasymys	d_N	n/a	0.042	0.087							
	d_S	n/a	0.040	0.040							
Aethomys	d_N	0.038	0.028	0.092	0.032						
	d_S	0.055	0.082	0.083	0.040						
Arvicanthis	d_N	0.058	0.088	0120	0.057	0.072					
	d_S	0.079	0.081	0.081	0.039	0.068					
Stenoceph	d_N	0.056	0.056	0.110	0.076	0.070	0.092				
	d_S	0.239	0.205	0.213	0.143	0.188	0.166				
Apodemus	d_N	n/a	0.057	0.119	0.057	0.072	0.103	0.094			
	d_S	n/a	0.081	0.081	0.040	0.082	0.080	0.205			
Dacnomys	d_N	0.019	0.028	0.088	0.052	0.051	0.083	0.070	0.063		
	d_S	0.025	0.096	0.013	0.053	0.096	0.094	0.220	0.095		
Maxomys	d_N	0.059	0.058	0.058	0.058	0.063	0.098	0.102	0.097	0.049	
	d_S	0.037	0.102	0.019	0.060	0.102	0.100	0.173	0.101	0.032	
Rattus	d_N	0.022	0.086	0.055	0.073	0.087	0.114	0.127	0.123	0.065	0.056
	d_S	0.011	0.073	0.005	0.031	0.073	0.071	0.190	0.072	0.018	0.025

The mean overall estimate of d_N for exon 7 was 0.0676, similar to the d_S value of 0.0615. Within the divisions, d_S was higher than d_N in four out of the six divisions with more than one species. Of the two divisions where d_N was higher than d_S, both involved South-east Asian divisions of Maxomys and Rattus, but both values were less than 0.06. Within the Stenocephalemys division, the d_S value of 0.239 was again considerably higher than the d_N value (0.056), although it was not as high as for that of exon 6.

Between divisions, d_N was lower than d_S in 23 out of the 45 pairwise comparisons. With the exception of RattusN (*Rattus norvegicus*) versus Stenocephalemys, all pairwise comparisons involving RattusN had

higher d_N than d_S values. In all pairwise comparisons between Stenocephalemys and the other divisions, the d_S value was again considerably higher than d_N, ranging from 0.143 (versus Dasymys) to 0.220 (versus Dacnomys). The d_S values were not as high as seen in exon 6 but in no pairwise comparison did they exceed the d_N value. The Apodemus division had higher mean d_N than d_S values in over half of the pairwise comparisons, with lower d_S values than was seen in exon 6.

Within the African divisions, d_N was higher than d_S in two out of the six divisions, suggesting that positive selection may only have been present in some of the lineages. Of the South-east Asian divisions, d_N was higher than d_S in all three divisions, suggesting that positive selection of *Zp3* may have occurred within these lineages.

5.3.2 Likelihood ratio tests (LRTs) of positive selection

To further assess the statistical evidence for positive selection, the maximum likelihood method of detecting positive selection using codon substitution models was applied to groups of species using *a priori* phylogenetic relationships. Likelihood ratio tests (LRTs) (see Chapter 2.10.2 for methods) were conducted on the following data sets, based on the following *a priori* phylogenetic relationships:

Table 5.5: Australian Old Endemic murines only, using the phylogeny based on the nucleotide sequence data from the mitochondrial control region (Ford 2006).

Table 5.6: Lorentzimys and Pogonomys divisions with some representative species from the Hydromys, Xeromys, Uromys and Pseudomys divisions, using the phylogeny based on microcomplement fixation of albumin (MCFA) data (Watts & Baverstock 1994b).

Table 5.7: African and South-east Asian divisions (including one New Guinean *Rattus* species) using a phylogeny based on MCFA data (Watts & Baverstock 1994a).

Table 5.8: Representative species from all African, Eurasian, South-east Asian, New Guinean and Australasian murine divisions using a phylogeny based on MCFA data (Watts & Baverstock 1995).

Table 5.9: Species from Africa and South-east Asia divisions using a phylogeny based on nucleotide sequence from one mitochondrial and three nuclear genes (Steppan *et al.* 2005).

Table 5.10: Representative species from most divisions, including Australasian and New Guinean Old Endemic species, using the same Steppan *et al.* (2005) phylogeny as above.

Table 5.11: Representative species from most divisions, using the phylogeny based on nucleotide data from the *AP5* gene and containing more species than the combined data tree above (Steppan *et al.* 2005).

Table 5.5. Likelihood ratio tests of models of codon substitution using PAML (version 3.15). Australasian old endemic species only, using the phylogeny based on data derived from the mitochondrial control region (Ford 2006).

Model	ℓ	2Δℓ			df	*p* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-458.67						0.22	p_0 = 0.85, p_1 = 0.15 ω_0 = 0.076, ω_1 = 1	
M2a	-458.66	0.02			2	0.99	0.22	p_0 =0.87, p_2 = 0.13 ω_0 = 0.08, ω_1 = 1, ω_2 = 1.13	None
M7	-458.83						0.21	p = 0.18, q = 0.66	
M8	-458.66		0.34		2	0.84	0.22	p = 8.20, q = 99 p_0 = 0.87, p_s = 0.13, ω_s = 1.13	None
M8a	-458.67			0.02	1	0.88	0.22	p = 8.30, q = 99 p_0 = 0.85, p_s = 0.15, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *p* (χ2) = *p* values of LRT under χ^2 distribution, significant values (<0.05); ω = d_N/d_S; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

The LRTs using a phylogeny containing 40 species from mainly Australian species in the Xeromys, Uromys and Pseudomys divisions (Ford 2006), failed to show support for positive selection occurring within the exon 6 and 7 sequence. All three LRTs consistently failed to reject the null model in favour of the positive selection model ($p > 0.05$). The average ω value across the tree was either 0.21 or 0.22 and, in models allowing for a proportion of sites to have ω > 1, the parameter estimate for ω was 1.13, which is close to the ratio expected under a neutral model of evolution. In addition, only one residue (mZP3-334S) was identified as being under possible positive selection, but the Bayesian posterior probability (PP) was less than 70% (Table 5.5). This residue is a serine or threonine in most murine species studied, and therefore appears to be specific to only one lineage, that of the pebble-mound mice in the genus *Pseudomys* (*P. calabyi, P. chapmani, P. johnsoni, P. laborifex* and *P. patrius*).

Table 5.6. Likelihood ratio tests of models of codon substitution using PAML (version 3.15). Selected representative New Guinean and Australasian old endemic species using a combined phylogenetic tree of Watts & Baverstock (1995) derived from MCFA data.

Model	ℓ	2Δℓ			df	*p* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-487.39						0.22	p_0 = 0.79, p_1 = 0.21 ω_0 = 0.011, ω_1 = 1	
M2a	-482.69	9.4			2	0.009	0.49	p_0 = 0.88, p_2 = 0.12 ω_0 = 0.06, ω_1 = 1, ω_2 = 3.57	324S, 325H
M7	-487.45						0.21	p = 0.005, q = 0.018	
M8	-482.71		9.48		2	0.008	0.49	p = 6.43, q = 99 p_0 = 0.88, p_s = 0.12, ω_s = 3.57	324S*, 325H*, 341P
M8a	-487.39			9.36	1	0.002	0.22	p = 1.15, q = 99 p_0 = 0.79, p_s = 0.21, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *p* (χ2) = *p* values of LRT under χ^2 distribution, significant values (<0.05); ω = d_N/d_S; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

LRTs, using representative species from all six divisions of New Guinean and Australian Old Endemic murines (those species studied in Chapter 3) and the phylogeny based on MCFA data (Watts and Baverstock 1995), supported evidence of positive selection. In all three tests, the null model (M1a, M7 and M8a) of no positive selection, was rejected ($p < 0.01$) in favour of the alternative model, that is, the model of positive selection fits the data significantly better than the null model. Models M2a and M8 identified 12% of codons as being under positive selection pressure (ω_s =3.57). Of these seven codon sites, two amino acid sites, mZP3-324 and mZP3-325, gave a posterior probability of greater than 95% for all models, with a third site identified only under the M8 model (mZP3-341) with a PP > 95%. All three sites are within the exon 7 coding region, with the first two being close to, but not within, the sequence of residues identified by Wassarman and Litscher (1995) as the combining-site for sperm in the mouse. Position 324 contains a serine residue in the majority of species, but this residue was replaced with an arginine (R) in one species in the Pogonomys division, and with either an arginine (2 species) or an isoleucine (I) (2 species) within the Uromys division. This common serine (S) residue was

substituted with either an isoleucine (I) residue (2 species) or a threonine(T) (two *Mesembriomys* species) within the large Pseudomys division.

The second site identified as being under positive selection, position 325, changed from an aspartic acid (D) in the majority of species to either an asparagine (N), alanine (A) or glutamic acid (E) within the Pogonomys division. However, within other divisions, only one species had a substitution at this site (*Pseudomys fieldi*: not included in PAML analysis). Residue mZP3-341 contains a proline in the majority of Lorentzimys and Pogonomys division species, with substitutions to an alanine (A) or a leucine (L) occurring. Within the other four divisions, the majority of species including the New Guinean *Melomys* species, have a serine in this position, with three Australian *Melomys* having an alanine (A).

Each of the above three sites show parallel site changes only within the New Guinean Old Endemics, where the same amino acid has independently evolved in different lineages.

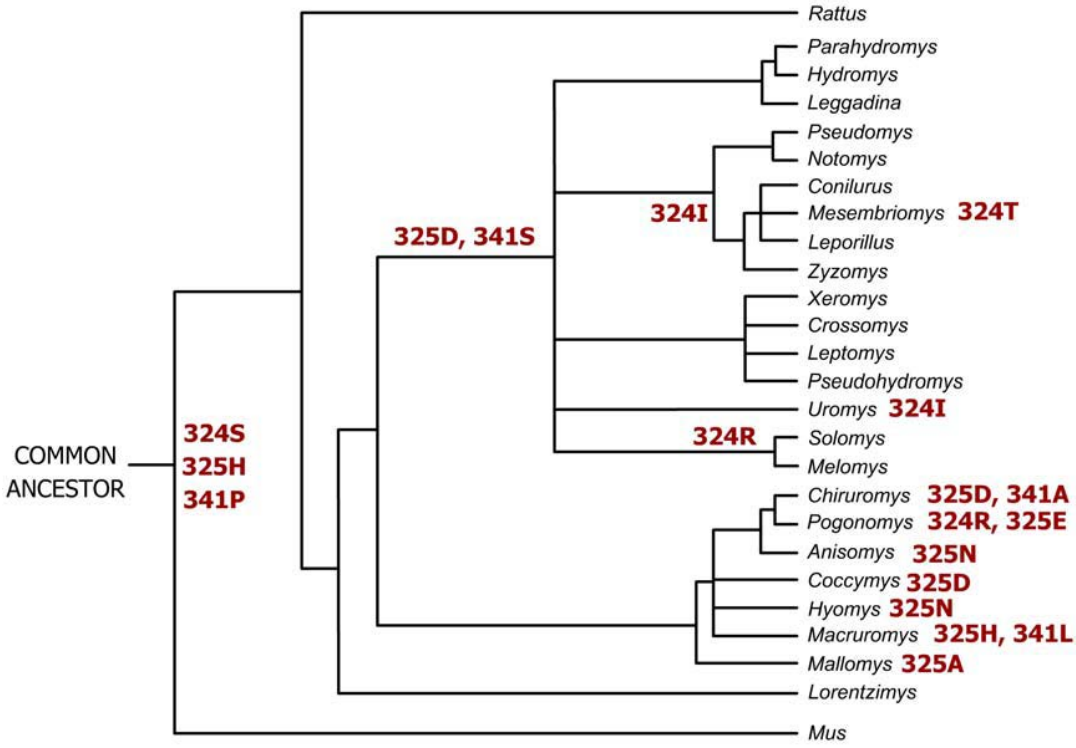


Fig. 5.1. Phylogeny based on microcomplement fixation of albumin data for representative New Guinean and Australian Old Endemic murine species (Watts & Baverstock 1995). The sites that were identified as being under positive selection (table 5.6) have been plotted against this phylogeny (in red).

Table 5.7. Likelihood ratio tests of models of codon substitution using PAML (version 3.15). African and South-east Asian species using the phylogeny based on microcomplement fixation of albumin data (Watts & Baverstock 1994a).

Model	ℓ	2Δℓ			df	*P* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-596.17						0.2601	p_0 = 0.7680, p_1 = 0.2320 ω_0 = 0.0366, ω_1 = 1	
M2a	-593.07	6.19			2	0.045	0.4591	$p_0 + p_1$ = 0.9058, p_2 = 0.0942 ω_0 = 0.0673, ω_1 = 1, ω_2 = 3.1720	335Q (0.92), 342H (0.93)
M7	-596.69						0.3179	p = 0.0149, q = 0.0296	
M8	-592.26		8.87		2	0.011	0.4339	p = 0.2927, q = 1.9088 p_0 = 0.8915, p_s = 0.1085, ω_s = 2.9371	335Q, 342H
M8a	-596.04			7.57	1	0.006	0.2497	p = 0.3381, q = 6.0883 p_0 = 0.7886, p_s = 0.2113, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *P* (χ2) = *P* values of LRT under χ^2 distribution, significant values (<0.05); ω = d_N/d_S; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

The phylogeny used in this LRT is that proposed by Watts and Baverstock (1994) and includes only those species investigated in Chapter 4 (African, Euroasian and South-east Asian species). All three LRTs showed good support for positive selection occurring, with the null model of no positive selection being rejected ($P < 0.05$) in all instances. Models M2a and M8 showed that ~ 10% of codons were under positive selection with ω of 3.17 and 2.93 respectively. However, only model M8 produced codons sites with posterior probabilities greater than 95%. These two sites (mZP3-335 and 342) were also identified in model M2a but with PP of 0.92 and 0.93 respectively.

Position 335 is a glutamine (Q) in the mouse sequence as well as in the sequence of Dasymys, Aethomys and Arvicanthis division species. Within the remaining African division of Stenocephalemys, *Mastomys natalensis* has a glutamine in this position, whereas *Hylomyscus alleni* has a tryptophan (W). *Apodemus chevrieri* has a leucine (L) in position 335. Of the South-east Asian divisions, the two *Leopoldamys* species have a glutamine, in common with the mouse, while the other member of this division (Dacnomys), *Niviventer fulvescens*, has an arginine (R). With the exception of *Maxomys*

bartelsii (with a glutamine) the remaining species, including all Rattus division species, have a glutamic acid (E). These results suggest that 335Q is the ancestral residue, and that lineage specific substitutions have occurred.

Position 342 contains an arginine (R) in the mouse, while *Rattus norvegicus* has an alanine (A). Other South-east Asian species have either a glycine (G) in the majority of species, an asparagine (N) in *Rattus exulans*, and a serine in two New Guinean *Rattus* species (*R. praetor* and *R. mordax*). Within the African divisions, *Dasymys incomtus* and *Micaelamys namaquensis* (Aethomys division) share a histidine (H) in this position, while the two *Aethomys* species have an arginine (R) in common with the mouse. All four species from the Arvicanthis and Stenocephalemys divisions have a tyrosine (Y) in position 342. *Apodemus chevrieri* has a proline (P) in this position.

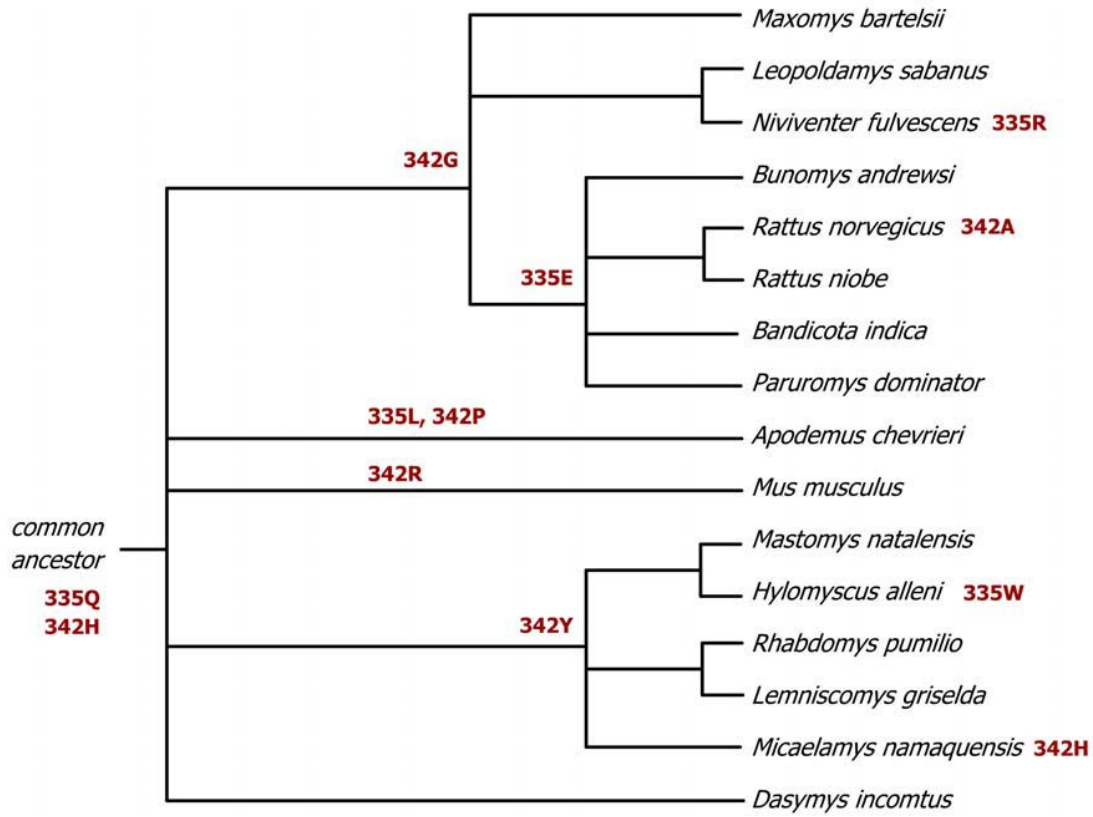


Fig. 5.2. Phylogeny proposed by Watts & Baverstock (1995), derived from microcomplement fixation of albumin from representative African, Asian and South-east Asian murine species (includes one New Guinean *Rattus* species). The sites that were identified as being under positive selection (table 5.7) have been plotted against this phylogeny and are in red.

Table 5.8 Likelihood ratio tests of models of codon substitution using PAML (version 3.15). African, Eurasian, South-east Asian, New Guinean and Australian murine species using the phylogeny based MCFA data (Watts & Baverstock 1995).

Model	ℓ	$2\Delta\ell$			df	P (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-893.06						0.2193	p_0 = 0.8312, p_1 = 0.1688 ω_0 = 0.0607, ω_1 = 1	
M2a	-883.19	19.75			2	<0.001	0.4600	$p_0 + p_1$ = 0.8997, p_2 = 0.1003 ω_0 = 0.0928, ω_1 = 1, ω_2 = 3.2587	311S, 325N*, 335Q*
M7	-896.62						0.2466	p = 0.1200, q = 0.3667	
M8	-883.10		27.04		2	<0.001	0.4576	p = 0.3710, q = 2.2292 p_0 = 0.8940, p_s = 0.1060, ω_s = 3.1440	311S*, 325N*, 335Q*, 342H
M8a	-892.80			19.40	1	<0.001	0.2173	p = 1.0484, q = 15.1292 p_0 = 0.8351, p_s = 0.1649, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; $2\Delta\ell$ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; P (χ2) = P values of LRT under χ^2 distribution, significant values (<0.05); $\omega = d_N/d_S$; ω for each branch calculated under each model; p = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

The phylogeny used in this analysis is based on the combined data phylogeny of Watts and Baverstock (1995) and includes 40 species from both Chapters 3 and 4. Strong support for positive selection was found with all three LRTs, allowing for the rejection of the null model of no positive selection, that is, the alternative model of positive selection (M2a and M8) fit the data statistically significantly better (P < 0.001) than the neutral models (M1a, M7 and M8a) of evolution. Both models M2a and M8 found 10% of codon sites to be under position selection, with three sites (M2a) and four sites (M8) having posterior probabilities (PP) of greater than 95%. Two of the sites (M2a: 335, M8: 335 and 342) were identified as being under positive selection using the data set and phylogeny set out in Table 5.7. Position 335, a glutamine (Q) in the mouse, and a glutamic acid (E) in Rattus division species, is a tryptophan (W) within all six divisions of New Guinean and Australasian old endemic murines. In position 342, an arginine (R) in the mouse and a glycine (G) in most *Rattus* species, is a proline (P) in species from all six divisions of New Guinean and Australasian Old Endemic murine species, with one exception being *Paramelomys rubex* (proline). These two sites do not appear to have been evolving under positive selective pressure

within the New Guinean and Australasian Old Endemic murines. However, when species from Africa and South-east Asia are included in the phylogeny, a signal of positive selection is detected.

The two other codon sites identified as being under positive selection are at positions 311 and 325. At position 311, the amino acid residue within the African and South-east Asian division is either a leucine (L) or a serine(S), and within the New Guinean and Australasian Old Endemic species either a serine (S) or tryptophan (W). All the *Pseudomys* and *Notomys* species, within the Pseudomys division, have a tryptophan in position 311, while most other species endemic to either New Guinea or Australasia have a serine, with the exception of *Lorentzimys nouhuysi* (leucine). In most instances, this substitution required the codon TTG (L) to change to TCG (S), a single substitution at the second codon base. However, the codon in this position in *Hylomyscus alleni* (of the Stenocephalemys division) is CTA (S), suggesting that two nucleotide substitutions, at least, have occurred. Parallel substitutions appeared to have occurred across different lineages.

The codon site in position 325 was also detected as being under positive selection when the phylogeny of Watts and Baverstock (1995) focusing only on New Guinean and Australasian Old Endemic species, was used (Table 5.6). Within the African groups of divisions, the amino acid residues at position 325 is either a histidine (H) (*Mus musculus*) and Stenocephalemys division species, an asparagine (N) (Dasymys and Micaelemys divisions) or an aspartic acid (D) (*Aethomys* species and the Arvicanthis division). Within the Eurasian and South-east Asian group of divisions, there is either a histidine (H) (Apodemus and Dacnomys divisions), or asparagine (N) (Maxomys and Rattus divisions).

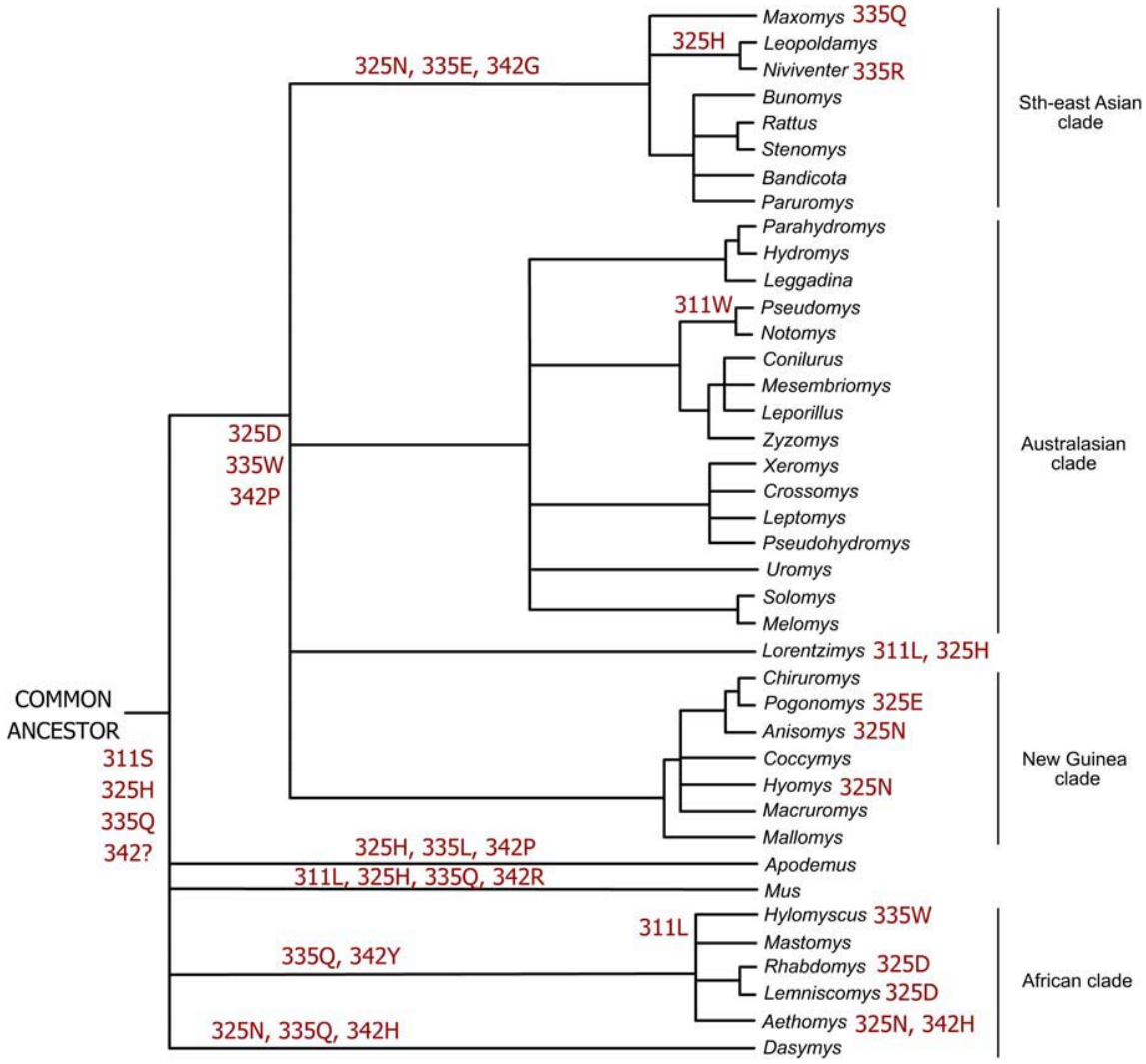


Fig. 5.3. Phylogeny based on microcomplement fixation of albumin data of Old World murine species (Watts & Baverstock 1995). The sites that were identified as being under positive selection (table 5.8) have been plotted against this phylogeny and are in red.

Fig.5.3 shows the sites identified as having evolved under positive selection plotted against the phylogeny used in the LRT. While three sites were identified in the LRT, when plotted against the phylogeny it is evident that only one codon site (311) varied across the Australasian Old Endemic lineage. However, the single substitution that occurred was specific only to the *Pseudomys/Notomys* lineage.

Table 5.9. Likelihood ratio tests of models of codon substitution using PAML (version 3.15). African, Eurasian and South-east Asian murine species using the phylogeny based on nucleotide sequence data from one mitochondrial and three nuclear genes (Steppan *et al.* 2005).

Model	ℓ	2Δℓ			df	*P* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-552.34						0.2738	p_0 = 0.7429, p_1 = 0.2571 ω_0 = 0.0225, ω_1 = 1	
M2a	-544.79	15.11			2	<0.001	0.5855	$p_0 + p_1$ = 0.9551, p_2 = 0.0449 ω_0 = 0.0192, ω_1 = 1, ω_2 = 7.1942	335Q, 342G*
M7	-552.46						0.2350	p = 0.0204, q = 0.0641	
M8	-544.62		15.67		2	<0.001	0.5530	p = 0.0693, q = 0.2088 p_0 = 0.9538, p_s = 0.0462, ω_s = 6.8292	335Q, 342G*
M8a	-552.24			15.24	1	<0.001	0.2613	p = 0.1200, q = 2.3175 p_0 = 0.7710, p_s = 0.2290, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *P* (χ2) = *P* values of LRT under χ^2 distribution, significant values (<0.05); $\omega = d_N/d_S$; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

The phylogeny used in this analysis is that proposed by Steppan *et al.* (2005) based on the combined maximum likelihood nucleotide data from one mitochondrial and three nuclear genes. For the purposes of this analysis, New Guinean and Australasian Old Endemic species were excluded. Steppan *et al.* (2005) suggested that, contrary to the Watts and Baverstock phylogeny, the African clades were separated by long intervening branches.

Each of the three LTRs produced strong support for positive selection, with the null model being rejected ($P < 0.001$) in all tests. Models M2a and M8 detected 4% of codon sites to be under position selection, with ω = 7.1 and 6.8 respectively, and two of these sites were identified as having posterior probabilities of greater than 95% (sites 335 and 342). These two sites were also identified by the analysis using the Watts and Baverstock data set (Table 5.8).

Position 335 appears to have as its ancestral amino acid a glutamine (Q). This residue is present in the Arvicanthis and Aethomys divisions, and in one out of two Stenocephalemys species. It is also present

in *Mus musculus* and some members of the South-east Asian divisions. However, this residue is a glutamic acid (E) in the Rattus division species.

Position 342 contains a glycine (G) in the majority of South-east Asian species, a tyrosine (Y) in the Arvicanthis and Stenocephalemys divisions, and a histidine (H) or arginine (R) in the two remaining African divisions. It is an arginine in Mus and a proline in Apodemus. Within the Rattus division there is an alanine (*Rattus norvegicus*), asparagine (*R. exulans*), serine (*R. mordax* and *R. praetor*) or a glycine (remainder of Rattus division species).

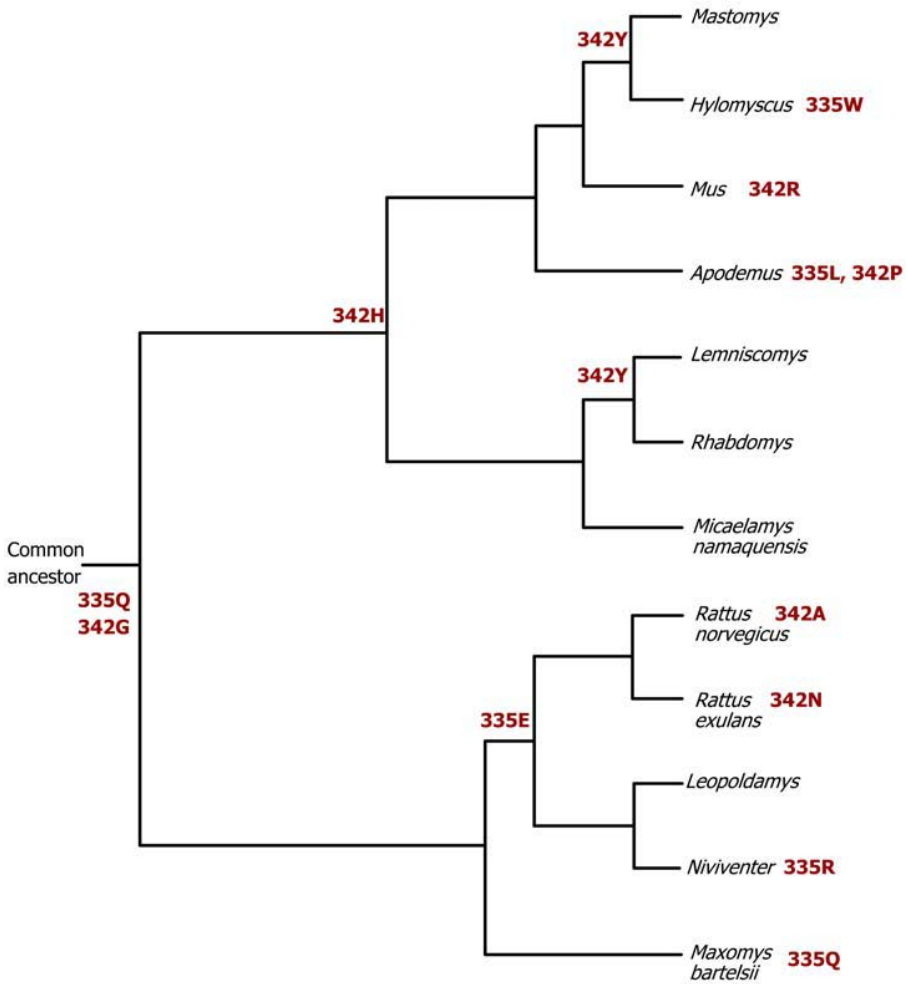


Fig. 5.4. Phylogeny based on the combined nucleotide sequence data of one mitochondrial and three nuclear genes from representative African, Eurasian and South-east Asian species (Steppan *et al.* 2005). The sites that were identified as being under positive selection (table 5.9) have been plotted against this phylogeny and are in red.

Table 5.10 Likelihood ratio tests of models of codon substitution using PAML (version 3.15). African, Eurasian, South-east Asian, New Guinean and Australian old endemic murine species using the phylogeny based on the combined nucleotide sequence data from one mitochondrial and three nuclear genes (Steppan *et al.* 2005).

Model	ℓ	2Δℓ			df	*P* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	*M7-M8*	*M8a-M8*					
M1a	-634.30						0.3122	p_0 =0.6957, p_1 = 0.3043 ω_0 = 0.0114, ω_1 = 1	
M2a	-623.56	21.48			2	<0.001	0.6870	$p_0 + p_1$ = 0.9480, p_2 = 0.0520 ω_0 = 0.0006, ω_1 = 1, ω_2 = 7.5088	335Q*, 342G*
M7	-634.02						0.3234	p = 0.0134, q = 0.0262	
M8	-623.41		21.21		2	<0.001	0.6256	p = 0.0266, q = 0.0754 p_0 = 0.9464, p_s = 0.0536, ω_s = 7.0130	335Q*, 342G*
M8a	-633.90			20.97	1	<0.001	0.2869	p = 0.0362, q = 0.6512 p_0 = 0.7435, p_s = 0.2565, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *P* (χ2) = *P* values of LRT under χ^2 distribution, significant values (<0.05); ω = d_N/d_S; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

The phylogeny used in this analysis is that proposed by Steppan *et al.* (2005), based on the combined maximum likelihood data (used in LRT Table 5.9), and contains representative species from all divisions in the present study. The LRTs produced similar results as the previous test (Table 5.9), although the 2Δℓ values were higher. The null model of no positive selection was rejected in all three tests (P < 0.001). A slightly higher number (5%) of codon sites were found to be under positive selection (ω = 7.5 (M2a) and 7.01(M8)), although the same two sites (335 and 342) were found to have PP of greater than 99%. When the codons 335 and 342 were plotted against the phylogeny used in the analysis, it is evident that although the substitutions occurred along the lineage that lead to the New Guinean and Australasian old endemic murines, no amino acid substitutions at these sites occurred during the large radiation of these species (with the single exception of *Paramelomys rubex*).

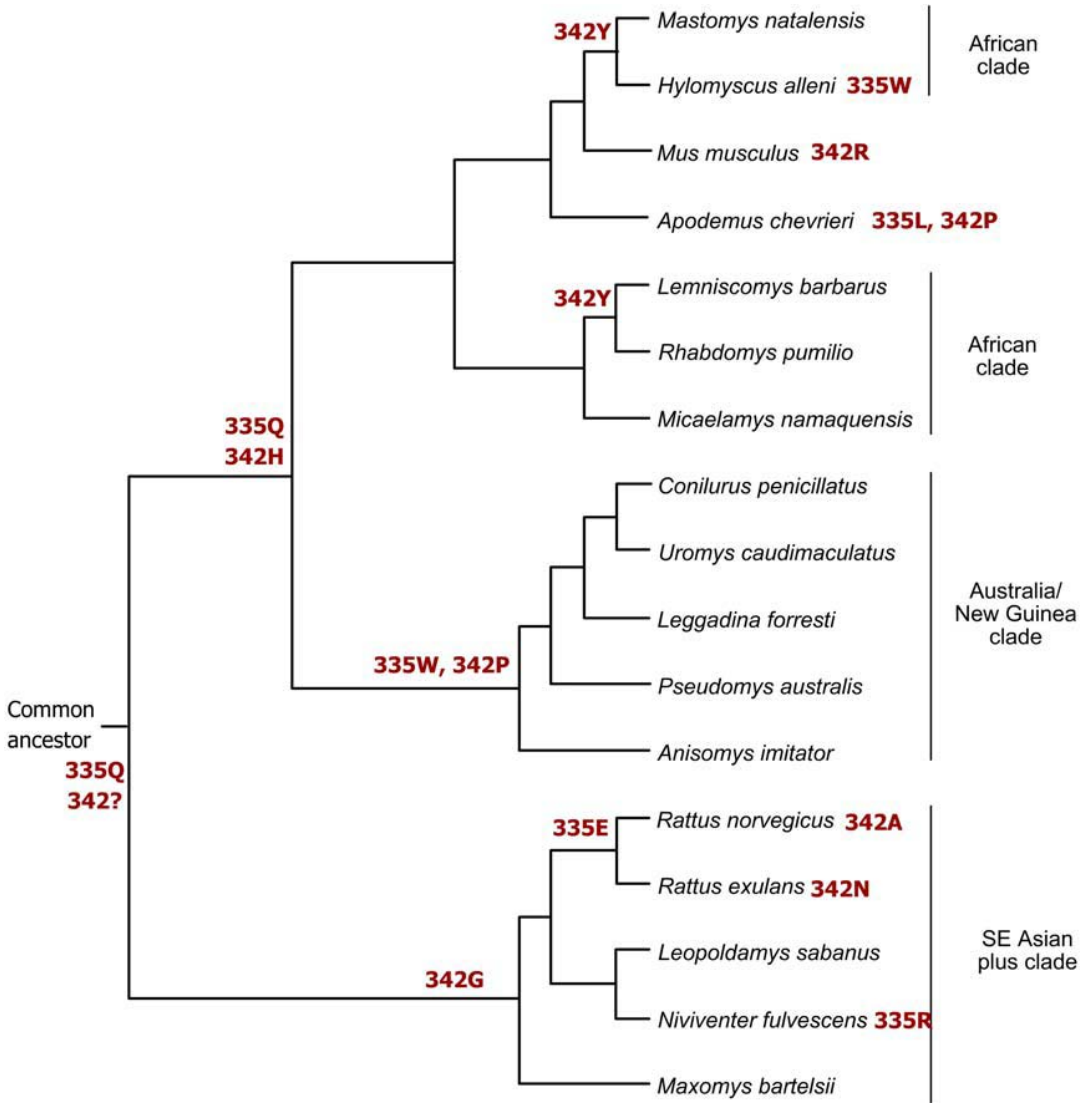


Fig. 5.5. Phylogeny based on the combined nucleotide sequence data of one mitochondrial and three nuclear genes from representative African, Eurasian, South-east Asian, New Guinean and Australasian Old Endemic murine species (Steppan *et al.* 2005). The sites that were identified as being under positive selection (table 5.10) have been plotted against this phylogeny and are in red.

Table 5.11 Likelihood ratio tests of models of codon substitution using PAML (version 3.15). African, Eurasian, South-east Asian, New Guinean and Australasian Old Endemic murine species using the phylogeny based on nucleotide sequence data from the AP5 gene (Steppan *et al.* 2005).

Model	ℓ	2Δℓ			df	*P* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-688.08						0.2478	p_0 =0.7804, p_1 = 0.2196 ω_0 = 0.0361, ω_1 = 1	
M2a	-680.78	14.6			2	<0.001	0.4647	$p_0 + p_1$ = 0.9499, p_2 = 0.0501 ω_0 = 0.0452, ω_1 = 1, ω_2 = 4.9692	335Q, 342H*
M7	-689.44						0.2528	p = 0.0726, q = 0.2145	
M8	-680.97		16.94		2	<0.001	0.4487	p = 0.1211, q = 0.4580 p_0 = 0.9475, p_s = 0.0524, ω_s = 4.7737	335Q*, 342H*
M8a	-688.09			14.8	1	<0.001	0.2473	p = 3.8584, q = 99 p_0 =0.7817, p_s = 0.2183, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *P* (χ2) = *P* values of LRT under χ^2 distribution, significant values (<0.05); ω = d_N/d_S; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

The phylogeny used in this analysis is that proposed by Steppan *et al.* (2005) based on nucleotide data from the AP5 gene and is included here due to the increased number of species represented (23 species as compared to 17 species in the combined data tree). There is very little difference between this analysis and that of the combined data analysis (Table 5.10), albeit the 2Δℓ values were lower. Strong support is again provided for positive selection occurring with significant differences between the fit of the null and alternative models ($P < 0.001$). The percentage of codons under positive selection was 5% with lower ω values (M2a = 4.9 and M8 = 4.7). The same two codons sites were found to be under positives selection with PP = < 95%.

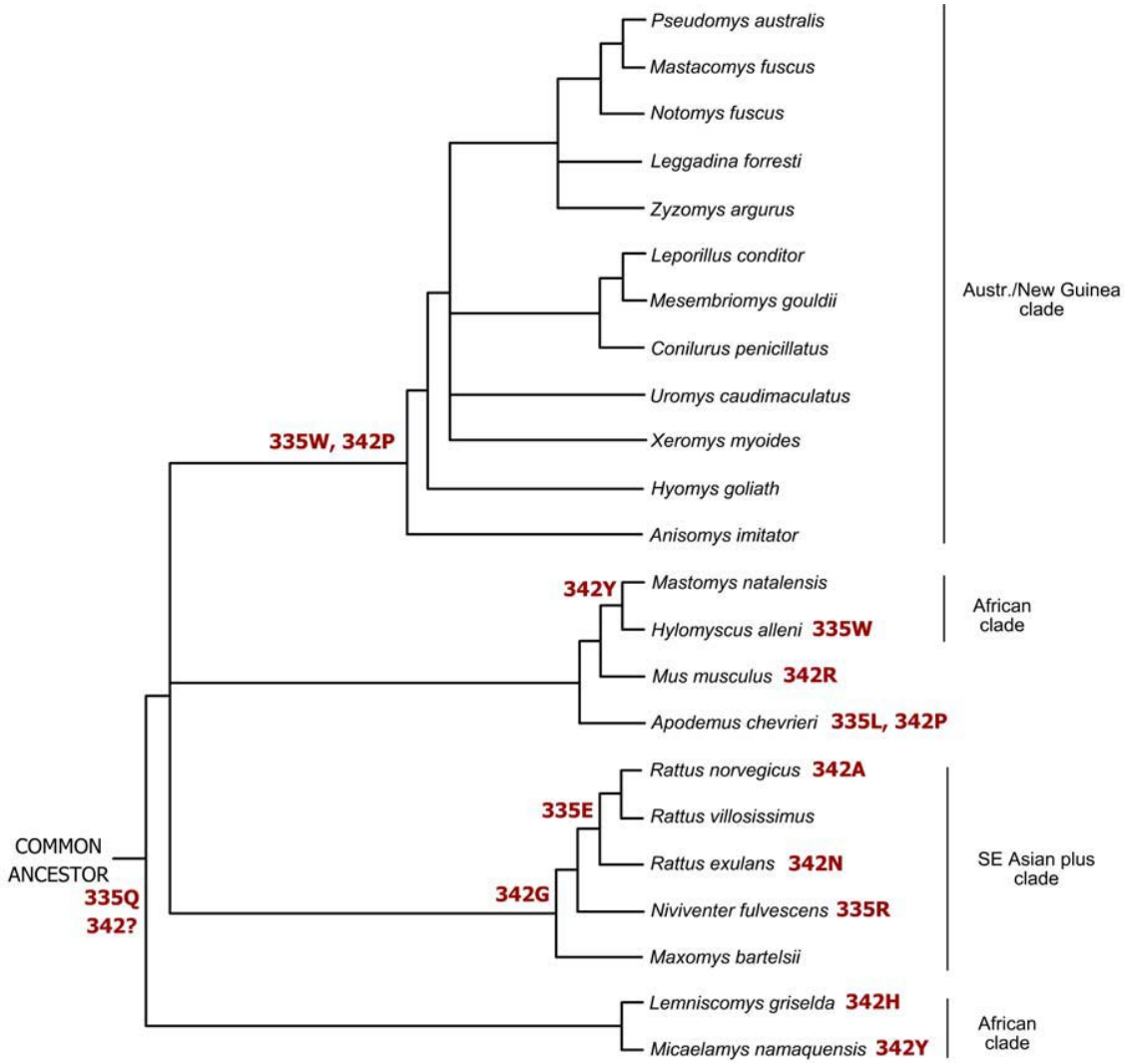


Fig. 5.6. Phylogeny based on nucleotide sequence data of the AP5 gene from representative African, Eurasian, South-east Asian, New Guinean and Australasian Old Endemic murine species (Steppan *et al.* 2005). The sites that were identified as being under positive selection (table 5.11) have been plotted against this phylogeny and are in red.

5.3.2.1 Summary of maximum likelihood analyses

Table 5.12. Summary of the results of the likelihood ratio tests for each codon substitution model.

Table No.	Phylogeny	Species	Positive selection	ω values	Sites with PP > 95%
5.5	Ford 2006	Australian Old Endemic only	No	0.22	None
5.6	Watts & Baverstock 1995	New Guinean and Australian Old Endemic only	Yes	3.57	324, 325, 341
5.7	Watts & Baverstock 1994a	African, Asian and SE-Asian	Yes	M2a = 3.17 M8 = 2.93	335 and 342
5.8	Watts & Baverstock 1995	All divisions	Yes		311, 325, 335 and 342
5.9	Steppan *et al.* 2005	African, Asian and SE-Asian combined ML data tree	Yes	M2a = 7.19 M8 = 6.82	335 and 342
5.10	Steppan *et al.* 2005	All divisions combined ML data tree	Yes	M2a = 7.5 M8 = 7.01	335 and 342
5.11	Steppan *et al.* 2005	All divisions AP5 data tree	Yes	M2a = 4.96 M8 = 4.77	335 and 342

5.4 Discussion

Conducting maximum likelihood analyses of codon substitution rates provides for statistical testing of models of selection. This method, using *a priori* phylogenies, plots the codon changes across hypothetical lineages. In this Chapter, different phylogenetic relationships, divided into two groups based on the authors of the respective phylogenies, were used in order to test the robustness of the methods: these were based on microcomplement fixation of albumin data (Watts & Baverstock 1995) and nucleotide sequence data from mitochondrial and nuclear genes (Steppan *et al.* 2005). In addition, for Australian Old Endemics a third phylogeny, based on nucleotide data, was used (Ford 2006). Using the codon substitution models, a strong signal of positive selection was detected occurring in New Guinean Old Endemic, African and South-east Asian murine species.

The results of the likelihood ratio tests (LRTs) showed a clear pattern of evolution. The LRT, using the Ford (2006) phylogeny for only Australian Old Endemic murines, failed to find statistical support for positive selection, with no codon sites selected as being under positive selection. However, when using the phylogeny of Watts and Baverstock (1995), the number of Australian species was reduced and grouped with New Guinean Old Endemic species, strong support for positive selection was suggested. Three sites (324, 325 and 341) were selected as having greater than 95% posterior probability of being under positive selection. Of these three sites only one (324) showed variation within the Australian Old Endemics. It is possible that this codon has been evolving under positive selection within the Australian Old Endemic murines but when only data from these species were used in the analysis the signal was not detected, perhaps due to the lack of power of the analyses when comparing these relatively closely related lineages.

When only African and South-east Asian species were tested using the Watts and Baverstock (1995) phylogeny, the null hypothesis of no positive selection was rejected in favour of the positive selection

model. Two sites (335 and 342) were selected as being under positive selection. However, when the New Guinean and Australasian old endemic species were included in this phylogeny, two more sites (311 and 325) were identified as being under positive selection. Therefore, using the Watts and Baverstock (1995) phylogeny, the LRTs suggested positive selection, with one set of residues evolving under positive selection among the New Guinean Old Endemics (324, 325 and 341) and another set evolving in African, Eurasian and South-east Asian lineages (335 and 342).

Three LRTs were conducted using the less species rich phylogenies of Steppan *et al.* (2005). Two phylogenies were used (based on different gene data sets) and all LRTs provided strong support for the positive selection model. Two codon sites were found to be under strong positive selection ($\omega = 4.77$ to 7.19). These two sites (335 and 342) were also found to under positive selection in the Watts and Baverstock (1995) phylogeny in the African, Asian and South-east Asian species. The codons 324, 325 and 341, evolving under positive selection in the New Guinean Old Endemic species were not found to be evolving under positive selection in the LRTs using the phylogenies of Steppan *et al.* (2005). This is perhaps not surprising as this group was only represented by two species.

The two codon sites of 335 and 342 are common to five of the LRTs. The substitutions that have occurred do not involve serine or threonine residues, with the exception of two species of *Rattus*. There appears to be no particular pattern to these substitutions. At position 335, the common residues are glutamine or glutamic acid, therefore introducing a change from an uncharged amino acid to a negatively charged residue. Other residues are either a tryptophan or leucine, both uncharged hydrophobic residues, or an arginine, a positively charged, highly hydrophilic residue. At position 342, there have been eight substitutions and only two involve a change in charge. However, it is possible that these changes either alter glycosylation sites by changing the overall charge of the primary sequence, or the hydrophobicity of the region of the protein. For this region to be effective in providing the ligand for the sperm proteins, it must be on the surface of the protein and hence, the zona matrix. Changing the

hydrophobicity of this region may thus alter the supramolecular structure of the glycoprotein, thereby changing the sperm-binding configuration.

The two residues at positions 335 and 342 were also found to be evolving under positive selection within the genus *Mus* (Jansa *et al.* 2003). These authors included the same two species from the Stenocephelamys division as has been used in the present study, and used *Rattus* as an outgroup. Residue 335 had a posterior probability of less than 95% but for residue 342 it was greater than 99%. A third site was also found to be under positive selection with a posterior probability of more than 95%. This site (337) was not found to be under positive selection in the LRTs conducted in this project but this does not mean that it is not evolving under positive selection, due to the limitations of this method. The maximum likelihood method of detecting positive selection when dealing with alignment gaps of the sort caused by the two codon deletion in the sequence from *Lemniscomys griselda* removes residue 337 from the analysis. This position in the African, Eurasian and South-east Asian species was variable and as there are a number of substitutions occurring along different lineages it is possible that the codon has evolved under positive selection. Therefore, the results of this project support the previous conclusions of Jansa *et al.* (2003), notwithstanding the fact that their results were challenged by Turner and Hoekstra (2006).

In their paper, Turner and Hoekstra (2006) used the same method to detect the occurrence of positive selection within the *Peromyscus* genus of North America. These species are members of the family Cricetidae and are therefore quite distantly related to species within the subfamily Murinae (present study). They are more closely related to the hamster than to the Old World mice and rats (Steppan *et al.* 2004). Within the *Peromyscus* genus, three codon sites (316, 343 and 345) were found to be evolving under positive selection, although two sites had only weak posterior probabilities (316 and 343). Due to indels, these three sites equated to mZP3-317, -344 and -346. The third codon, 346, had a posterior probability of greater than 95%, but as this site was not investigated in the present study, no parallels

can be drawn. However, in the present study, two other sites (317 and 344) were not found to be evolving under positive selection. In fact, in the old world murines, the residues at these sites remained unchanged.

Swanson *et al.* (2001) also used the maximum likelihood method of detecting positive selection within the *Zp3* sequence of a diverse range of mammals. They detected five residues (331, 333, 340, 341 and 345) under positive selection within the same region studied in the present study although only one (345) had a posterior probability greater than 90%. These residues are different from those found in the present study and are also different from those found by Turner and Hoekstra (2006) in their study. The d_N/d_S ratio was 1.7, a rate considerably lower than that found in the present analysis. Berlin and Smith (2005) repeated the LRT's of Swanson *et al.* (2001), using the same data sets with additional mammalian sequences, and applied improved algorithms. They also simulated data to test the robustness of the LRT's and found that the pattern of evolution in *Zp3* could generate an excess of false positives thus raising doubts about the conclusion of Swanson *et al.* (2001). In Chapter 6, a larger list of mammals will be used to test for positive selection but will focus on three mammalian orders: Rodentia, Carnivora and Primates .

In summary, in Old World murines, with the exception of the large group of Australasian Old Endemic species, there is strong support from three LRTs using several hypothetical phylogenies, that positive selection is occurring on a subset of codons. The codons found to have evolved under positive selection are all within the region identified by Wassarman and colleagues as being involved in sperm-ZP binding, although it appears that the sites under positive selection have not remained constant nor has it occurred consistently at the same codon sites throughout the adaptive radiation of the Old World murines.

Chapter 6

Comparison of amino acid sequences of the exon 7 coding region of *Zp3*, and detection of positive selection, between mammalian species

Image on reverse: Australian Old Endemic rodent, *Pseudomys nanus*.
Modified image from the private collection of Assoc. Prof. Bill Breed

Chapter 6

Comparison of amino acid sequences of the exon 7 coding region of Zp3, and detection of positive selection, between mammalian species.

6.1 Introduction

In the previous chapter, evidence for positive selection was detected within the region encoded by exon 7 of *Zp3* throughout the adaptive radiation of the murine rodents. However, the sites that have been evolving under positive selection are variable and are not under positive selection in all lineages. For example, evidence for positive selection was found within New Guinean Old Endemic murines but not within the Australasian murines. However, it was found that species of the same genus tended to share the same amino acid sequence within the exon 7 coding region. These findings are in contrast to those of Turner and Hoekstra (2006), who detected positive selection occurring within this region among species of the American rodent genus, *Peromyscus* and also found intraspecific variation in sequence.

As indicated in the previous chapter, Swanson *et al.* (2001) found evidence of positive selection within this region among a group of disparate mammalian species. However, the same data (with additional sequences) were re-analysed by Berlin and Smith (2005) who found no convincing evidence of positive selection, although this may have possibly been because the codon substitutions models had become more conservative since the Swanson *et al.* study.

If positive selection, albeit lineage variable, has been found in murine rodents, does it actually occur within other Orders and families of mammals? The data set used by Swanson *et al.* (2001) included species such as the mouse, rat, human, marmoset, bonnet monkey, dog, cat and pig. Berlin and Smith (2005) included in their reanalysis, additional sequences from the hamster, cow, rabbit, fox, ferret,

lemming and Brandt's vole. Although these fifteen sequences were used, there was no attempt to compare the amino acid sequences, nor assess the pattern of evolution in each taxonomic Order. Since 2001, more *Zp3* sequence have been determined and placed on GenBank, and it is now possible to analyse six species from the Order Primates, five species of the Order Carnivora and a collection of murines and cricetids from the Order Rodentia. This chapter focuses on these three orders by comparing the amino acid sequences, computing rates of nonsynonymous and synonymous substitutions rates and applying the codon substitution models to the data set.

6.2 Materials and Methods

6.2.1 DNA Sequences

DNA sequences for the species studied in this chapter were obtained from the National Centre for Biological Information data base (GenBank). Table 6.1 provides the full list, including common names and GenBank accession numbers. Nucleotide sequences were first aligned with Clustalw (Thompson *et al.* 1994), and then visually aligned to correspond with published amino acid sequences.

Table 6.1. List of mammalian species used in this chapter. Genus, species and common names were taken from the 3rd Edition of Mammal Species of the World (2005) and do not necessarily reflect the name used by authors on GenBank.

Order Suborder Family	Genus	Species	Common name	GenBank Accession No.
Rodentia				
Myomorpha				
Muridae	*Mus*	*musculus*	House mouse	NM_0011775
	Rattus	*norvegicus*	Brown rat	NM_053762
	Rattus	*rattus*	Roof rat	Y10823
	Rattus	*tanezumi*	Oriental house rat	AY338396
Cricetidae	*Lasiopodomys*	*brandtii*	Brandt's vole	AF304487
	Lagurus	*lagurus*	Steppe vole	AF515621
	Mesiocricetus	*auratus*	Golden hamster	M63629
	Onychomys	*torridus*	Southern grasshopper mouse	DQ668293 DQ668343
	Peromyscus	*polionotus*	Oldfield mouse	DQ668287 DQ668303
Primates				
Haplorrhini				
Hominidae	*Homo*	*sapiens*	human	NM_007155
	Pan	*troglodytes*	Chimpanzee	XM_001157669
Cercopithecidae	*Macaca*	*fascicularis*	Crab-eating macaque	AY222644
	Macaca	*mulatta*	Rhesus monkey	XM_001114760
	Macaca	*radiata*	Bonnet monkey	X82639
Cebidae	*Callithrix*	*jacchus*	Marmoset	S71825
Carnivora				
Caniformia				
Canidae	*Canis*	*lupus*	Dog	NM_001003224
	Vulpes	*vulpes*	Red fox	AY598032
	Mustela	*erminea*	Ermine	AY648050
	Mustela	*putorius*	Domestic ferret	AY702973
Feliformia				
Felidae	*Felis*	*catus*	Cat	NM_001009330

The aligned nucleotide sequence data are shown in Appendix 3. The predicted amino acid sequence is shown in Figure 6.1.

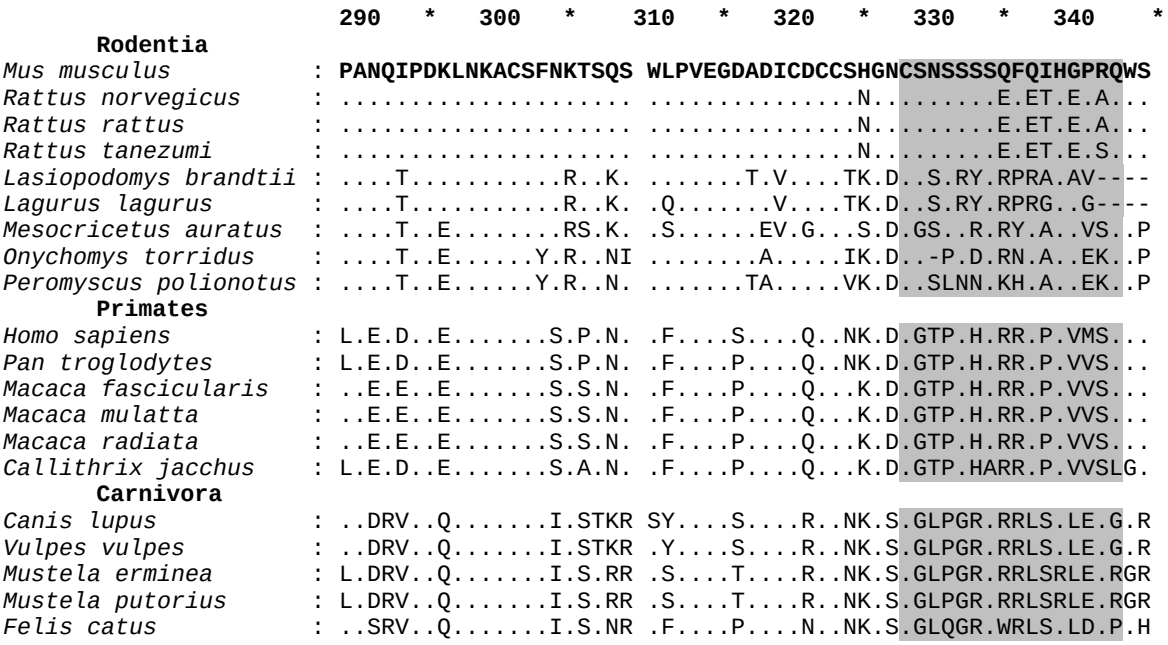


Fig. 6. 1. Alignment of predicted peptide sequence of exon 6 and exon 7 of *Zp3* from a range of species from three Orders: Rodentia, Primates and Carnivora. Genus and species names are according to the 3rd Edition of Mammal Species of the World 2005. Single dots (.) represent conservation of amino acid residues at any given position. The exon boundary is designated by the gap between residues 309 and 310. The area shaded in grey is the combining-site for sperm identified by Rosiere & Wassarman (1992).

6.2.2 Nonsynonymous and synonymous substitution rates

All but three sequences were used to calculate nonsynonymous substitution (d_N) and synonymous substitution (d_S) rates as detailed in chapter 2.10.1. This was due to the fact that the sequences from three species, all from the Order Rodentia (*Lasiopodomys brandtii, Lagurus lagurus* and *Onychomys torridus*), contained deletions within the exon 7 coding region of *Zp3*. As the codon substitution models implemented by the PAML software does not recognize alignment gaps, caused by either a deletion or an insertion, and removes them from the analysis, five informative codon sites would also have been removed across all species.

6.2.3 Analyses to test for evidence of positive selection

The method used to test for evidence of positive selection is that detailed in Chapter 2.10.2. For the same reasons discussed above, the sequences from three rodent species were omitted from the analysis using the codon substitution models.

6.2.4 Phylogenetic relationships

The phylogenetic trees used in the PAML analysis were derived from a number of different sources. The mammalian phylogeny was taken from Murphy *et al*. (2001). The phylogenies for the individual Orders were taken from the following papers:

- Rodentia: Steppan *et al.* (2004)
- Primates: Purvis (1995)
- Carnivora: Bininda-Emonds *et al.* (1999).

As the phylogeny of *Rattus* is poorly understood and no phylogeny exists for all of the extant *Rattus* species, only *Rattus norvegicus* was used in the analysis. With the removal of *Lasiopodomys brandtii, Lagurus lagurus* and *Onychomys torridus*, only four species represented the order Rodentia in the analysis.

6.3 Results

6.3.1 Amino acid sequence comparisons

The predicted amino acid sequence of the exon 6 and exon 7 coding regions of *Zp3* from twenty species were aligned. The sequence from *Mus musculus* was used as a reference to be consistent with Chapter 3 and 4. Sequences from nine species from the order Rodentia, six species from the order Primates and five species from the order Carnivora were included. Nucleotide sequences are available on GenBank from the cow, pig, rabbit and brushtail possum. These sequences were not included in the present analysis as they were single representatives from their Orders and therefore did not provide sufficient information for comparative purposes.

6.3.1.1 Rodentia

Three species of the genus *Rattus* were available (*R. norvegicus, R. rattus* and *R. tanezumi*). The sequence for *Rattus tanezumi* has been listed on GenBank as *R. rattus diardii*. This subspecies is now listed by Musser and Carleton (2005) as *R. tanezumi,* and this taxonomy has been used in this present study. The predicted amino acid sequence was identical between *R. norvegicus* and *R. rattus*. *R. tanezumi* shared all but one amino acid with *R. norvegicus* and *R. rattus*, having a serine in position 342. Together with *Mus musculus*, these four species are all from the Family Muridae.

There are five species from the Family Cricetidae. However, there is considerably more variation in the sequences within this family than was seen within Muridae. Within the exon 6 coding region there was a common shared threonine (T) in position 293, and a common shared arginine (R) in position 305. *Lasiopodomys brandtii* (listed in GenBank as *Microtus brandtii*), *Lagurus lagurus* and *Mesocricetus auratus* shared a lysine (K) in position 308, while the two remaining species shared an asparagine (N) in this position. *Mesocricetus auratus, Onychomys torridus* and *Peromyscus polionotus* all shared a glutamic acid (E) in position 296.

The coding region of exon 7 within Cricetidae showed more variability than was seen within Murinae. In particular, of the five serine residues conserved within the majority of murine species (Chapters 3 and 4), only Ser-334 was conserved in all five cricetids. The asparagine (N) residues in positions 327 and 330, found to be *N*-linked glycosylated in the mouse (Boja *et al.* 2003), and conserved in other murines species (Chapters 3 and 4), were not conserved in the cricetids, with all five species having an aspartic acid (D) in position 327 and four of the five species having a serine (S) in position 330. The fourth species, *Onychomys torridus* had a single amino acid deletion occurring in this position. All four cysteine (C) residues occurring within the exon 7 coding region were conserved. *Lasiopodomys brandtii* and *Lagurus lagurus* shared a four amino acid deletion from position 341.

6.3.1.2 Primates

While the predicted amino acid sequence of the five primate species showed considerable variation relative to the reference sequence of *Mus musculus*, within the Order there was marked conservation. Within exon 6, *Homo sapiens* and *Pan troglodytes* shared an identical sequence, as did the three *Macaca* species, albeit a different one. The amino acid sequence of *Callithrix jacchus* differed from that of *Homo sapiens* and *Pan troglodytes* by only one residue.

Within the exon 7 coding region, all cysteine residues found to be disulfide linked in the mouse (Boja *et al.* 2003) were conserved in all primates. The *N*-linked asparagines (N) at positions 327 and 330, common to the murines, have been substituted with an aspartic acid (D) in position 327 and a threonine (T) in position 330 in all six species. Of the five serine (S) residues at positions 329, 331 to 334, common to the murines, only Ser-332 has been conserved in all five primates and all but *Callithrix jacchus* have a serine in position 334. All primates have a glycine (G) in position 329, a proline (P) in position 331 and a histidine (H) in position 333. Between *Homo sapiens* and *Pan troglodytes* the amino acid sequence is identical in all but two positions: 316 and 341. The three species of *Macaca* all had an identical amino acid sequence.

6.3.1.3 Carnivora

The amino acid sequence from four species of the Family Canidae and one species from the Family Felidae were available. Again, the amino acid sequence was considerably different from that observed within the murines, but was reasonably conserved within the Order. All cysteine residues were conserved and the two *N*-linked asparagine (N) residues at positions 327 and 330 were substituted with a serine (S) and leucine (L) respectively. Of the five serine residues at positions 329, 331 to 334, only position 334 contained a serine in all five species. The amino acid sequence between *Canis lupus* and *Vulpes vulpes* was identical across both exon 6 and exon 7 coding regions with one exception: *C. lupus* has a serine in position 310, while *V. vulpes* has a tryptophan (W), a residue that is otherwise conserved across all three Orders (including those species investigated in Chapters 3 and 4). The two *Mustela* species have an identical sequence across both exon coding regions. The amino acid sequence of *Felis catus* shared some common residues with the other Carnivores although the sequence was quite variable within the exon 7 coding region.

6.3.2 Possible effects of amino acid change

6.3.2.1 Isoelectric points, potential glycosylation sites and mean hydropathic profiles

To assess the possible effects amino acid sequence change may have on a protein, relative serine/threonine composition (potential *O*-linked glycosylation sites); isoelectric point (change in overall charge) and hydropathy profiles (changes in hydrophobicity and hydrophilicity) were calculated using the methods discussed in Chapter 2.9.1. and 2.9.2. To be consistent with Chapters 3 and 4, only the carboxy terminal region of exon 7 (eighteen residues between 328 and 345) has been analysed. Table 6.2 provides the isoelectric point, relative percentages of serine/threonine residues and average hydropathy index per sequence for the three mammalian Orders.

Table 6.2. Isoelectric point, relative serine/threonine percentages and the average hydropathy index of the region of Zp3 from residues 328 to 345 from species belonging to the Orders Rodentia, Primates and Carnivora.

	Isoelectric point	Relative serine/threonine composition (%)		Average hydropathy index
		Serine	Threonine	
Rodentia				
Mus musculus	8.4	33	0	-1.09
Rattus norvegicus	4.3	33	6	-1.2
Rattus rattus	4.3	33	6	-1.2
Rattus tanezumi	4.3	39	6	-1.34
Lasiopoldomys brandti	10.9	29	0	-0.89
Lagurus lagurus	10.9	29	0	-1.54
Mesocricetus auratus	9.7	28	0	-1.07
Oncyhomys torridus	7.0	18	0	-1.86
Peromyscus polionotus	8.4	17	0	-1.61
Primates				
Homo sapiens	10.5	22	6	-1.23
Pan troglodytes	10.5	22	6	-1.11
Macaca fascicularis	10.5	22	6	-1.11
Macaca mulatta	10.5	22	6	-1.11
Macaca radiata	10.5	22	6	-1.11
Callithrix jacchus	10.5	17	6	-0.53
Carnivora				
Canis lupus	11.8	11	0	-1.14
Vulpes vulpes	11.8	11	0	-1.14
Mustela ermine	12.1	11	0	-1.42
Mustela putorius	12.1	11	0	-1.42
Felis catus	10.4	11	0	-1.04

The isoelectric point showed a range of values between the three Orders. The Rattus species showed the lowest value, with the Mustela species having the highest. The serine/threonine composition was also variable with the murine rodents having the highest percentage than the other species. The Carnivores had a much reduced value (11%) suggesting potential differences in glycosylation of this region

6.3.2.2 Hydropathy profiles for each Order

The hydropathy profile for species from the order Rodentia shows that there is reasonable conservation of hydropathy within the exon 6 coding region (292 to 310) (Fig.6.2). Between the residues 328 to 345 most sequences are hydrophilic. However, the amino acid sequence from the cricetid species have a few residues that make stretches of residues hydrophobic.

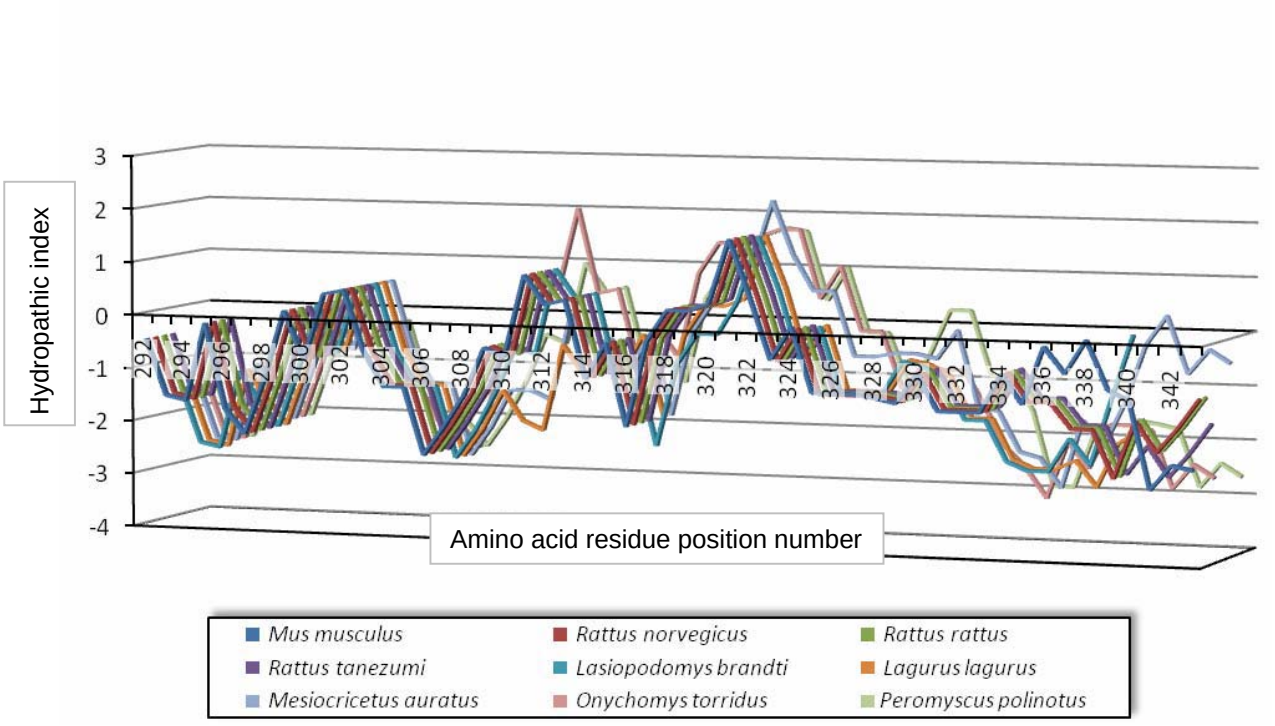


Fig. 6.2. 3D graphical representation of the hydropathy profile of the exon 6 and 7 coding region of *Zp3* from species from the order Rodentia. Below the horizontal axis (midline) are those residues that are hydrophilic and are therefore possibly on the surface of the protein. Above the horizontal axis (midline) are those residues that are hydrophobic and may be buried within the protein.

The hydropathy profile for species from the order Primates (Fig. 6.3) shows that there is good conservation of hydropathy across the exon 6 and 7 coding region. For the residues between positions 339 and 345, there is a small stretch of hydrophobicity for all six species, although the degree varies slightly, with *Callithrix jacchus* having the highest hydropathic index for this region.

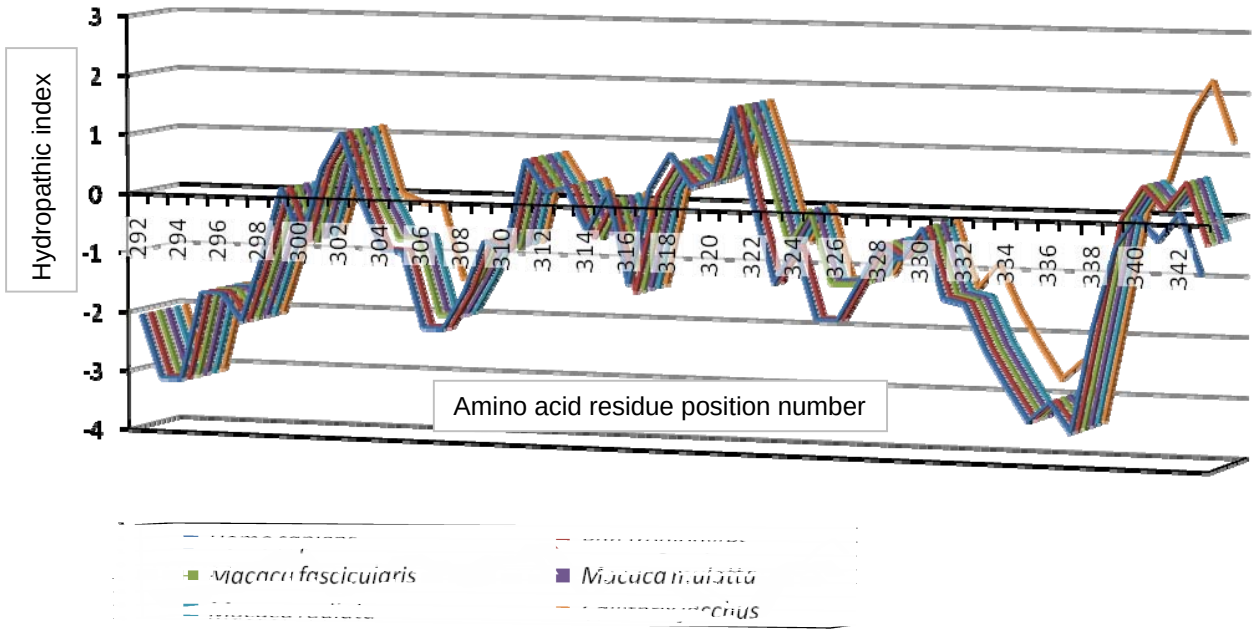


Fig. 6.3. 3D graphical representation of the hydropathy profile of the exon 6 and 7 coding region of *Zp3* from species belonging to the Order Primates. Below the horizontal axis (midline) are those residues that are hydrophilic and are therefore possibly on the surface of the protein. Above the horizontal axis (midline) are those residues that are hydrophobic and may be buried within the protein.

The hydropathy profile for species from the order Carnivora (Fig. 6.4) shows that there is a high degree of conservation of hydropathy across the exon 6 and 7 coding region. In contrast to the primates, there is a hydrophobic stretch of residues from 328 to 331. However, from 332 to 345, the residues are hydrophilic.

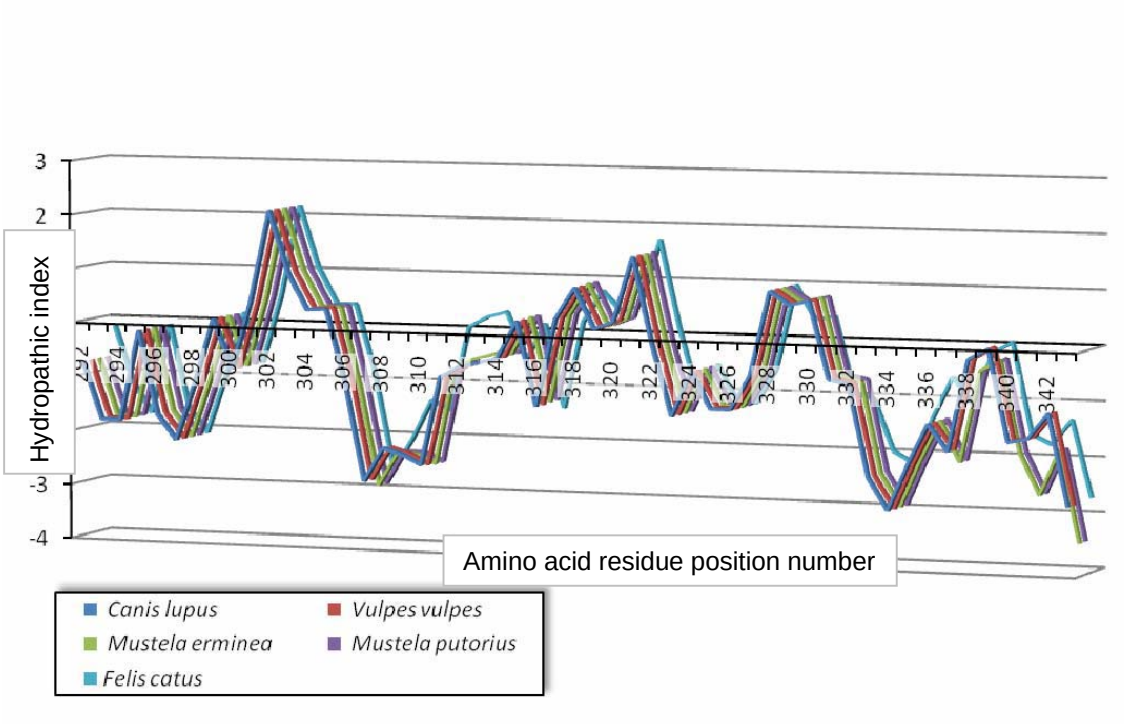


Fig. 6.4. 3D graphical representation of the hydropathy profile of the exon 6 and 7 coding region of *Zp3* from species belonging to the order Carnivora. Below the horizontal axis (midline) are those residues that are hydrophilic and are therefore possibly on the surface of the protein. Above the horizontal axis (midline) are those residues that are hydrophobic and may be buried within the protein.

To gain a clearer appreciation of the differences in hydropathy between the three Orders, Fig. 6.5 shows the hydropathy index from one species of each Order.

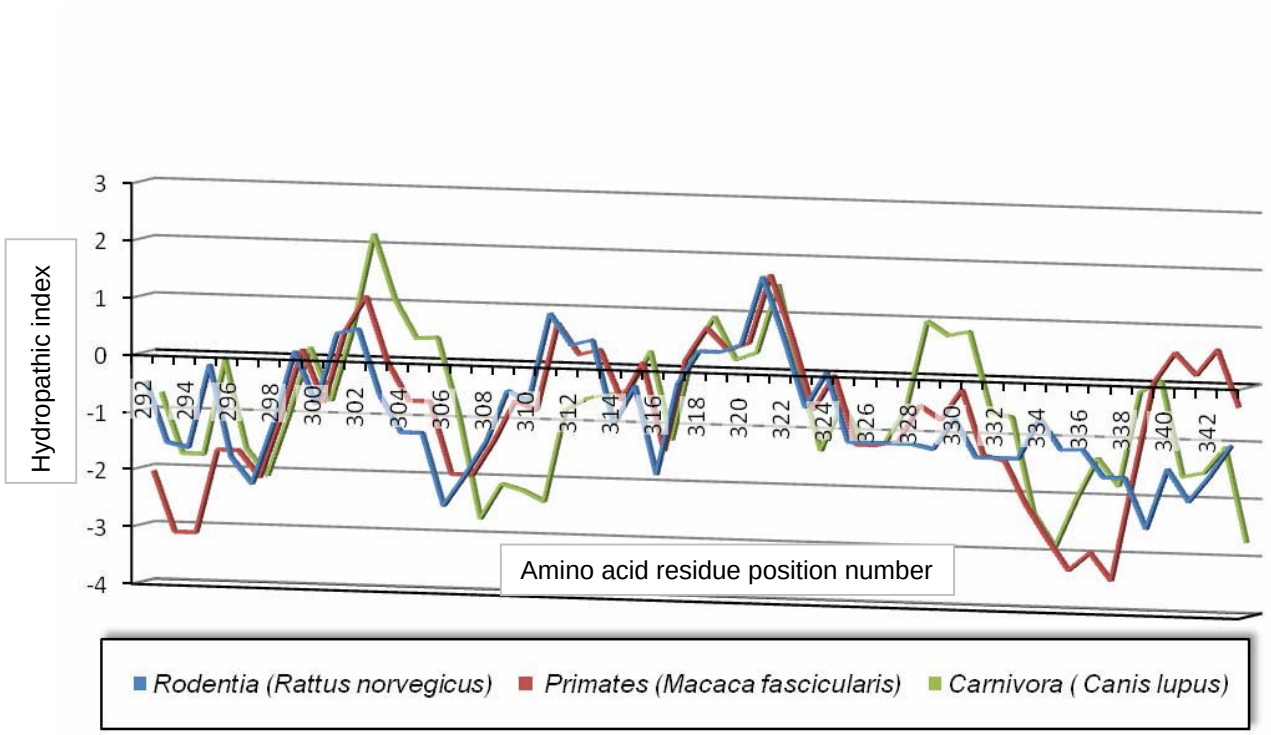


Fig. 6.5. 3D graphical representation of the hydropathy profile of the exon 6 and 7 coding region of *Zp3* from one species from each Mammalian order. Below the horizontal axis (midline) are those residues that are hydrophilic and are therefore possibly on the surface of the protein. Above the horizontal axis (midline) are those residues that are hydrophobic and may be buried within the protein.

Fig. 6.5 highlights the fact that there are regions of conserved hydropathy along the exon 6 and 7 coding regions of *Zp3*. The plot also clearly shows the differences in hydropathy in the region of residues from 328 to 345. Combined with varying isoelectric points and relative composition of serine/threonine residues between Orders, the differences in hydropathy within the region 328 to 343, suggests that this region of the glycoprotein may fold differently between Orders.

6.3.3 Rates of nonsynonymous and synonymous substitutions

Mean rates of nonsynonymous substitutions (d_N) and synonymous (d_S) substitutions were obtained within and between Orders.

6.3.3.1 Exon 6 of Zp3

Table 6.3. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 6 coding region of *Zp3* within and between each of the mammalian Orders.

Overall mean d_N = 0.1913, d_S = 0.6845

		Within	Between		
			Rodentia	Primates	Carnivora
Rodentia	d_N	0.0740			
	d_S	0.2296			
Primates	d_N	0.0422	0.2294		
	d_S	0.0869	0.4486		
Carnivora	d_N	0.0768	0.2625	0.2456	
	d_S	0.3733	1.2370	1.0451	

Within each order the d_S values were considerably higher than d_N values. The rate of nonsynonymous substitutions was below 0.1 for all three orders. The d_N values between orders were all within the range .2294 to 0.2625, while the d_S values were considerably higher, suggesting that the exon 6 coding region has evolved under purifying selection.

6.3.3.1.1 Rodentia

Table 6.4. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for exon the 6 coding region of *Zp3* between six species from the Order Rodentia.

Overall mean d_N = 0.0740, d_S = 0.2296

				Between			
		Mus musculus	*Rattus norvegicus*	*Rattus rattus*	*Rattus tanezumi*	*Mesocricetus auratus*	*Peromyscus polionotus*
M. musculus	d_N						
	d_S						
R. norvegicus	d_N	0.0000					
	d_S	0.0639					
R. rattus	d_N	0.0000	0.0000				
	d_S	0.0639	0.0000				
R. tanezumi	d_N	0.0000	0.0000	0.0000			
	d_S	0.1379	0.0666	0.0666			
M. auratus	d_N	0.1144	0.1130	0.1130	0.1122		
	d_S	0.3222	0.2405	0.2405	0.3649		
P. polionotus	d_N	0.1487	0.1470	0.1470	0.1464	0.0679	
	d_S	0.4450	0.3651	0.3651	0.4854	0.2160	

A closer inspection of the pairwise comparison of species within the order Rodentia shows that within the Muridae, only synonymous substitutions have taken place. While d_N values between the two cricetid species and the murines were relatively high with the majority having a d_N of more than 0.1, the d_S values were considerably greater.

6.3.3.1.2 Primates

Table 6.5. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 6 coding region of *Zp3* between six species from the Order Primates.

Overall mean d_N = 0.0422, d_S = 0.0869

		Between					
		Homo sapiens	*Pan troglodytes*	*Macaca fascicularis*	*Macaca mulatta*	*Macaca radiata*	*Callithrix jacchus*
H. sapiens	d_N						
	d_S						
P. troglodytes	d_N	0.0000					
	d_S	0.0712					
M. fascicularis	d_N	0.0647	0.0645				
	d_S	0.1548	0.0742				
M. mulatta	d_N	0.0647	0.0645	0.0000			
	d_S	0.1548	0.0742	0.0000			
M. radiata	d_N	0.0664	0.0662	0.0000	0.0000		
	d_S	0.0674	0.0000	0.0696	0.0696		
C. jacchus	d_N	0.0214	0.0214	0.0656	0.0656	0.0673	
	d_S	0.1426	0.0681	0.1461	0.1461	0.0647	

Within the order Primates, the pairwise comparisons shows that the d_S value was higher than the d_N value in most instances, with the exception of *Pan troglodytes* versus *Macaca radiata* (d_N = 0.0662, d_S = 0.0000) and *Callithrix jacchus* versus *M. radiata* (d_N = 0.0673, d_S = 0.0647).

6.3.3.1.3 Carnivora

Table 6.6. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 6 coding region of *Zp3* between five species from the Order Carnivora.

Overall mean d_N = 0.0768, d_S = 0.3733

				Between		
		Canis lupus	*Vulpes vulpes*	*Mustela erminea*	*Mustela putorius*	*Felis catus*
C. lupus	d_N					
	d_S					
V. vulpes	d_N	0.0000				
	d_S	0.0630				
M. erminea	d_N	0.0657	0.0648			
	d_S	0.5058	0.4977			
M. putorius	d_N	0.0657	0.0648	0.0000		
	d_S	0.5058	0.4977	0.0000		
F. catus	d_N	0.0918	0.0908	0.1621	0.1621	
	d_S	0.2912	0.3180	0.5268	0.5268	

Within the Carnivora Order, all d_S values in pairwise comparisons were higher than d_N values. Apart from d_S = 0.00 between the two *Mustela* species, the d_S values ranged from 0.0630 (*Canis lupus* versus *Vulpes vulpes*) to 0.5268 (*Mustela putorius* versus *Felis catus*).

6.3.3.2 Exon 7 of Zp3

Table 6.7. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 7 coding region of *Zp3* within and between each of the Mammalian orders,

Overall mean d_N = 0.3564, d_S = 0.5305

		Within	Between		
			Rodentia	Primates	Carnivora
Rodentia	d_N	0.2250			
	d_S	0.2686			
Primates	d_N	0.0278	0.3994		
	d_S	0.1344	0.7332		
Carnivora	d_N	0.0991	0.6484	0.3288	
	d_S	0.1962	0.7955	0.4628	

Within each order the d_S value was again higher than the d_N values, although the difference between the two values was less than seen in exon 6. Within Rodentia, the d_N and d_S values were similar with d_S being slightly higher. Within the Primate order, d_N was very low (0.0278) compared to that within Rodentia (0.2250) and the d_S rate was higher than that for exon 6. Between orders, the mean d_S values were again higher than d_N although the difference between the two values was considerably less than

that seen in exon 6. This high d_S value relative to d_N, suggests that the majority of the exon 7 coding region has evolved under purifying selection.

6.3.3.2.1 Rodentia

Table 6.8. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 7 coding region of *Zp3* between six species from the order Rodentia.

Overall mean d_N = 0.2250, d_S = 0.2686

				Between			
		Mus musculus	*Rattus norvegicus*	*Rattus rattus*	*Rattus tanezumi*	*Mesocricetus auratus*	*Peromyscus polionotus*
M. musculus	d_N						
	d_S						
R. norvegicus	d_N	0.0950					
	d_S	0.0719					
R. rattus	d_N	0.0950	0.0000				
	d_S	0.0719	0.0000				
R. tanezumi	d_N	0.0951	0.0130	0.0130			
	d_S	0.0716	0.0000	0.0000			
M. auratus	d_N	0.3058	0.3458	0.3458	0.3426		
	d_S	0.1977	0.1495	0.1495	0.1494		
P. polionotus	d_N	0.3590	0.3517	0.3517	0.3498	0.3116	
	d_S	0.6562	0.6166	0.6166	0.6168	0.6619	

The pairwise comparisons between the four Muridae species shows that d_N is higher than d_S in all comparisons, although within the *Rattus* genus the d_N value was low. In Chapter 5, the d_N value for murines was variable with approximately half of the pairwise comparison having a lower d_N than d_S value. *Peromyscus polionotus*, in all five pairwise comparisons, had very high d_S values relative to the d_N values and to those in exon 6, ranging from 0.6166 to 0.6619. These values are in contrast to the pairwise values of *Mesocricetus auratus* whereby the d_N values were higher in all comparisons except with *P. polionotus*.

6.3.3.2.2 Primates

Table 6.9. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 7 coding region of *Zp3* between six species from the Order Primates.

Overall mean d_N = 0.0278, d_S = 0.1344

				Between			
		Homo sapiens	*Pan troglodytes*	*Macaca fascicularis*	*Macaca mulatta*	*Macaca radiata*	*Callithrix jacchus*
H. sapiens	d_N						
	d_S						
P. troglodytes	d_N	0.0252					
	d_S	0.0000					
M. fascicularis	d_N	0.0380	0.0125				
	d_S	0.1197	0.1186				
M. mulatta	d_N	0.0380	0.0125	0.0000			
	d_S	0.1197	0.1186	0.0000			
M. radiata	d_N	0.0380	0.0125	0.0000	0.0000		
	d_S	0.1197	0.1186	0.0000	0.0000		
C. jacchus	d_N	0.0771	0.0506	0.0375	0.0375	0.0375	
	d_S	0.2972	0.2931	0.2367	0.2367	0.2367	

In only one pairwise comparison (*Homo sapiens* versus *Pan troglodytes*) was the d_N value higher than the d_S value, and in this instance the d_N value was quite low (d_N = 0.0252, d_S = 0.0000). In all other pairwise comparisons, including those between *Homo sapiens* versus *Callithrix jacchus*, the d_S value exceeded the d_N value by over 100%. The three species of *Macaca* all had no synonymous or nonsynonymous substitutions in pairwise comparisons.

6.3.3.2.3 Carnivora

Table 6.10. Mean estimated pairwise comparisons of nonsynonymous substitutions per nonsynonymous site (d_N – in bold) and synonymous substitutions per synonymous site (d_S) for the exon 7 coding region of *Zp3* between five species from the Order Carnivora.

Overall mean d_N = 0.0991, d_S = 0.1962

				Between		
		Canis lupus	*Vulpes vulpes*	*Mustela erminea*	*Mustela putorius*	*Felis catus*
C. lupus	d_N					
	d_S					
V. vulpes	d_N	0.0203				
	d_S	0.0154				
M. erminea	d_N	0.0928	0.0703			
	d_S	0.2302	0.2107			
M. putorius	d_N	0.0931	0.0706	0.0000		
	d_S	0.2734	0.2527	0.0299		
F. catus	d_N	0.1668	0.1424	0.1674	0.1676	
	d_S	0.3009	0.2778	0.1635	0.2070	

The pairwise comparison between *Canis lupus* and *Vulpes vulpes* showed that the d_N value of 0.0203 was higher than the d_S value of 0.0154. Apart from this, all other pairwise comparisons showed higher d_S than d_N values, although d_S values were lower than was seen in exon 6.

6.3.4 Likelihood ratio tests (LRTs) of positive selection.

To further assess the evidence for positive selection using a statistical approach, LRTs of codon substitution models were conducted (see Chapter 2.10.2 for theory and method).

Table 6.11 shows the LRTs using all species with the exceptions of the three Rodentia species aforementioned (see 6.2.3). The phylogeny used in the analysis is that proposed by Murphy *et al.* (2001) (Fig. 6.6).

Table 6.11. Likelihood ratio tests of models of codon substitution using PAML (version 3.15: Yang 1997) of exon 6 and 7 coding sequences of *Zp3* from species belonging to the Orders Rodentia, Primates and Carnivora and computed using the phylogeny of Murphy *et al.* (2001).

Model	ℓ	2Δℓ			df	*P* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-1019.99						0.5375	p_0 = 0.5314, p_1 = 0.4686 ω_0 = 0.1297, ω_1 = 1	
M2a	-1019.99	0.00			2	1.00	0.5419	p_0 =0.5335, p_1 = 0.4665, p_2 = 0.4665 ω_0 = 0.13117,ω_1 = 1, ω_2 = 1.011	none
M7	-1020.04						0.5047	p = 0.3428, q = 0.3365	
M8	-1019.65		0.8		2	0.99	0.5460	p = 0.9096, q = 4.055 p_0 = 0.5847, p_s = 0.4153, ω_s = 1.0608	None above 68%
M8a	-1019.67			0.00	1	1.00	0.5276	p = 0.9833, q = 4.9453 p_0 = 0.5639, p_s = 0.4361, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *P* (χ2) = *P* values of LRT under χ^2 distribution, significant values (<0.05); $\omega = d_N/d_S$; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

Three LRTs were conducted, comparing null models of no positive selection with the alternative models of positive selection. In all three LRTs, the null model could not be rejected in favour of the alternative

model of positive selection (P = 1). The positive selection models, despite not fitting the data significantly better than the no positive selection model, indicated that 46% of codons were evolving at a nearly neutral rate (ω = 1.01). Those codons selected did not have posterior probabilities of evolving under positive selection above 68%.

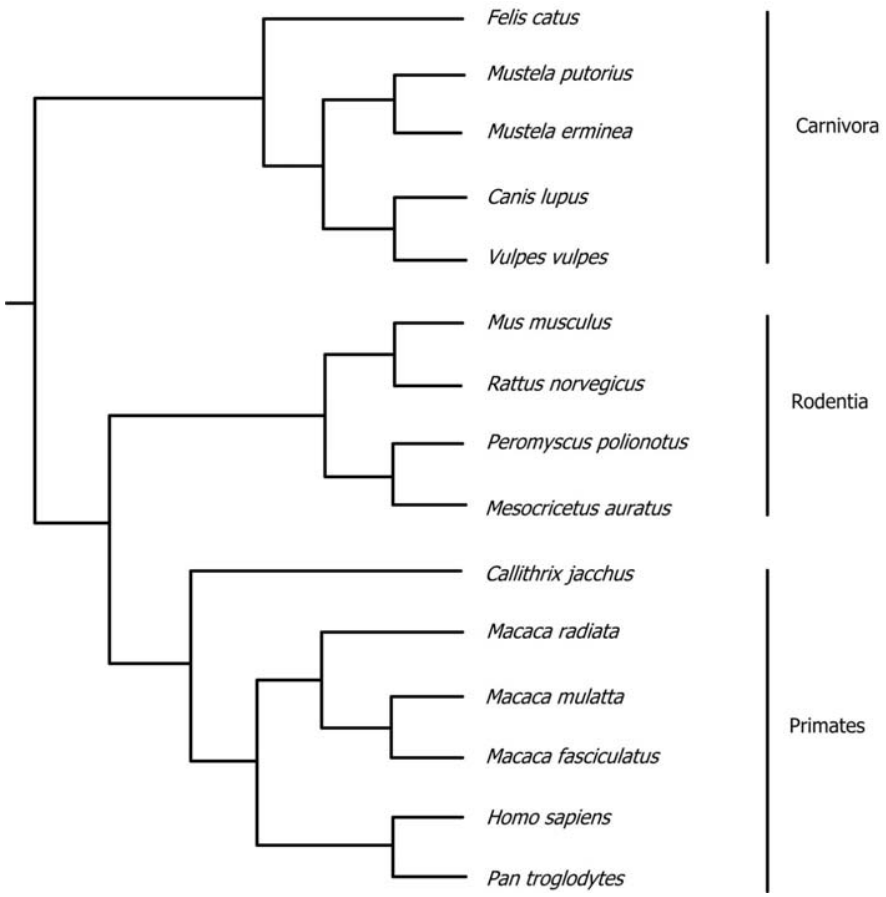


Fig. 6.6. Mammalian phylogenetic tree based on hypothetical relationships proposed by Murphy *et al*. (2001).

In order to investigate whether there are different patterns of adaptive evolution occurring within the different orders, LRTs of the three models were conducted on each individual order.

6.3.4.1 Rodentia

The LRTs were conducted using the phylogeny of Steppan *et al.* (2005) (Fig. 6.7).

Table 6.12. Likelihood ratio tests of models of codon substitution using PAML (version 3.15) of exon 6 and 7 coding sequences of *Zp3* from species belonging to the Order Rodentia and computed using the phylogeny of Steppan *et al.* (2005).

Model	ℓ	$2\Delta\ell$			df	$P(\chi 2)$	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-483.94						0.5246	p_0 = 0.5370, p_1 = 0.4630 ω_0 = 0.1148, ω_1 = 1	
M2a	-479.57	8.74			2	< 0.05	1.2310	p_0 =0.8769, p_1 = 0.0000, p_2 = 0.1231 ω_0 = 0.4051, ω_1 = 1, ω_2 = 7.1156	333S (93.5%) 335Q (82.4%) 341P (86.8%) 342R
M7	-484.32						0.5696	p = 0.1334, q = 0.1007	
M8	-479.51		9.59		2	< 0.05	1.2515	p = 1.7606, q = 2.2607 p_0 = 0.8848, p_s = 0.1152, ω_s = 7.5072	333S 335Q (87.1%) 341P (90.4%) 342R
M8a	-483.95			8.88	1	< 0.05	0.5251	p = 13.0182, q = 99.0000 p_0 = 0.5371, p_s = 0.4629, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; $2\Delta\ell$ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; $P(\chi 2)$ = P values of LRT under χ^2 distribution, significant values (<0.05); $\omega = d_N/d_S$; ω for each branch calculated under each model; p = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

All three LRTs showed strong support for the rejection of the null hypothesis of no positive selection ($P < 0.05$) in favour of the positive selection model. Between 11% and 12% of codons were evolving under positive selection although only two were found to have posterior probabilities of greater than 95%. The ω value was between 7.11 and 7.5, suggesting that the percentage of codons were evolving under strong positive selection. In total, four codons were found to be evolving under positive selection (with posterior probabilities greater than 82%) and of these two (335 and 342) were found to be under positive selection within the murine rodents in Chapter 5.

Alone, these results should be treated with some caution due to the low representation of species (four). However, when added to the results of Chapter 5 whereby strong support for positive selection was

found within the Murinae, it does appear that exon 7 of *Zp3* has evolved under positive selection among species within the order Rodentia.

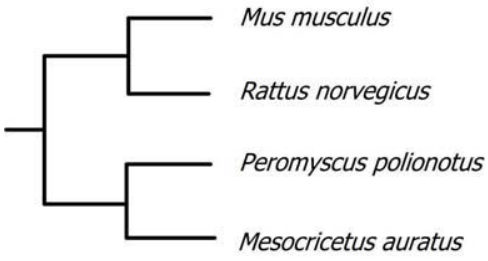


Fig. 6.7. Rodentia phylogenetic tree based on hypothetical relationships proposed by Steppan et al. (2004).

6.3.4.2 Primates and Carnivora

LRTs of the three models of codon substitution were conducted using the phylogeny of Purvis (1995) for the primates (Table 6.13 and Fig. 6.8) and the phylogeny of Bininda-Emonds *et al.* (1999) for the Order of Carnivora (Table 6.14 and Fig. 9).

Table 6.13. Likelihood ratio tests of models of codon substitution using PAML (version 3.15) og exon 6 and 7 coding sequences of *Zp3* from species belonging to the Order Primates and computed using the phylogeny of Purvis *et al.* (1995).

Model	ℓ	2Δℓ			df	*P* (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-334.65						0.3054	p_0 = 0.7673, p_1 = 0.2327 ω_0 = 0.0948, ω_1 = 1	
M2a	-334.58	0.00			2	1.00	0.3344	p_0 =0.9492, p_1 = 0.0000, p_2 = 0.0508 ω_0 = 0.1968, ω_1 = 1, ω_2 = 2.907	None
M7	-334.67						0.3080	p = 0.1562, q = 0.3511	
M8	-334.58		0.18		2	0.98	0.3344	p = 24.4430, q = 99.0000 p_0 = 0.9496, p_s = 0.0504, ω_s = 2.9098	None above 70%
M8a	-334.65			0.18	1	1.00	0.3055	p = 0.9833, q = 4.9453 p_0 = 0.5639, p_s = 0.4361, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; 2Δℓ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; *P* (χ2) = *P* values of LRT under χ^2 distribution, significant values (<0.05); ω = d_N/d_S; ω for each branch calculated under each model; *p* = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

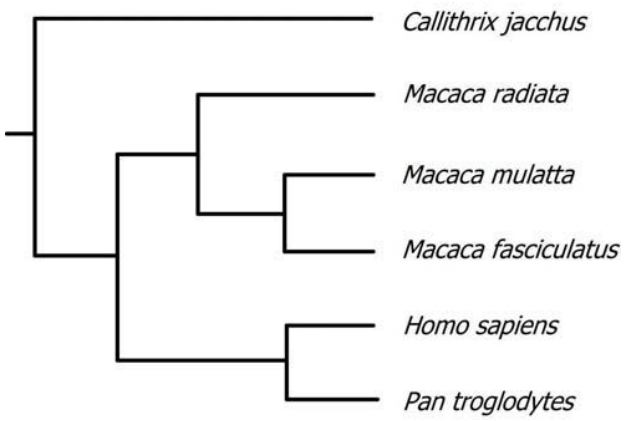


Fig. 6.8. Primate phylogenetic tree based on hypothetical relationships proposed by Purvis 1995

Table 6.14. Likelihood ratio tests of models of codon substitution using PAML (version 3.15) of exon 6 and 7 coding sequences of *Zp3* from species belonging to the Order Primates and computed using the phylogeny of Bininda-Emonds *et al.* (1999).

Model	ℓ	$2\Delta\ell$			df	P (χ2)	ω for each branch	Parameter estimates	Positively selected sites (PP > 95%, PP > 99%*)
		M1a-M2a	M7-M8	M8a-M8					
M1a	-412.96						0.3928	p_0 = 0.6072, p_1 = 0.3928 ω_0 = 0.0000, ω_1 = 1	
M2a	-412.96	0.00			2	1.00	0.3928	p_0 =0.6072, p_1 = 0.2610, p_2 = 0.1318 ω_0 = 0.0000, ω_1 = 1, ω_2 = 1	None above 51%
M7	-412.96						0.3946	p = 0.0127, q = 0.0214	
M8	-412.96		0.00		2	1.00	0.3928	p = 0.0050, q = 2.4316 p_0 = 0.6072, p_s = 0.3928, ω_s = 1	None above 70%
M8a	-412.96			0.00	1	1.00	0.5276	p = 0.0050, q = 11.5099 p_0 = 0.6072, p_s = 0.3928, ω_s = 1	

Notes: In the above tables, ℓ = log likelihood ratio; $2\Delta\ell$ (-2 lnL) = twice the difference between log likelihood ratios; df = degrees of freedom; P (χ2) = P values of LRT under χ^2 distribution, significant values (<0.05); $\omega = d_N/d_S$; ω for each branch calculated under each model; p = parameters; PP = posterior probability calculated under Bayes Empirical Bayes method.

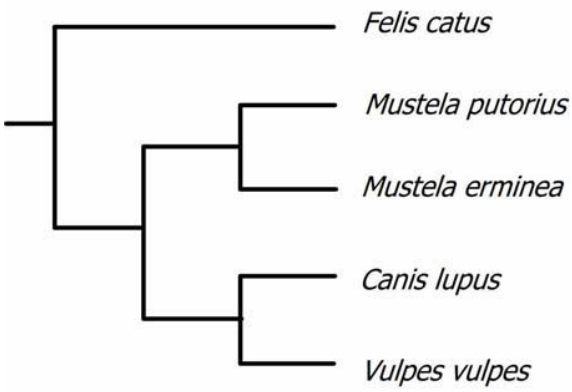


Fig. 6.9. Carnivora phylogenetic tree based on hypothetical relationships proposed by Bininda-Emonds *et al.* 1999

In contrast to the LRTs of *Zp3* for the species in the Order Rodentia, the LRTs for species from the Primate and Carnivora Orders did not support rejection of the null model of no positive selection ($P = 1$). While the ω value in respect of the Primates was 2.9, no codon site was selected with a posterior probability greater than 70%. In respect of the Carnivora Order the percentage of codons in the class for positive selection was high (39%) and yet the ω value was 1. No codons were selected with posterior probabilities greater than 70%. This suggests that the whole of the exon 7 coding region is evolving under neutral selection within the group of carnivores investigated, in the present study at least.

6.4 Discussion

In Chapters 3 and 4, amino acid comparisons of the exons 6 and 7 coding regions of *Zp3* of murines from Africa, South-east Asia, New Guinea and Australasia showed that closely related species generally shared the same amino acid sequence, with divergence only apparent among more distantly related species. Hence species that have recently speciated and occur in sympatry over part of their region, do not appear to have amino acid changes in the region of *Zp3* thought to be involved in sperm-zona pellucida binding in the mouse. A similar pattern is also evident in the comparison of the exon 7 coding region between a limited number of species amongst the species from the Orders Primates and Carnivora. For example, the three species of the genus *Macaca* share an identical amino acid sequence, with no synonymous substitutions occurring within exon 7. Similarly, in the Order Carnivora, two species from the genus *Mustela* share an identical sequence within the exon 7 coding region with only a small number of synonymous substitutions being evident.

Within both the Orders Primates and Carnivora high synonymous substitution and low nonsynonymous substitution rates were evident. When the models of codon substitution were tested, the first LRT used species from all three orders and no support was found for the occurrence of positive selection within the exons 6 and 7 of *Zp3*. However, when the sequences from each order were used, only the LRTs using species from the Order Rodentia provided support for positive selection with four sites identified at which positive selection has taken place. Two of these sites were also found to be have evolved under positive selection within the Muridae (see Chapter 5).

The results from the LRTs using data from all three Orders supports the findings of Berlin and Smith (2005). A limitation of the codon substitution models, as stated previously, is the inability of the computing program to handle alignment gaps. Berlin and Smith used sequences from species that contained deletions such as those from the cow, pig and rabbit. This meant that potentially informative codon sites were removed from the analysis and therefore the analyses may have failed to detect

evidence of positive selection at those sites. Two of the three Rodentia species that were removed from the analysis in the present study (those of *Lasiopodomys brandtii* and *Lagurus lagurus*) both contained deletions at codon site 342 (Fig. 6.1). Therefore, alignment gaps and the low sampling of murine species are possible reasons for the failure to detect evidence of positive selection in the Berlin and Smith (2005) analysis. Notwithstanding these limitations, the LRTs using data from the three Orders did not contradict those of Berlin and Smith.

When the individual Orders of mammals were subjected to the codon substitution models only sequences from rodent species provided evidence for the positive selection model, thus supporting the results of Chapter 5. Of the four species used in the present analysis, *Peromyscus polionotus*, an American cricetid, was one of the species that were used by Turner and Hoekstra (2006) to demonstrate that *Zp3* was evolving under positive selection within that genus.

The observation that the amino acid sequence encoded by exon 7 of *Zp*3 shows little or no divergence between closely related species, appears to be a general trend across mammals. The LRTs of rodent sequences supported the conclusion of Chapter 5 that positive selection has occurred during the evolution of *Zp3* within the murine adaptive radiation, although does not occur in all lineages. There appears to be no support for positive selection occurring within the other two mammalian Orders (Primates and Carnivora) as the LRTs for these groups not support the hypothesis that codons within *Zp3* are evolving under positive selection. However, it is not possible to state that positive selection is not occurring due to the limited number of species used in the analysis. It may be that support for positive selection was not found because of the low number of species sampled or that those species selected are either too closely or too distantly related. Rodents are also believed to evolve at a faster rate than other mammals (Berry & Scriven 2005), thus perhaps increasing the chances of a signal of positive selection being detected. It is also possible that the model of fertilization whereby sperm bind to a site within the exon 7 coding region of *Zp3* applies only to rodents and hence it is possible that there is

no selective pressure on this region of the gene in lineages of other Orders. Another contributing factor may be the as yet unknown role ZP4 plays in fertilization. Expression of *Zp4* has been found in the oocytes of rats and humans (Lefievre *et al.* 2004, Hoodbhoy *et al.* 2005) but not in the mouse, and therefore the models of fertilization, developed in the mouse, do not involve ZP4.

In conclusion, the pattern of low sequence divergence between closely related species is observed within other mammalian Orders. Positive selection has only been detected in lineages from the Order Rodentia. This supports findings from Chapter 5 and although it failed to be detected within the other two Orders, it cannot be ruled out completely, due to the limited sampling of taxa in these Orders and, perhaps, the failure of analyses to handle alignment gaps.

Chapter 7

General discussion

Image on reverse: Australian Old Endemic rodent, *Mesembriomys macrurus.*
Modified image from the private collection of Assoc. Prof. Bill Breed

Chapter 7

General discussion

7.1 General discussion

The 'single glycan' model of sperm-ZP binding in the mouse, proposed by Wassarman and colleagues, hypothesizes that the exon 7 coding region of *Zp3* contains the *O*-linked glycosylation site(s) (serine residues) essential for primary sperm-ZP binding. By contrast, the 'zona scaffold' model, proposed by Rankin, Dean and colleagues suggests that the supramolecular structure of the ZP, influenced by the cleavage status of ZP2, results in sperm-ZP binding, with no particular glycan or glycoprotein being involved. What is not in dispute is the importance of mouse ZP3 in the assemblage and architecture of the zona matrix, or that the region encoded by exon 7 of *Zp*3 is present in the matrix but is not essential for either the secretion or assembly of the matrix. This region also lacks sequence conservation, relative to the rest of ZP3, between distantly related species (Wassarman & Litscher 1995; Wassarman *et al.* 1999).

ZP3 is a highly glycosylated protein. The arrangement of glycosides is influenced, in part, by the primary structure of the glycoprotein as well as ionic interactions, hydropathy and spatial distribution of the asparagine residues (*N*-linked glycosylation sites) and serine/threonine residues (*O*-linked). Thus, changes to the primary structure of this region may alter glycosylation patterns of the ZP glycoproteins, affecting structural aspects of the matrix such as zona thickness, pore diameter and, in turn, the refractoriness of the ZP to binding of sperm from other species. It is, therefore, possible that changes to the amino acid sequence of ZP3 results in species specificity of sperm-ZP binding if that occurs.

The observation that the region encoded by exon 7 of *Zp3* is highly divergent between disparate groups of mammals has led to the often quoted suggestion that this divergence may contribute to species

specificity of sperm-ZP binding (Kinloch *et al.* 1995; Wassarman 1995b, 1999, 2002; Wassarman & Litscher 1995; Wassarman *et al.* 1999, 2001, 2004b; Williams *et al.* 2003, 2006). In addition, there is some suggestion that codons within exon 7 have rapidly evolved due to positive Darwinian selection, although at least in mammals good evidence for this appears to be restricted to the *Peromyscus* lineage (Turner & Hoekstra 2006).

If the amino acid sequence of this region has undergone rapid divergence, and this has resulted in species specificity of primary sperm-ZP binding, then it might be expected that sequence divergence would have taken place between closely related species, especially those living in sympatry, due to positive selection. Rapid divergence of amino acids that are involved in sperm recognition, may evolve under selective pressure in order to prevent hybridization. The specific hypotheses tested in this thesis are that:

- There is a high level of sequence divergence within the region encoded by exon 7 of *Zp3* between closely related species which contributes to potential species specific sperm-ZP binding; and

- The region encoded by exon 7 of *Zp3* has undergone rapid evolution due to positive selection.

In chapters 3 and 4, the nucleotide and predicted amino acid sequences of the regions encoded by exons 6 and 7 of *Zp3* for 96 murine species from Africa, Eurasia, South-east Asia, New Guinea and Australasia were determined. Comparison of the amino acid sequences showed that, in general, closely related species showed a lack of amino acid sequence variation in this region, a finding contrary to the prediction of the hypothesis that in order to prevent interspecific fertilization rapid divergence of these amino acids would occur between closely related species. In fact, within the Australasian taxa even relatively distantly related species, such as *Hydromys chrysogaster* and *Notomys alexis* (Fig. 5.1),

showed no nucleotide sequence divergence within the exon 7 region of *Zp3* (Appendix 1), a finding which contrasts to the sequence divergence within the *Peromyscus* genus (Turner & Hoekstra 2006).

This pattern of similarity of sequence of exon 7 of *Zp3* between species within the same genus is repeated across all murine divisions albeit that sampling of the African and South-east Asian divisions was not as extensive as those from the Australasia. Hence, the two *Aethomys* species from Africa share the same amino acid sequence of the exon 7 coding region, and yet *Micaelamys namaquensis* (a member of the same division) does not. Within the South-east Asian species group the two *Leopoldamys* species have in common all but one amino acid within this region. Three South-east Asian species of *Bandicota indica*, *Bunomys andrewsi* and *Paruromys dominator* all share the same sequence. Within the Pogonomys division (New Guinean Old Endemics), species of the same genus share the same amino acid sequence (for example, *Mallomys* and *Mammelomys* genera). Furthermore, amongst the Hydromys and Xeromys divisions (Australasian Old Endemics) no divergence is seen and, in fact, all six species share the same amino acid sequence despite being members of different genera and divisions. A large percentage of species within the Pseudomys division also share an identical amino acid sequence, regardless of whether they are allopatric or sympatric. Chapter 6 shows that this pattern of conservation of amino acids between closely related species is also found to be the case between Primate and Carnivora species.

This lack of sequence divergence between closely related species negates support for the first hypothesis, and indicates that in general there is sequence similarity or even identity, of the region encoded by exon 7 of *Zp3* in closely related species.

Nevertheless, in a few lineages sequence divergence between closely related species has occurred. Where this is the case, it tends to be lineage specific. For example, *Rattus* species that are endemic to Australia share a single amino acid insertion that is not present in species of *Rattus* occurring in New

Guinea (for example, *R. leucopus*). Three *Melomys* species endemic to the Australian mainland have an alanine in position 341 while species from all other Australasian Old Endemic divisions have a serine. Furthermore, five *Pseudomys* species with distinctive pebble-mound building behaviour, all share a proline in position 334 which contains a serine in all other murine species investigated.

Despite little, or no, sequence divergence within genera, strong statistical evidence was detected for positive selection of the exon 7 coding region having occurred within the African, South-east Asian and New Guinean divisions. The results of the maximum likelihood analyses that detected evidence of positive selection were consistent despite using a range of different phylogenetic hypotheses, suggesting that the detection of positive selection was unlikely to have been as a result of phylogenetic error. Like sequence divergence, those codons that have apparently evolved under positive selection vary across lineages. Hence the two codon sites that have evolved under positive selection in species from the African and South-east Asian divisions (mZP3-335 and -342) appear not to have rapidly diverged within the New Guinean Old Endemic divisions. Within this latter group a different subset of codons (mZP3-324, -325, -341) was identified as evolving under positive selection, with only one of these (324) diverging in a few lineages within species from the Australasian Old Endemic divisions. These results suggests that positive selection of exon 7 of *Zp3* may have only taken place in the lineages leading to extant taxa within the Pogonomys and Lorentzimys divisions. These findings support the second hypothesis, that codons within the region encoded by exon 7 of *Zp3* have undergone rapid evolution due to positive selection, albeit that it appears to have occurred only in some lineages and not in others.

The codon substitution models employed to detect evidence of positive selection have some limitations that are worth discussing. Evidence of adaptive evolution can only be detected if there is an excess of nonsynonymous substitutions relative to synonymous substitutions (Wong *et al.* 2004). In addition, rates of synonymous and nonsynonymous substitutions need to be reasonably high for these methods to

detect selection. When divergence levels are low there is often not enough information (such as between species from the Australasian taxa), whereas synonymous substitutions are likely to be saturated at high levels (such as between species from different mammalian Orders) (Wong *et al.* 2004). Despite these limitations, Turner and Hoekstra (2006) detected evidence of positive selection within species of the same genus. Therefore, a failure to detect evidence of positive selection within a data set does not necessarily mean that it is not occurring. Certainly, these methods should not be used to test a hypothesis that positive selection is not occurring.

Notwithstanding the strong statistical support for positive selection acting on a subset of codons in a few lineages, there is also good evidence for purifying selection, as demonstrated by the high synonymous to nonsynonymous substitutions rates seen within, and even between a few, divisions. One particular example of this is seen within the Stenocephalemys division (*Mastomys natalensis* and *Hylomyscus alleni*) where there is a high synonymous substitution rate (0.2390) compared to the rate of nonsynonymous substitution (0.0561). This pattern of high d_S and low d_N occurred in all pairwise comparisons involving the two species in the Stenocephalemys division. However, in most pairwise comparisons between divisions, d_N exceeded d_S.

The different codons within exon 7 of *Zp3* identified as evolving under positive selection during the adaptive radiation of the Old World murines suggest that the causative selective pressure(s) may also vary across the different lineages. There are several hypothetical models for selective pressures that result in rapid divergence of reproductive proteins due to positive selection. These include reinforcement due to selection against unfit hybrids, co-evolution of gametes due to intra- and inter-male sperm competition, sexual selection and conflict, and immune response to microbial attack.

7.1.2 Reinforcement

Reinforcement is the rapid evolution of reproductive barriers to minimize the occurrence of unfit hybrids (Clark *et al.* 2006). Sequence divergence of the exon 7 coding region of *Zp3* may, therefore, evolve by positive selection between closely related species that are present in sympatry especially if their reproductive anatomy is similar. In three pairs of Australasian Old Endemic species that occur in sympatry in part of their range, *Notomys alexis* and *N. mitchelli*, *Pseudomys hermannsburgensis* and *P. bolami*, and the two *Leggadina* species, there is an identical amino acid sequence of the exon 7 coding region of *Zp3.* This finding indicates that positive selection resulting from reinforcement has not resulted in divergence of exon 7 of *Zp3* in these species.

Furthermore, in *Rattus,* viable hybrid offspring between different crosses from various Australian species have been produced in the laboratory (Baverstock *et al.* 1983) suggesting that there is no barrier to interspecific fertilization at the sperm-ZP binding level in these species. The fertility and fecundity of their offspring varied although detailed data was only available for some of the crosses. The amino acid sequence of the exon 7 coding region of *Zp3* varies by one residue between all the Australian *Rattus* species (Fig. 4.2). The ability of these rodents to cross-fertilize and the lack of sequence divergence within the exon 7 coding region supports the above suggestion that there is little selective pressure on the exon 7 coding region of *Zp3* to prevent hybridization. At least two of the *Rattus* species occur in sympatry but as yet hybrids have not been found in the wild (Baverstock *et al.* 1983). In addition, viable offspring were produced from a cross between *Rattus fuscipes* and *R. leucopus* (Baverstock *et al.* 1983). Although these two species live in sympatry in the north-east of Australia, *R. leucopus* is a member of a largely New Guinean species group (see Chapter 1.2.8.2.4), and lacks the single amino acid insertion within the exon 7 coding region of *Zp3* present in all other Australian *Rattus* species. Therefore, the divergence within this region does not appear to prevent *in vivo* cross-fertilization between these two species, at least within the laboratory. These findings suggest that, at least within the Australian Old,

and New Endemic murine rodents, positive selection for divergence of the exon 7 coding region of *Zp3* to minimize the occurrence of hybridization has not taken place despite a number of these species occurring in sympatry.

A contrast between species whose populations are allopatric and those in sympatry may provide a framework for testing reinforcement as a driving force of rapid evolution (Clark *et al.* 2006). The advantage of determining the amino acid sequence of the exon 7 coding region of *Zp3* within such a large group of murine species allows such a comparison. In species that are clearly allopatric, such as the New Guinean Old Endemic *Leptomys elegans* and the Australian Old Endemic *Leggadina forresti,* an identical exon 7 coding region sequence occurs. Two allopatric *Pseudomys* species, *P. delicatulus* from Northern Australia and *P. higginsi* from Southern Australia/Tasmania, share an identical amino acid sequence with two sympatric *Notomys* species, *N. alexis* and *N. mitchelli*. This would appear to rule out reinforcement as a driving force of rapid evolution of the exon 7 coding region of *Zp3*, at least in this group of murine rodents.

7.1.3 Co-evolution of gametes

One form of sperm competition that may occur is between sperm within an ejaculate from the same male. Thus, there may be selection on sperm within the female tract, which may occur at the level of the zona pellucida, in order to prevent sperm from binding to, and penetrating the zona matrix, after fertilization. It is possible, within this model, there may be selection on the ZP composition and molecular organization to minimize the occurrence of multiple sperm interacting with the oocyte (polyspermy). Thus increasing sperm competition and selection within the female tract may have a positive feedback effect on both sperm quality and quantity (Cummins 1990).

An example of the co-evolution of gametes is in the development of a diverse range of sperm head shapes. In mammals, it has been suggested that interspecific differences in sperm head shape may relate to the variation in the structural organization of the ZP (Rankin & Dean 2000). Most species of

murine rodents contain a sperm head that is hook shaped or falciform with extensive cytoskeletal material extending into the hook as a perforatorium. This sperm head morphology occurs in the laboratory mouse and rat and is also present in most South-east Asian species investigated in this study, with the exception of *Bandicota indica* which has a conical or bulbous head shape (Fig. 7.1A). Nevertheless, there is no sequence divergence seen in the exon 7 coding region of *Zp3* between *Bandicota indica* and that of other South-east Asian murines such as *Paruromys dominator* and *Bunomys andewsi* (Fig. 7.1C and I). Likewise, in the two African *Aethomys* species an identical amino acid sequence of the exon 7 coding region of *Zp3* occurs in spite of marked differences in sperm head shape and tail length (Fig. 7.2D and E).

NOTE: This table is included on page 206 of the print copy of the thesis held in the University of Adelaide Library.

Fig 7.1. Sperm heads of South-east Asian murines, including Rattus. Images A) to E) are scanning electron microscopy images and images F) to J) are light microscope images.

1. Taken from Breed and Yong (1986)
2. Taken from Breed and Musser (1991)
3. Taken from Breed and Aplin (1994)
4. Taken from Breed (1997)
5. Taken from Breed (2004)

NOTE: This figure is included on page 207 in the print copy of the thesis held in the University of Adelaide Library.

Fig. 7.2. Sperm head shapes of African murines rodents. Images from A) to C) are light microscope images and D) to E) are scanning electron microscopy images. Abbreviations: AH = apical hook, VP = ventral process.

1. Taken from Breed (1995)
2. Taken from Breed (2005)

NOTE: This figure is included on page 207 in the print copy of the thesis held in the University of Adelaide Library.

Fig. 7.3. Sperm head shapes of New Guinean Old Endemic rodents. Images from A) to G) are light microscope images. Abbreviations: AH = apical hook, VP = ventral process

1. Taken from Breed and Aplin (1994)

NOTE: This table is included on page 208 of the print copy of the thesis held in the University of Adelaide Library.

Fig. 7.4. Sperm head shapes of representative species from Australasian Old endemicmurine rodents. Images A) to S) are scanning electron images, while images T) to V) are light microscopy images. Abbreviations: AH = apical hook, ES = equatorial segment, VH = ventral hook, VP = ventral process, VS = ventral spike.

1. Taken from Breed (1983)
2. Taken from Breed (1984)
3. Taken from Breed (1997)
4. Taken from Breed and Aplin (1994)

The morphology of the sperm head in most species of Australasian Old Endemic rodents is more complex than that of the African and South-east Asian murines, due to the presence of two ventral processes that extend from the upper concave surface (Breed 1997) (Fig. 7.4). Electron microscopy of eggs and sperm recovered from recently mated females suggest that these ventral processes may contain molecules for zona binding (Breed & Leigh 1991) albeit that a few species in the genera *Pseudomys* and *Notomys* do not have these extensions present in the sperm head (*P. delicatulus, P. novaehollandiae, P. pilligaensis, P. shortridgei* and *Notomys alexis* Fig. 7.4). Nevertheless, the results of the present study indicate that, in general, the Australasian Old Endemic rodents with divergent sperm morphology have little or no variation in the amino acid sequence within the exon 7 coding region of *Zp3*. While it is tempting to suggest that the divergent sperm heads have co-evolved with the changes to the combining-site for sperm within exon 7 of *Zp3* (in particular the two serine residues at positions 336 and 341), the results indicate that this is probably not the case. Amongst the New Guinean Old Endemic murines two species do not have the serine residues in positions 336 and 341, yet have the additional ventral processes (Fig. 7.3) similar to that in Australasian Old Endemic murines.

Thus, these comparative findings on sperm head structure from African, Eurasian, New Guinean and Australasian murine rodents do not support the notion that there is an association between the sperm head shape and the amino acid divergence of the ZP3 region purported to be involved in sperm-ZP binding.

Furthermore, McGregor *et al.* (1989) found that, in three species of Australasian Old Endemic murines, the species with the least divergent sperm head cytoskeleton, *Notomys alexis*, had a thicker zona matrix compared to that of *Pseudomys australis.* Since the amino acid sequence of the exon 7 coding region between these two species is identical, there also appears to be no relationship between the zona thickness and the primary sequence of this region of ZP3.

If the co-evolution of gametes is a driving selective force then it could be predicted that positive selection may also be detected in male reproductive proteins. There have been many candidate male reproductive proteins involved in fertilization over the past two decades, and, unlike ZP3, no single protein has been established as the essential sperm zona-binding protein. The difficulty in locating this protein may be due to the possibility that there are numerous receptors on the sperm head, which would account for the multiple affinities of ZP3 to sperm (Thaler & Cardullo 1996). A few of these candidate proteins (PH20, fertilin β and α, zonadhesin, acrosin and SP17) appear to be evolving under positive selection (Swanson *et al.* 2003) although not all are known to interact with ZP3 as some are located within the acrosome. The presence of β1,4-galactosyltransferase (GalTase) on the outer acrosomal membrane of mouse sperm, has been found to be involved in sperm-ZP3 binding and induction of the acrosome reaction (Miller *et al.* 1992) but transgenic mice with no active sperm GalTase were still fertile (Lu *et al.* 1997) suggesting that GalTase could be one of many proteins interacting with the zona matrix (Miller *et al.* 2002). Swanson *et al.* (2003) did not find evidence of positive selection occurring within the GalTase gene. Unfortunately, it is not known which of the candidate sperm zona-binding proteins interact with the exon 7 coding region of *Zp3*, making it difficult to search for co-evolving gamete proteins.

7.1.4 Inter-male sperm competition

The molecular structural organization of the ZP matrix could also evolve to reduce the incidence of inter-male sperm competition (Swanson *et al.* 2003). If this is the case, rapid evolution of ZP glycoproteins may tend to occur where there is inter-male sperm competition to fertilize recently ovulated oocytes. This model would predict that rapidly evolving zona pellucida glycoproteins might occur in species that have a polyandrous or multi-male mating system (Clark *et al.* 2006). In comparative studies of eutherian mammals, relative testis size has been shown to relate to the intensity of potential inter-male sperm competition with species that have a polyandrous mating system having larger testes mass than those

that do not (Harcourt *et al.* 1981; Kenagy and Trombulak 1986). The Australian endemic rodents have a very large range of relative testes size across species. For instance, within the Pseudomys division most *Notomys* have relative testes masses of only 0.1 to 0.2% of body mass, in *Pseudomys shortridgei* and *P. novaehollandiae* it is 0.35 to 0. 45% of body mass whereas in other *Pseudomys* species, e.g. *P. australis, P. desertor, P. fumeus, P. nanus* and *Mastacomys fuscus*, it varies between 1.4 and 3.6% of body mass (Breed & Taylor 2000). A comparison of exon 6 and 7 coding regions of *Zp3* of these species does not show any association between testis size and divergence in the putative combining-site for sperm. This finding is similar to the conclusions drawn by Turner and Hoekstra (2006) in their study of *Zp3* in the genus *Peromyscus.*

7.1.5 Sexual conflict

Another model of evolutionary pressure that may result in rapid divergence of reproductive proteins is sexual conflict. Sexual conflict occurs when traits that promote the reproductive success of one sex reduce the fitness of the other (Gavrilets 2000). In respect of reproductive proteins this could occur when the female reproductive proteins, such as the ZP glycoproteins, continuously evolve in order to reduce the costs of fertilization, while the male reproductive proteins evolve to increase the rate of fertilization (Gavrilets 2000). The prediction of this model is that there may be a rapid rate of evolution of the exon 7 coding region of *Zp3* within and between closely related species in order to reduce the occurrence of frequent matings. However, this is not the pattern found within the murine rodents and thus it would appear that sexual conflict does not drive the divergence of the exon 7 coding region of *Zp3*.

7.1.6 Microbial infection

A final possible selective pressure is that for avoidance of parasitic or viral infections. Microbial attacks could exert a constant selective pressure for gamete surface proteins to evolve in order to evade attackers (Clark *et al.* 2006). In mammals, sexually transmitted pathogens may result in rapid evolution

of surface proteins to minimize the chances of infection. This model would predict rapid evolution both within and between closely related species, although it is possible this would vary in different lineages depending upon the pathogen load. The present study suggests that the pattern of evolution of exon 7 in murine rodents does not result in pathogens being a driving force for the evolution of this region of *Zp3*, although without knowledge of the degree of pathogen risk to a specific population it cannot be completely eliminated as a possible selection agent.

While it is not possible to determine the actual selective pressure that has lead to rapid divergence of the exon 7 coding region of *Zp3* in some lineages, it may be a combination of some, or even all, of the above factors, acting at different times and on different lineages. Mice and rats are thought to be a rapidly evolving group of mammals (Berry & Scriven 2005) and selective pressures may differ biogeographically and over time.

The suggestion that positive selection appears to be acting on different codons of ZP3 and on different lineages within murines, together with the lack of sequence divergence between closely related species, could be due to a number of reasons that do not involve selective pressure, as explored in the next section.

7.1.7 Other reasons for absence of positive selection

Firstly, there could be no selective pressure on a particular lineage due to, for example, a lack of sperm competition, and hence the selective pressure could vary from one lineage to another due to environmental, behavioural or physiological reasons. Secondly, this region of ZP3 may not be involved in sperm-ZP binding in murine rodents. Evidence suggests that there are both low and high affinity binding sites for the sperm on the ZP (Thaler & Cardullo 1996; Castle 2002), and the lack of a definitive zona-binding sperm protein would indicate that there is considerable redundancy in the primary sperm-ZP binding event. With such redundancy in the system, it may be that some lineages lose one molecular interaction yet remain fertile. Experimental evidence from transgenic mice, where candidate

sperm proteins were 'knocked out' and fertility is still retained, supports this (acrosin: Baba *et al.* 1994, GalTase: Lu *et al.* 1997).

A third possibility is that in those species that express a fourth ZP glycoprotein, there is an absence of selective pressure on the carboxyl region of ZP3. The evidence that the exon 7 coding region of *Zp3* is involved in sperm-ZP binding in the mouse is derived mainly from *in vitro* sperm-inhibition studies using ZP3 constructs expressed in either mammalian or prokaryotic expression systems. Early evidence, using similar technology, also supported the same role for this region of ZP3 in the hamster (Moller *et al.* 1990; Kinloch *et al.* 1990, 1992) and, recently, in the bonnet monkey (Gahlay *et al.* 2005). However, it is not known if the mouse model of fertilization applies to all species of mammals. One reason for doubting this is that the ZP of the mouse contains only three glycoproteins. The presence of a fourth ZP glycoprotein in the rat and human, as well as other primates, suggests that in these species the zona matrix is much more complex than in the mouse and, as a role for ZP4 (ZPB) in fertilization has been demonstrated in a range of species, including the pig (Yurewicz *et al.* 1998), the rabbit (Prasad *et al.* 1996), the cow (Topper *et al.* 1997) and human (Govind *et al.* 2000), the molecular events at fertilization may also be more complex than in the mouse. It is possible that a combination of ZP3 and ZP4 may contain recognition sites for the sperm as has been proposed in the pig (Yurewicz *et al.* 1998) and, in this situation, positive selection may not be detected within the carboxyl terminal of ZP3. An example of where this may have occurred is between the human and chimpanzee exon 7 coding region, where there is only one residue difference between the sequences of the two species (Table 6.1). As ZP4 has been found to be expressed in the oocytes of both species (Conner *et al.* 2005), it may be that, like in the pig, a combination of both ZP glycoproteins are involved in sperm-ZP binding. Hence positive selection may not act on *Zp3* alone. ZP4 is also expressed in rat oocytes but not in mouse, and it is not known whether the *Zp4* pseudogene in mice was derived from its ancestor or occurred solely within this species. The exon 7 coding region of *Zp3* within the Australasian Old Endemic murines is more similar

to that of the mouse than it is to the rat. As it is not clear if the Australasian Old Endemics shared a more recent ancestor with the mouse, it is possible that the *Zp4* gene is expressed within this group. This could conceivably be why positive selection within *Zp3* is not detected. It would be interesting to determine if *Zp4* is a pseudogene in the Old Endemic murines of New Guinea and Australasia and whether there is an association between the lack of selective pressure on *Zp3* with the presence of an expressed form of *Zp4*.

Species specificity of sperm-ZP binding, of course, may not be due to changes to the amino acid sequence of *Zp3*, but due to differential glycosylation. It is possible that there are interspecific differences in the oligosaccharides of the zona matrix, and in particular to the terminal sugars that interact with sperm. This may be due to the glycosylation mechanisms of a particular species allowing sperm to bind to a certain pattern of sugars found only within the same species. With this in mind, a preliminary study was conducted, comparing lectin staining of ZP from follicular oocytes of two species of Old Endemic Australian murines (*Notomys alexis* and *Pseudomys australis*) and the laboratory mouse. Although some differences were seen between the mouse and the Australian species, results from repeat experiments were variable and unreliable. In order to get reliable information of the sugar composition of the zona pellucida from comparative species, mass spectrometry needs to be performed. This was not possible in the present study. Therefore, due to the unreliability of the lectin histochemical results, it was decided not to pursue this approach in the present study.

Swanson *et al.* (2001) and Turner and Hoekstra (2006) found evidence of positive selection occurring within ZP2. This protein, in the mouse, has been implicated in secondary binding of sperm to the zona, and also in a cleavage event after fertilization rendering the zona matrix refractory to continued sperm binding. No region of this gene has been found to have functional importance in these events and therefore, the implications of the detection of positive selection occurring within this gene are not known.

7.2 Positive selection and the models of sperm-ZP binding

The single glycan model and the zona scaffold model of sperm-ZP binding differ in the role that sugars play in mediating sperm-ZP interactions. However, the importance of sugar cannot easily be dismissed from the zona scaffold model due to the high level of sugar in the composition of the matrix. As previously stated, changes to the primary sequence of the ZP glycoprotein may introduce differences in charge, hydropathy and glycosylation thereby altering the structure of the matrix. The zona scaffold model proposes that sperm are induced to undergo the acrosome reaction when they enter a pore of the zona matrix (Baibakov *et al.* 2007). Conceivably, the pore diameters, which show interspecies differences (Sinowatz *et al.* 2001), may be influenced by differential glycosylation and thus sperm from one species may not be able to interact with the ZP pores of another species. It is not known whether there is a correlation between diameter of pore size and sperm head shape or size. Therefore, although the detection of positive selection occurring within a few lineages of murine rodents does tend to support the single glycan model of sperm-ZP binding and the involvement of the exon 7 coding region in this process, support for the zona scaffold model cannot be completely ruled out.

7.3 Conclusion

The aim of this project was to test the hypothesis that there is a high rate of divergence within the exon 7 coding region of *Zp3* between closely related murine species and evidence of positive selection acting on codons with this region. It is generally accepted that ZP3 is important for the structural integrity of the ZP matrix and that the exon 7 coding region is present in the secreted form of the glycoprotein. Therefore, it is possible that this region does indeed come into contact with the cell membrane of the sperm and may play a part in species specific recognition between the gametes.

However, in the Old World murines a comparison of the amino acid sequence of the exon 7 coding region of *Zp3* reveals limited divergence of this region in closely related species and among genera. Notwithstanding this observation, evidence for positive selection was detected within the African/South-east Asian and New Guinean murine lineages, albeit at different codon sites. These results support the conclusions of Jansa *et al.* (2005) and Turner and Hoekstra (2006). However, within the Australian Old Endemic murines there was a low level of sequence divergence, with no evidence of positive selection, thus suggesting that different evolutionary forces may be present in these various divisions. Finally, the highly similar amino acid sequence of the exon 7 coding region of *Zp3* across closely related species, despite evidence of positive selection within a few lineages, suggests that selection would appear to be unlikely to be related to molecular co-evolution of binding sites between sperm and the ZP3.

7.4 Future directions

The apparent higher rate of divergence of exon 7 of *Zp*3 observed within the African and South-east Asian murine divisions may be a result of poorly representative sampling of species as compared to the Australian taxa. This would be resolved by obtaining DNA from more South-east Asian species, in particular more species from the same genus. In addition, there have been some interesting results from molecular phylogenetics regarding the probable sister taxa of the New Guinean and Australasian Old endemic murines being species endemic to the Philippines (Steppan *et al.* 2005; Jansa *et al.* 2006).

Although the present study did not sample murines from the Philippines due to the unavailability of specimens, obtaining the DNA sequence of the exon 7 coding region from species of genera such as *Apomys* and *Rhynchomys* would be interesting, especially in the light of recent findings where sperm of these species have a highly complex head morphology similar to that of most Australian Old Endemics (Breed & Leigh 2007). This would be particularly interesting as there is currently some dispute as to whether the sister taxa for the Australian Old Endemic murines occur in the Phillipines or in South-east Asia (see Chapter 1.2.8).

In addition to determining the amino acid sequence from more murine species, it would be interesting to investigate whether *Zp4* is a functional gene or pseudogene within the New Guinean and Australasian Old Endemic murines. It would appear that the active form of *Zp4* has been lost at some stage in the evolution of *Mus* after it diverged from a common ancestor with the rat. At this stage it is not known how many murine species express *Zp4*. Determining the presence/absence of *Zp4* in a broad range of murine rodents, may provide insight into the evolution of the *Zp3* gene in these various murine rodent groups.

A limitation of the method used to detect evidence of positive selection is the requirement of a robust *a priori* hypothetical phylogeny. Currently, published phylogenies are limited either in species numbers or in the data used (such as microcomplement fixation of albumin). The present study attempted to overcome this limitation by using as many phylogenies as were available, and was able to obtain consistency in results. With publication of more robust trees, these analyses may be revisited.