

INVESTIGATING THE ASSOCIATION
BETWEEN GERM LINE SPECIFICATION AND
SEQUENCE EVOLUTION IN VERTEBRATES

Teri Evans (née Forey), BSc. MRes.

Thesis submitted to the University of Nottingham for the degree of

Doctor of Philosophy

JULY 2015

Abstract

Within vertebrates the primordial germ cells (PGCs) can either be induced by embryonic signals (known as epigenesis), or predetermined by maternally deposited germ plasm (preformation). Epigenesis is known to be the ancestral mechanism, while preformation has evolved multiple times. Epigenesis has been proposed to enforce a developmental constraint on the evolution of somatic structures that is released in species which acquired preformation. In accordance with this hypothesis, the mesoderm gene regulatory network is conserved between urodeles and mammals, which have retained epigenesis, but has diverged in anurans (preformation). An increase in speciation has also been shown in vertebrates which have acquired preformation. Our aims were to investigate whether the mode of PGC specification associates with the molecular evolution of protein-coding genes.

We downloaded all publicly available vertebrate sequences. These were combined with our three novel transcriptomes from axolotl, sturgeon and lungfish. In line with previous analyses, we built 4-taxon trees to investigate the extent of phylogenetic incongruence. This revealed a bias associated with the mode of PGC specification, caused by a significant difference in the rate of evolution. Many genes in species that have acquired preformation are evolving significantly faster than in their sister taxa undergoing epigenesis. These sequences are typically expressed in early development, and are ancient genes with known orthologs at the base of Eukaryotes. Additionally, we show that Oct4 and Nanog, which are crucial for pluripotency, have been lost in taxa using preformation. Therefore our results are consistent with the proposal that developmental constraint, imposed by epigenesis, is released in species undergoing preformation.

Acknowledgements

First of all I'd like to thank my supervisor Matt Loose for all his help and advice over the past years. Thank you for teaching me bioinformatics, being a great boss and getting impatient with me when I needed it. If it weren't for you I might still be attempting to assemble the chicken genome by accident. Thanks also to Chris Wade for all your help and support with understanding molecular evolution and how to build phylogenetic trees. I'd also like to thank Andrew Johnson for the hypothesis that started all this work and his constant enthusiasm and advice, as well as a few good pub trips. Thanks also to everyone past and present around the QMC, especially to Laura, Kayleigh, Carolin and Becca. You kept me sane, listened to me rant and always offered a cup of tea (or pint) when necessary.

Huge thanks to my family and friends who've kept me laughing and appreciating life outside of work. Particular thanks go to Mick, Linda, Emily and Alice for your constant support over the past four years. Thank you for providing me a shoulder to cry on in the hard times and always being ready to rush over for chinese and champagne when things went well. Finally, I'd like to thank my husband, Ian, you've always encouraged me and supported me; I could never have completed this PhD without you, so thank you.

Contents

1	Introduction	1
1.1	Vertebrate evolution	2
1.2	Germ line specification	5
1.2.1	Preformation	5
1.2.2	Epigenesis	12
1.3	PGC specification across vertebrates	17
1.4	Developmental Constraint	22
1.4.1	Developmental Constraint and the Germ Line	23
1.5	Genetic Evolution	26
1.5.1	Phylogenetic Trees	26
1.5.2	Rate of Molecular Evolution	28
1.6	Hypothesis and Aims	30
2	Materials and Methods	33
2.1	Programming and data storage	33
2.2	Sequence Library	34
2.2.1	ESTs, mRNAs and cDNAs	34
2.2.2	Transcriptome sequencing	35
2.2.3	Single Genomes for Mapping	36
2.2.4	Additional Transcriptomes	36
2.2.5	Whole Genomes	37
2.2.6	BLAST databases	39
2.3	Orthology finding	39
2.4	Locating the open reading frame	43
2.5	Four-taxon Phylogenies	46
2.5.1	Alignments	46
2.5.2	Building Trees	49
2.6	Whole Gene Phylogenies	50
2.7	Relative Rate Test	51

2.8	Gene Ontology	52
2.9	Synteny	53
3	Global Analysis of Vertebrate Protein-Coding Genes	55
3.1	Data Evaluation	56
3.1.1	Refining the Quality Parameters	58
3.1.2	Protein coding results	61
3.1.3	Likelihood tests of the tree topology	67
3.1.4	Pipeline Summary	70
3.2	Four-Taxon Phylogenetic Trees	70
3.2.1	Amphibians	70
3.2.2	Actinopterygii	76
3.2.3	Sauropsids	80
3.2.4	Conclusion	84
3.3	Distance Matrix	85
3.4	Relative Rate Test	87
3.4.1	Amphibians and Actinopterygii	87
3.4.2	Sauropsids	90
3.4.3	Conclusion	92
3.5	Rate of evolution and tree topology	94
3.6	Conclusion	97
4	Characterising Genes with a Change in Molecular Evolution	101
4.1	Mapping to a Single Genome	102
4.2	Gene Function	105
4.2.1	Gene Ontology	106
4.2.2	Transcription Factors	107
4.3	Gene Expression	109
4.3.1	Time of Expression	109
4.3.2	Location of Expression	114
4.4	Gene Age	117
4.5	Conclusions	121
5	Expanding the Global Analysis	123
5.1	Four-taxon Trees	124
5.2	Relative Rate Test	130

5.2.1	Analysing the rate of evolution within lineages	137
5.3	Gene Characterisation	141
5.4	Coelacanth, Lungfish and Sharks	146
5.4.1	Four-taxon trees	146
5.4.2	Relative Rate Test	152
5.5	Conclusion	153
6	Analysing Pluripotency Genes	155
6.1	Oct4	157
6.1.1	POU gene family	160
6.1.2	The POU5 class	161
6.2	Sox2	171
6.3	Klf4	174
6.4	Nanog	179
6.5	Conclusion	185
7	Discussion	187
7.1	Phylogenetic Incongruence	187
7.2	The rate of evolution and PGC specification	189
7.3	Future work	193
	Bibliography	197
	Appendices	219
A	Additional Tables	221
B	Additional Figures	247

Evans T, Wade C M, Chapman F A, Johnson A D and Loose M (2014).
 Acquisition of germ plasm accelerates vertebrate evolution. *Science*,
 344(6180):200-203

CHAPTER 1

Introduction

There are two fundamental cell types in multicellular organisms; the somatic cells and the germ cells. The germ cells develop into the gametes, which go on to form the next generation, while the somatic cells develop into the rest of the organism. Segregation of these cell types occurs by two distinct mechanisms in vertebrates; epigenesis and preformation. The latter, preformation, begins prior to fertilisation when germ line determinants are deposited and localised in the oocyte (Ikenishi, 1998). During embryo development these determinants asymmetrically segregate and the cells that retain them form the germ line. In contrast, epigenesis occurs relatively late in development as signals within the embryo induce a subpopulation to become germ cells (Ohinata et al., 2009).

Epigenesis is the ancestral mechanism and is known to have been retained in mammals and urodeles (Extavour and Akam, 2003; Johnson et al., 2003b). Preformation has evolved independently multiple times within vertebrates, for example in birds, anurans and teleost fish (Hashimoto et al., 2004; Tada et al., 2012; Tsunekawa et al., 2000). It has been observed that preformation is associated with an increase in speciation as well as changes to gene regulatory networks (Crother et al., 2007; Swiers et al., 2010). However, no study has investigated whether there is an association between the mode of germ cell specification and molecular evolution. It is that which this thesis attempts to answer.

Before undertaking a large study across vertebrates it is crucial to understand the natural history of these organisms and how they have evolved, as covered in Section 1.1. This forms the reference point for all molecular evolutionary analyses. We will then discuss details on how the germ line specification mechanisms differ (Section 1.2) and therefore where in the vertebrate natural history preformation has evolved (Section 1.3). Section 1.4 discusses a current hypothesis on how the mode of primordial germ cell specification might impact molecular evolution. Finally, we comment on how sequence evolution can be measured and the procedures used within this study (Section 1.5).

1.1 Vertebrate evolution

Vertebrates are a clade within the deuterostomes, which are in turn a member of the Bilateria (Minelli, 2008). The word deuterostome comes from the Greek "second mouth" and describes those organisms where the blastopore, the initial opening in the embryo, develops into the anus. This is in contrast to the protostomes where the blastopore develops into the mouth. The deuterostomes contain many taxonomic clades as can be seen in Figure 1.1; the relationships among them are described below.

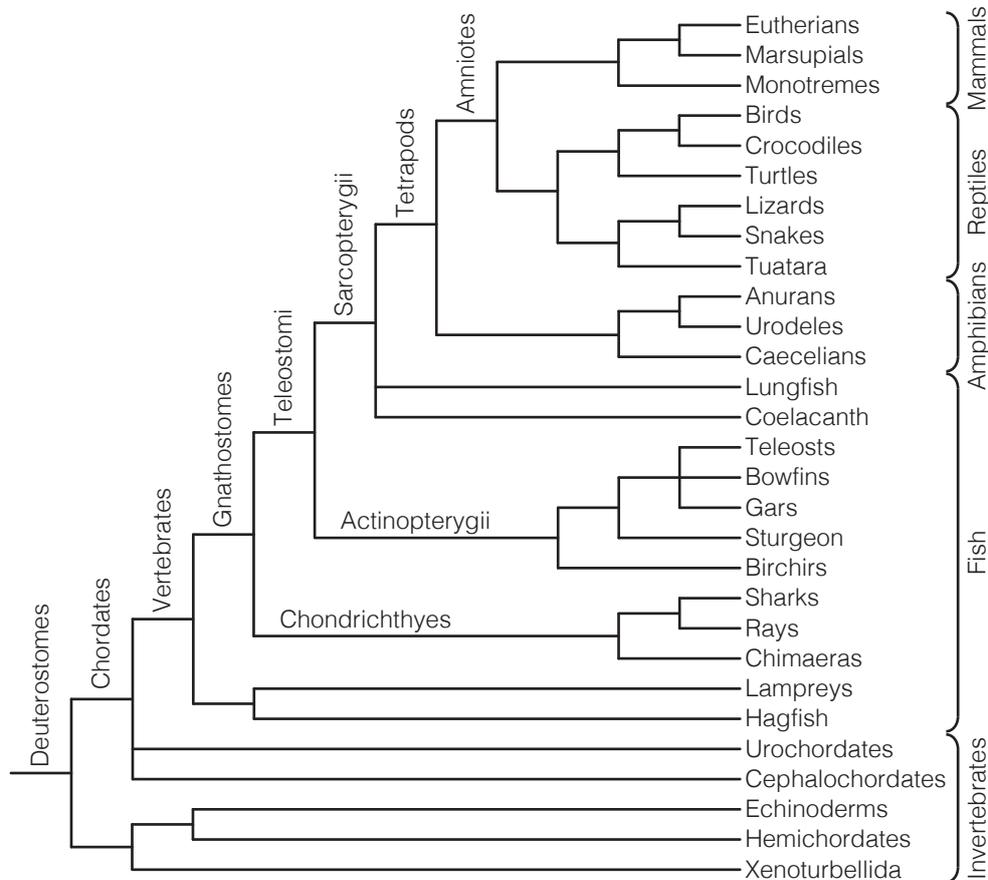


Figure 1.1: Deuterostome evolution. The relationship between the major clades of deuterostome species are shown, those relationships that are still to be resolved are shown as trifurcating branches. (See text for references for tree topology.)

The basal clades within deuterostomes are the chordates, echinoderms, hemichordates and Xenoturbellida (Minelli, 2008). Most studies show the echinoderms and hemichordates form a monophyletic clade, known as the Ambulacraria (Blair and Hedges, 2005; Peterson, 2004; Turbeville et al., 1994; Wada and Satoh, 1994). *Xenoturbella bocki* was initially placed within the protostome

bivalve molluscs (Norén and Jondelius, 1997) but has recently been reclassified as a close relation to the Ambulacraria group (Bourlat et al., 2006, 2003).

The chordates consist of three clades, the cephalochordates, urochordates and the vertebrates (Cowen, 2013). All three have a notochord during embryo development, the key characteristic of chordates. The cephalochordates retain the notochord throughout life and were originally thought to be the sister group to vertebrates (Adoutte et al., 2000; Cameron et al., 2000), but more recent studies have placed urochordates together with vertebrates (Blair and Hedges, 2005; Bourlat et al., 2006; Delsuc et al., 2006).

Vertebrates are comprised of hagfish, lampreys and gnathostomes (jawed vertebrates). Initial morphological and molecular data placed hagfish as the ancestral group (Maisey, 1986; Rasmussen et al., 1998; Suzuki et al., 1995). However, it is now established that hagfish and lampreys form a monophyletic group, known as Cyclostomata (Blair and Hedges, 2005; Delarbre et al., 2002; Heimberg et al., 2010; Oisi et al., 2013; Takezaki et al., 2003). Current evidence suggests that two rounds of whole genome duplication took place at the base of vertebrates, although their precise location relative to Lampreys and Hagfish is unknown (Dehal and Boore, 2005; Panopoulou et al., 2003; Putnam et al., 2008). Within the gnathostomes there are the cartilaginous fish, Chondrichthyes, and bony vertebrates, Teleostomi. The two major subclasses of Chondrichthyes are the Holocephali (chimeras) and Elasmobranchii (sharks and rays) (Inoue et al., 2010). The Teleostomi are comprised of the ray-finned fish, Actinopterygii, and the lobe-finned fish, Sarcopterygii.

The Actinopterygii consist of 5 major groups; birchirs, sturgeons, gars, bowfins and teleosts. Most studies concur that birchirs are the ancestral group and that sturgeons are basal to gars, bowfins and teleosts (Chiu et al., 2004; Hoegg et al., 2004; Hurley et al., 2007; Near et al., 2012). The placement of bowfins and gars in relation to the teleosts is poorly resolved but many studies group the bowfins and gars together (Hurley et al., 2007; Inoue et al., 2003; Near et al., 2012). A further whole genome duplication is known to have taken place at the base of teleosts (Hoegg et al., 2004; Jaillon et al., 2004; Taylor et al., 2003).

Within the Sarcopterygii are the lobe-finned fish groups, coelacanth and lungfish, as well as the land animals, tetrapods. The relationships between these three groups has been highly debated (Fritzsich, 1987; Hedges et al., 1993;

Liang et al., 2013; Shan and Gras, 2011; Zardoya and Meyer, 1996). However, the recent release of the coelacanth genome, as well as lungfish RNA-seq data, supports lungfish being the sister taxa to tetrapods (Amemiya et al., 2013).

Tetrapods are comprised of amphibians and amniotes, the amphibians are divided into three orders, caecilians, urodeles and anurans. Some studies have suggested that the amphibians are polyphyletic and that the caecilians group with amniotes (Anderson, 2008; Anderson et al., 2008). However, most phylogenetic studies suggest a monophyly, although there is still some doubt as to whether urodeles group with the caecilians (Feller and Hedges, 1998; Hedges et al., 1990), or anurans (Frost et al., 2006; Hillman et al., 2009; Pough et al., 2009; Pyron and Wiens, 2011; Trueb and Cloutier, 1991; Zardoya and Meyer, 2001; Zhang et al., 2005). The most common result is for an anuran-urodele group but with only mediocre support.

The amniotes are defined by the presence of the amniotic membrane which allows eggs to be laid away from water. They are composed of the mammals and sauropsids, the latter of which can be further subdivided into testudines, archosaurs and lepidosaurs. Although still highly debated, recent studies have concluded that turtles are the sister group of archosaurs, i.e. birds and crocodiles (Crawford et al., 2012; Iwabe et al., 2005; Kumazawa and Nishida, 1999; Laurin and Reisz, 1995; Rest et al., 2003; Shedlock et al., 2007; Wang et al., 2013). Within the lepidosaurs, the tuatara are a small clade consisting of only 2 extant species (Daugherty et al., 1990), it is well established that they are the basal lepidosaur (for review, see Evans, 2003). The remaining lepidosaurs, known as squamates, can be broadly classified into the lizards and snakes. It is thought that lizards are polyphyletic, and that snakes evolved secondarily (see Figure 1.9; Bergmann and Irschick, 2011; Fry et al., 2006; Lee, 2000).

The final group of deuterostome species are the mammals, these consist of monotremes, marsupials and eutherians. Although there were suggestions that monotremes were a highly derived group of marsupials, most morphological and molecular studies place the monotremes at the base of the mammalian phylogeny (Janke et al., 1997; Killian et al., 2001; Luo et al., 2001; Phillips and Penny, 2003).

Although there are still some points of contention within vertebrate evolution the majority of the taxa discussed above are well documented and their

relationships established. This information forms the backbone of our molecular evolutionary analyses as we expect it to reflect the underlying species phylogeny. It also provides the framework to extrapolate where and how often the mode of germ line specification has altered.

1.2 Germ line specification

To classify the mode of germ cell specification in each of the vertebrate divisions described above, it is important to understand how these two mechanisms function. Therefore, the formation of the earliest cells of the germ line, the primordial germ cells (PGCs), is described below for the vertebrate model species. In each of these species the PGCs are either specified by epigenesis or preformation. By knowing how these mechanisms differ and the key characteristics within them it will be possible to deduce the probable mode of PGC specification in non-model organisms.

1.2.1 Preformation

The key model species in which preformation occurs are the anuran *Xenopus laevis*, the bird *Gallus gallus* and the teleost fish *Danio rerio*. For each of these, the early embryology and PGC specification mechanisms are described. The common aspect is germ plasm, the maternally deposited determinants of PGCs.

Xenopus

The *Xenopus* oocyte is already patterned before fertilisation, with the pigmented animal pole and the yolky vegetal pole (for review, see Gilbert, 2006). Fertilisation occurs in the animal pole and the point of sperm entry will later on determine the dorsal-ventral axis. The first two mitotic cell divisions begin at the animal pole and extend to the vegetal pole. The third division is equatorial but is displaced towards the animal pole because of the concentrated yolk in the vegetal region. This divides the egg into four small blastomeres in the animal region and four large blastomeres at the vegetal pole. By the 16-64 cell stage the embryo is referred to as a morula and by the 128-cell stage the blastocoel is apparent and the embryo is classed as a blastula. At this stage the three germ layers, ectoderm, endoderm and mesoderm are specified. The animal and vegetal regions will develop into the ectoderm and endoderm respectively

due to the presence of localized maternal factors (Sinner et al., 2006; Zhang and King, 1996). The formation of the third germ layer, the mesoderm, occurs in the animal cells exposed to signals from the vegetal region (Slack et al., 1987). The dorsal region, marked by the site 180 degrees opposite to sperm entry, will form the organizer and signals originating here establish the dorsal-ventral axis (Figure 1.2). The dorsal endodermal cells, adjacent to the organizer, invaginate and the process of gastrulation begins.

In early *Xenopus* oocytes, mitochondria aggregate into a structure known as the mitochondrial cloud (Billett and Adam, 1976). As well as mitochondria the cloud contains vesicles of endoplasmic reticulum and germ plasm. The germ plasm is thought to aggregate into the mitochondrial cloud in early stage I oocytes (Zhou and King, 1996). The germ plasm persists in the vegetal region in small 'islands' through oocyte maturation, release and fertilization; after cortical rotation they begin to aggregate in the vegetal pole (Heasman et al., 1984; Savage and Danilchik, 1993). After the first cleavage they aggregate further along the cleavage membrane, before being divided into the four vegetal blastomeres (Whittington and Dixon, 1975). In the following divisions the germ plasm is segregated into only one daughter cell, keeping the number of presumptive PGCs (pPGCs) constant. By the blastula stage the pPGCs have moved away from the pole towards the blastocoel. During gastrulation they move into the embryo along with the endoderm into a position below the presumptive gut, it is at this time that they begin to divide symmetrically (Whittington and Dixon, 1975). From this point they move into the dorsal crest of the posterior endoderm. The PGCs then move into the lateral plate mesoderm of the dorsal mesentery and continue to migrate into the genital ridges (Wylie and Heasman, 1976).

Xenopus germ plasm contains RNAs such as *Nanos* (Mosquera et al., 1993), *Xpat* (Hudson and Woodland, 1998) and *Dazl* (Houston et al., 1998). *Nanos* is localised to germ plasm from stage I oocytes through to early stage embryos (Forristall et al., 1995; Zhou and King, 1996). *Nanos* is translated at late blastula stage though gastrulation as the germ plasm moves to a perinuclear position. The protein contains an RNA-binding motif and was proposed to act as a translational regulator during PGC proliferation (MacArthur et al., 1999). It

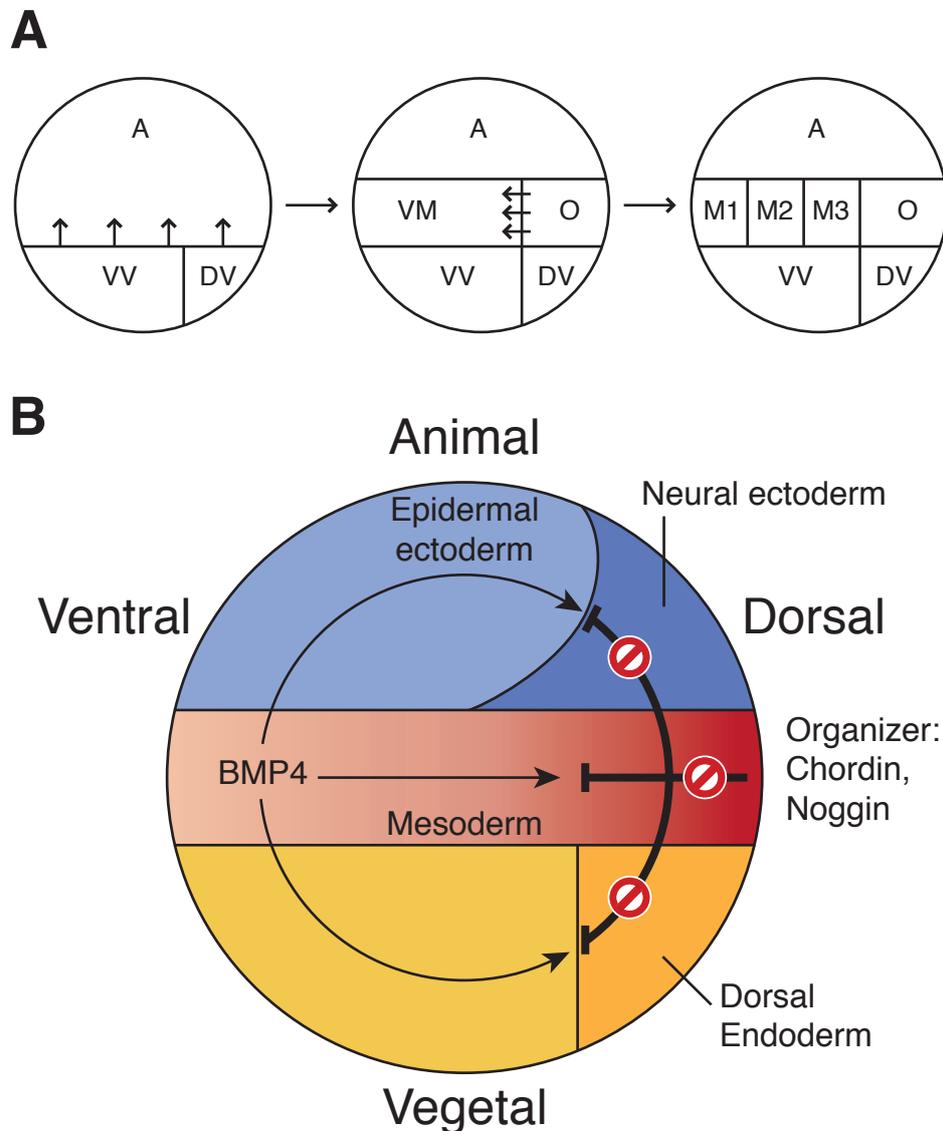


Figure 1.2: *Xenopus laevis* mesoderm induction. (A) The original 3-signal model of mesoderm induction (Slack et al., 1987; Smith and Slack, 1983), the first two signals come from the ventral vegetal (VV) and dorsal vegetal (DV) regions, these specify the ventral mesoderm (VM) and organizer (O) respectively. The third signal comes from the organizer and separates the mesoderm into three sections, muscle (M3), pronephros (M2) and ventral mesoderm (M1). (B) The organizer emits signalling molecules such as Chordin and Noggin which block the action of BMP4 being expressed from the ventral pole. This establishes a gradient across the ventral-dorsal axis. This gradient specifies cell fate in all three germ layers (Figure adapted from Gilbert, 2006).

has since been shown to repress the translation of *VegT*, an endodermal progenitor, and so preserves the germ line identity (Lai et al., 2012). *Xpat* is thought to be unique to anurans and to have a role in the organisation and positioning of germ plasm. Exogenous *Xpat* protein is capable of forming germ plasm-like 'islands' and of incorporating mitochondria (Machado et al., 2005). *Dazl* is localised to the germ plasm in oocytes and depletion of this maternal RNA leads to a near total loss of PGCs due to a failure in their migration towards the dorsal crest (Houston and King, 2000; Houston et al., 1998). The maintenance of *Xpat* and *Dazl* through the germ plasm and pPGCs means these genes are good markers for PGC specification and development in *Xenopus*.

These RNA molecules, along with others within the germ plasm, are involved in the repression of somatic signals, the maintenance and proliferation of PGCs, as well as their later migration and differentiation (Houston and King, 2000; Ikenishi and Tanaka, 1997; Lai et al., 2012; Venkatarama et al., 2010; Yamaguchi et al., 2013). When germ plasm is transplanted to an ectopic site, it is able to produce functional PGCs. Germ plasm is therefore sufficient for germ cell determination in *Xenopus* (Tada et al., 2012).

Zebrafish

Danio rerio development begins very differently to that of *Xenopus*, the cleavages only occur in the small region of non-yolky cytoplasm at the animal pole known as the blastodisc (Figure 1.3A). After 10 rounds of division the mound of cells, known as a blastoderm, can be distinguished as 3 different cell types. The most vegetal cells of the blastoderm fuse with the yolk cell to form the yolk syncytial layer (YSL). The remaining cells of the blastoderm form the one cell thick extra embryonic enveloping layer, and the deep cells that will form the embryo proper. These cell layers will then undergo gastrulation by migrating towards the vegetal pole, covering the yolk cell (for review, see Gilbert, 2006).

Germ plasm in zebrafish can be observed using markers for *Dazl* and *Nanos1*, as in *Xenopus*. It is also possible to use *Vasa*, which is localised in zebrafish but not *Xenopus* germ plasm (Ikenishi et al., 1996; Yoon et al., 1997). These three RNAs are localised to the mitochondrial cloud in stage I oocytes, however by stage II *Dazl* remains localised in the vegetal cortex while *Vasa* and *Nanos1* acquire a more distributed pattern around the cortex (Kosaka et al.,

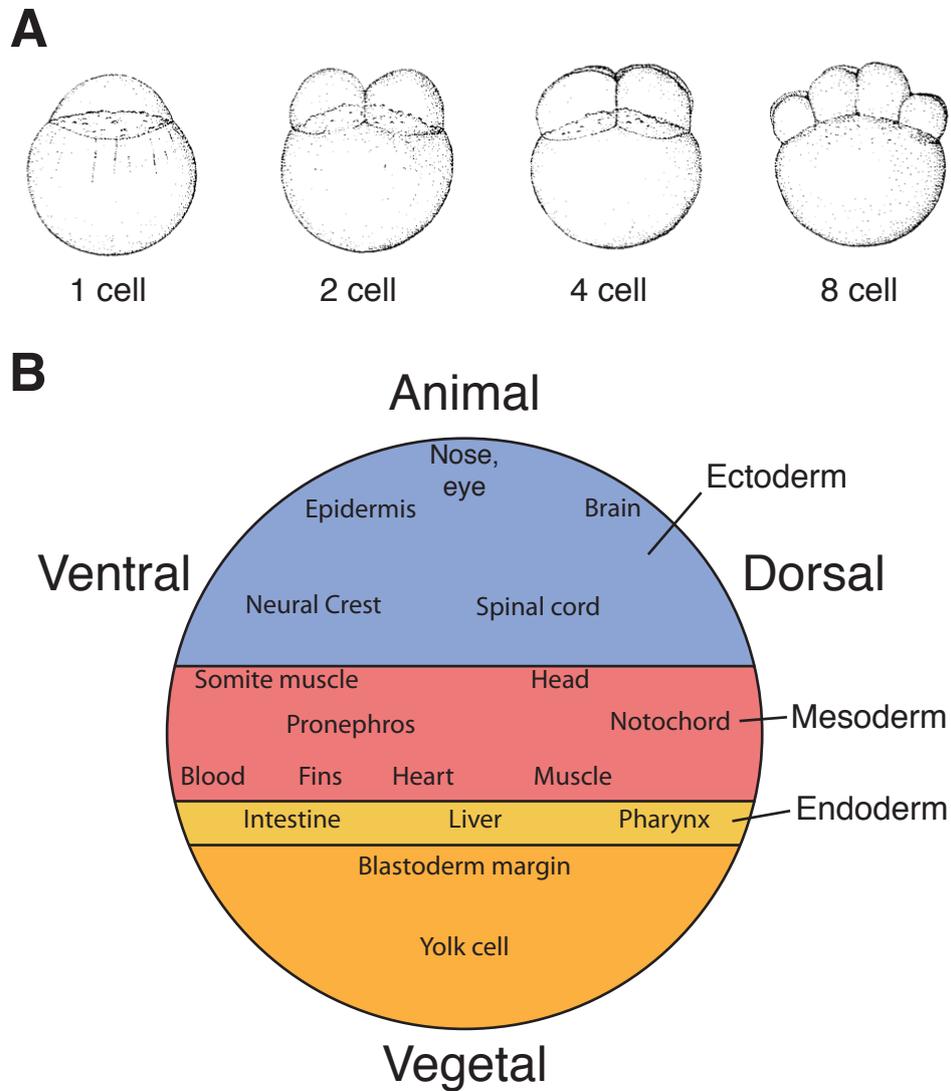


Figure 1.3: Zebrafish embryogenesis. (A) The first four stages of zebrafish development, the cleavages occur on the blastodisc, the protruding non-yolky cytoplasm. These meroblastic cleavages form a mound of cells on top of a large yolk cell (Images from Kimmel et al., 1995). (B) A fate map of the three germ layers and eventual tissues from the deep cells of a zebrafish blastula just before gastrulation; adapted from Gilbert, 2006.

2007). By the first cell division *Vasa* is localised between the yolk and cytoplasm compartments (Braat et al., 1999) and by the second cleavage it is localised along with *Nanos1* at the cleavage furrows (Kopranner et al., 2001; Yoon et al., 1997). Meanwhile, *Dazl* translocates from its position in the vegetal cortex towards the animal pole and becomes localised in an overlapping but distinct pattern to *Vasa* (Kosaka et al., 2007; Theusch et al., 2006). The germ plasm is then divided asymmetrically so that by the 1k-cell stage only 4 cells contain the germ plasm. Unlike in *Xenopus laevis*, these four cells are at opposite corners of the embryo and their location differs each time (Figure 1.4; Weidinger et al., 1999; Yoon et al., 1997). The PGCs migrate towards the dorsal pole, and form two populations either side of the notochord. These populations will then migrate posteriorly after 24 hours post-fertilisation into the developing gonad (Yoon et al., 1997).

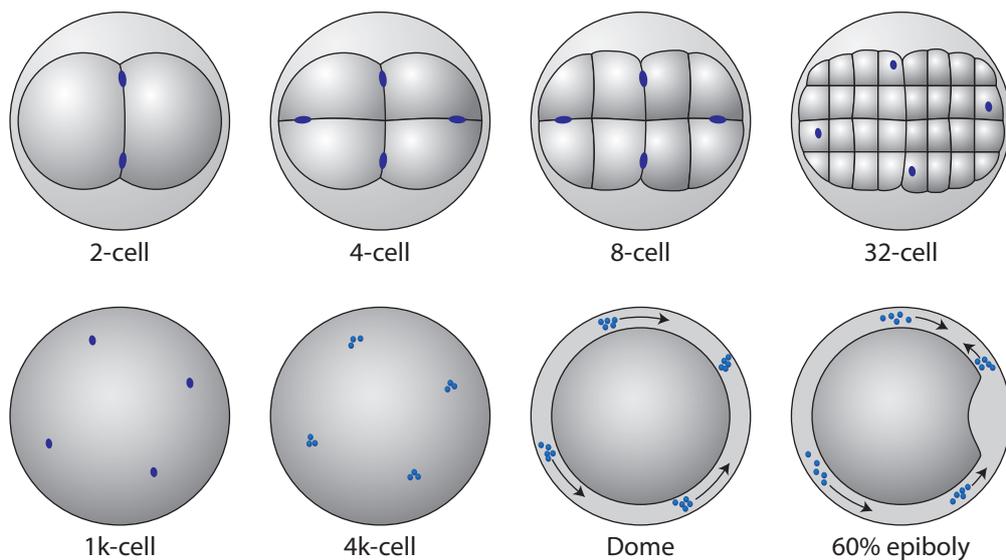


Figure 1.4: Zebrafish Germplasm. The germ plasm is shown in blue at the cleavage furrows during the first rounds of cell division. The cells that have inherited the germ plasm by the 4k-cell stage are the PGCs and they migrate towards the dorsal end of the embryo. From here they will form two populations either side of the notochord, after 24 hours they will make their final migration posteriorly into the developing gonad (Gilbert, 2006; Weidinger et al., 1999; Yoon et al., 1997).

When the germ plasm in the early cleavage furrows is ablated, the number of PGCs decreases (Hashimoto et al., 2004). This demonstrates that, like in *Xenopus*, the germ plasm is required for the development of PGCs in zebrafish.

Chicken

Chicken eggs contain an enormous amount of yolk relative to the size of the embryo. The early development of chicken embryos therefore resembles fish, with cleavage occurring on the blastodisc at the animal pole. The resulting blastoderm separates from the yolk, forming the subgerminal cavity. Cells within the centre of the blastoderm are then shed, leaving behind a one-cell thick area pellucida which will go on to form most of the embryo. The ring of surrounding cells is known as the area opaca, and the cells between them as the marginal zone. Thereafter cells from the area pellucida delaminate and migrate into the subgerminal cavity. These cells form islands of 5-20 cell clusters and will form the primary hypoblast. Soon afterwards, cells from the posterior margin of the blastoderm, known as Koller's sickle, will migrate anteriorly, pushing the primary hypoblast and forming the secondary hypoblast. The resulting blastoderm contains two cell layers, the epiblast and hypoblast, with a blastocoel between the two. The embryo will then undergo gastrulation by forming a primitive streak anterior to Koller's sickle (for review, see Gilbert, 2006).

The germ plasm in chickens can be observed in oocytes using *Vasa* staining; this shows localization to the mitochondrial cloud. As in fish, the germ plasm is localized to the cleavage furrows during the first rounds of cell division. It is then asymmetrically segregated until stage V (approximately 600 cells), when *Vasa* can be detected in the ventral cytoplasm of 6-8 cells which are located in the center of the blastodisc (Tsunekawa et al., 2000). These cells form a population at the centre of the area pellucida within the epiblast layer, they then migrate to a crescent-shaped zone in the hypoblast and begin to proliferate (Eyal-Giladi et al., 1981; Ginsburg and Eyal-Giladi, 1987; Kagami et al., 1997; Tsunekawa et al., 2000). The PGCs then migrate, through the bloodstream into the mesentery, and then finally into the genital ridges (Kuwana, 1993; Tsunekawa et al., 2000).

In all three model species the early development and localization of germ plasm differs. However in each case germ plasm is maternally deposited in the oocyte and asymmetrically divided into a few pPGCs which will later migrate and develop into the PGCs. It is the presence of germ plasm that determines

the specification of PGCs, these cells are therefore predetermined prior to fertilisation. This mechanism strongly differs from the epigenesis mode of PGC specification.

1.2.2 Epigenesis

The majority of information on the epigenesis mode of PGC specification has come from the mouse model species, *Mus musculus*. There has also been some work on humans embryos and the urodele, *Ambystoma mexicanum* (axolotl).

Mouse

Mouse development involves a number of different tissues only present during embryogenesis. At embryonic day (E) 4-4.25 the blastocyst is implanting, it contains the inner cell mass and primitive endoderm surrounded by the trophoctoderm. The inner cell mass, primitive endoderm and polar trophoctoderm then elongate and form the ectoplacental cone (which attaches the embryo to the uterus), the extra-embryonic ectoderm, the epiblast and visceral endoderm (Figure 1.5). This elongation establishes the proximal-distal axis. By E5.5 the distal region of the embryonic visceral endoderm will start to thicken, becoming known as the distal visceral endoderm (DVE). Towards E6.0 the DVE will migrate toward the anterior, establishing the anterior-posterior axis. It then becomes known as the anterior visceral endoderm (AVE) and specifies a signalling gradient across the epiblast. By E6.5 the primitive streak is formed at the posterior pole and the embryo will begin gastrulation. The proximal and posterior epiblast cells ingress through the primitive streak forming the extra-embryonic mesoderm, embryonic mesoderm and definitive endoderm. The remaining epiblast cells will form the ectoderm (for review, see Tam and Loebel, 2007), the whole process can be seen in Figure 1.5.

The PGCs are first identified as a population of cells that are scattered in the epiblast close to the extra-embryonic ectoderm before primitive streak formation at E6.0 (Lawson and Hage, 1994). They will then pass through the primitive streak early in gastrulation to lie within the extra-embryonic mesoderm (Chiquoine, 1954; Ginsburg et al., 1990). Clonal experiments showed that the PGCs are not lineage restricted until E7-7.5, when there are approximately 40

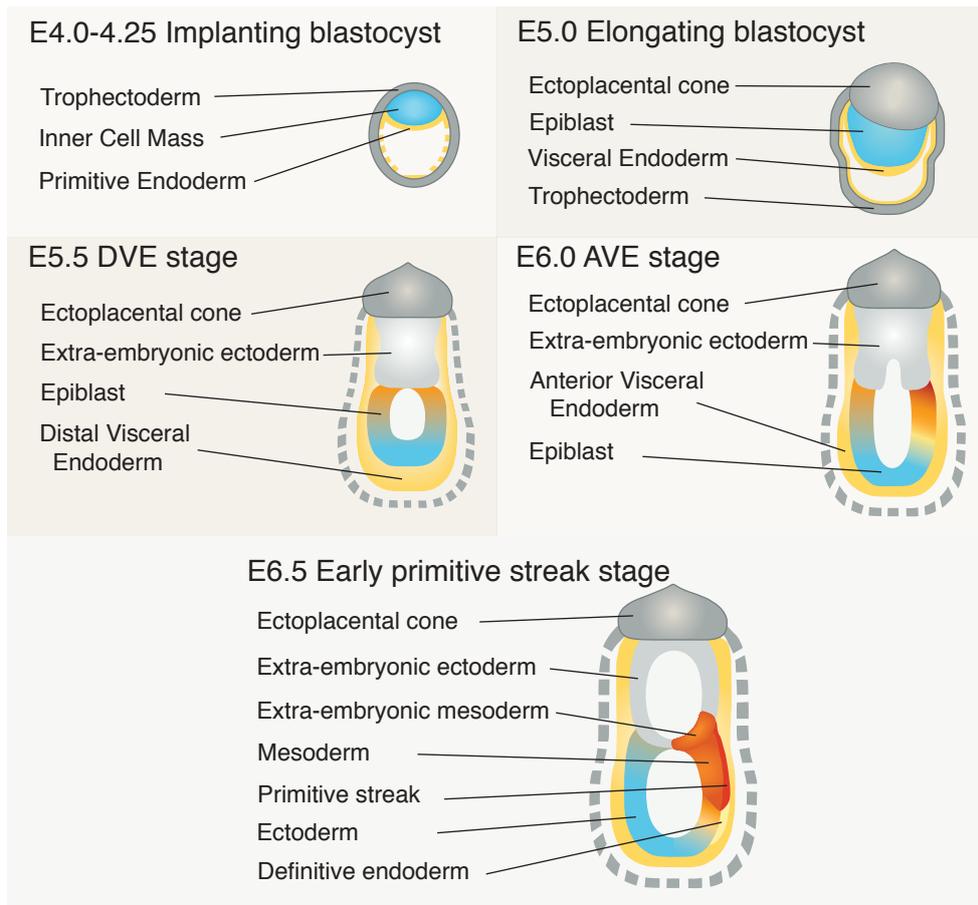


Figure 1.5: Early development in the mouse embryo. This figure shows the process of mouse embryogenesis from implantation to the earliest stage of gastrulation. It shows the formation of the three germ layers, ectoderm, mesoderm and endoderm as well as the large amount of extra-embryonic tissues which are unique to mammals. (Figure adapted from Tam and Loebel, 2007.)

cells (Lawson and Hage, 1994). Tam and Zhou (1996) showed that distally derived donor cells could be injected into the proximal epiblast and form PGCs. This demonstrates that the PGC fate is plastic and is dependent upon the local environment.

The signals required to form PGCs are emitted from the AVE, extra-embryonic ectoderm and epiblast. Bone morphogenic protein 4 (Bmp4) is required for PGC specification, it is expressed in the extra-embryonic ectoderm before gastrulation and in the extra-embryonic mesoderm in mid-to-late stage gastrulation (Lawson et al., 1999). It is not clear whether the Bmp4 signal is mediated by Alk2 in the visceral endoderm (De Sousa Lopes et al., 2004) or acts directly on the epiblast via Alk3 or Alk6 (Ohinata et al., 2009). What is known is that the dose-dependent Bmp4 signal will activate Prdm1 (also known as Blimp1) and Prdm14 in the posterior proximal epiblast (Ohinata et al., 2009). This requires Wnt3 expression for the epiblast to be susceptible to Bmp4 (Ohinata et al., 2009). Prdm1 acts as a suppressor of somatic genes and is essential for PGC development (Kurimoto et al., 2008; Ohinata et al., 2005; Vincent et al., 2005). Prdm1 is a target of the microRNA let-7 and so the let-7 repressor protein Lin28 is required for correct Prdm1 expression (West et al., 2009). Prdm14 activates pluripotency genes such as Sox2 (Yamaji et al., 2008). Prdm1 and Prdm14 are therefore key regulators of PGC specification in mouse. The area of Prdm1 and Prdm14 expression is limited by the inhibitory signals expressed in the AVE, such as Cer1, Dkk1 and Lefty1 (Lewis et al., 2008; Ohinata et al., 2009; Perea-Gomez et al., 2002; Tam and Loebel, 2007). This is in turn regulated by Bmp8b signals from the extra-embryonic ectoderm, confining the inhibitory signals to an appropriate level (Ohinata et al., 2009). A diagram of the processes required to induce the PGCs in mouse epiblast is shown in Figure 1.6.

By E7.25 the PGCs have passed through the primitive streak and are within the extra-embryonic mesoderm at the base of the allantois. The PGCs are now restricted to the germ cell lineage and have begun to express the germ cell specific gene *Dppa3*, also known as *Stella* or *Pgc7* (Saitou et al., 2002; Sato et al., 2002). Around E7.5, Prdm14 will repress *Glp* causing the global demethylation of H3K9me, beginning the process of epigenetic reprogramming (Seki et al., 2007; Yamaji et al., 2008). The expression of the pluripotency gene *Oct4* (*POU5F1*) will become restricted to the PGCs at this time (Kehler et al., 2004;

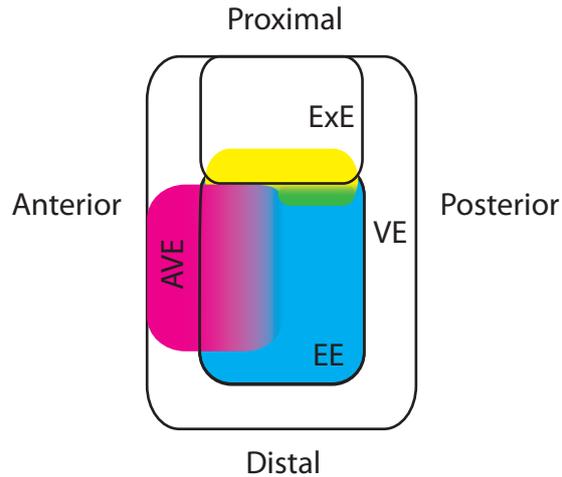


Figure 1.6: PGC induction in mouse. This figure shows a diagrammatic view of how PGCs are specified in a E6.25 mouse embryo. The blue area represents Wnt3 signalling in the epiblast, magenta the inhibitory signals emitted from the AVE and yellow the Bmp4 signal coming from the extra-embryonic ectoderm. The proximal, posterior epiblast with high Bmp4 signalling, primed by Wnt3 and without the inhibitors is where PGCs (green) are specified in mice. ExE, extra-embryonic ectoderm. VE, visceral endoderm. EE, epiblast. AVE, anterior visceral endoderm. (Figure adapted from Ohinata et al., 2009; for review, see Saitou and Yamaji, 2010.)

Schöler et al., 1990a; Yeom et al., 1996). From this point onwards the PGCs will migrate through the hindgut endoderm into the germinal ridges arriving around E10.5 (Chiquoine, 1954; Ginsburg et al., 1990; Lawson and Hage, 1994). The germ cell marker genes *Vasa* and *Dazl* are turned on at E11.5 whereupon the PGCs begin to develop into sex-specific germ cells (Fujiwara et al., 1994; Haston et al., 2009; Lin and Page, 2005; Ruggiu et al., 1997; Seligman and Page, 1998; Tanaka et al., 2000; Toyooka et al., 2000).

Humans

There is no information known on how the PGCs are specified in humans pre gastrulation, although it can be predicted that there will be some differences to mouse since humans lack any extra-embryonic ectoderm tissue (Gilbert, 2006). What is known, is that the PGCs can be first identified in the dorsal wall of the yolk sac near the developing allantois (reviewed by De Felici, 2013). This is roughly equivalent to the time and location of the mouse E7.25 *Dppa3* expressing PGCs. The PGCs will then migrate along nerve fibers from the dorsal

hindgut mesentery to the gonadal ridges (Møllgård et al., 2010). By this time they are expressing *Vasa*, as they are in mouse (Castrillon et al., 2000).

Axolotl

In the urodele model species, *Ambystoma mexicanum*, early development resembles that of *Xenopus*. There is a pigmented animal cap and signals from the vegetal pole induce the formation of mesoderm. There are however some key differences, such that PGC are also induced during mesoderm formation (Boterenbrood and Nieuwkoop, 1973). There are also differences within the oocyte; *VegT* one of the mesendoderm determinants is not localized in axolotl oocytes as it is in *Xenopus* (Nath and Elinson, 2007). There is also no localization of maternally expressed *Dazl* and *Vasa* (Bachvarova et al., 2004; Johnson et al., 2001).

Axolotl PGCs originate from the presumptive lateral plate mesoderm (Nieuwkoop, 1974), and are induced, along with somatic cells, by Fgf and Bmp4 signals (Chatfield et al., 2014). This signalling is regulated by the actions of Brachyury, Mix and Nodal (Chatfield et al., 2014; Swiers et al., 2010). The pPGCs are uncommitted until tailbud stages, as Mix over-expression and MAPK inhibition can cause them to differentiate into somatic cells (Chatfield et al., 2014). Since all other cell lineages are established during gastrulation in axolotls, the PGCs are the last cell lineage to be committed. This differs to the situation in mouse where *Prdm1* suppresses somatic genes earlier in development (Kurimoto et al., 2008; Ohinata et al., 2005; Vincent et al., 2005). Interestingly, *Prdm1* does not appear to have a role in PGC specification in axolotls (Chatfield et al., 2014). By late tail bud stage the PGCs can be positively identified in the dorsal edge of the lateral plate by the expression of *Dazl* and *Vasa* (Bachvarova et al., 2004; Johnson et al., 2001). The pluripotency gene *Oct4* is not expressed in PGCs until they begin developing into the germ cells (Bachvarova et al., 2004).

The unlocalized *Dazl* and *Vasa* in the axolotl oocyte suggests that this organism does not have germ plasm (Bachvarova et al., 2004; Johnson et al., 2001). The lack of determinants in the pPGC population is also demonstrated by the loss of PGCs when somatic signals are altered in the embryo (Chatfield et al.,

2014). Together these data show that PGCs in axolotl are specified by epigenesis.

1.3 PGC specification across vertebrates

To understand which PGC specification mechanism is utilised within non-model vertebrates, we can look for the molecular signatures as identified above. For example, localized *Dazl* and *Vasa* RNAs in the oocyte suggest the presence of germ plasm. This has been used to show that other species of teleost fish (Herpin et al., 2007; Peng et al., 2009) and anurans (Elinson et al., 2011; Marracci et al., 2011; Nath et al., 2005) undergo preformation. Conversely, the absence of *Dazl* and *Vasa* localization in oocytes suggests the organisms undergo epigenesis. This has been shown in another urodele species (Tamori et al., 2004), as well as a species of marsupial (Hickford et al., 2011), turtle (Bachvarova et al., 2009), sturgeon (Johnson et al., 2011) and lizard (Maurizii et al., 2009). Figure 1.7 shows the modes of PGC specification on the known vertebrate phylogeny according to the model species analysed in Section 1.2 and the data described above.

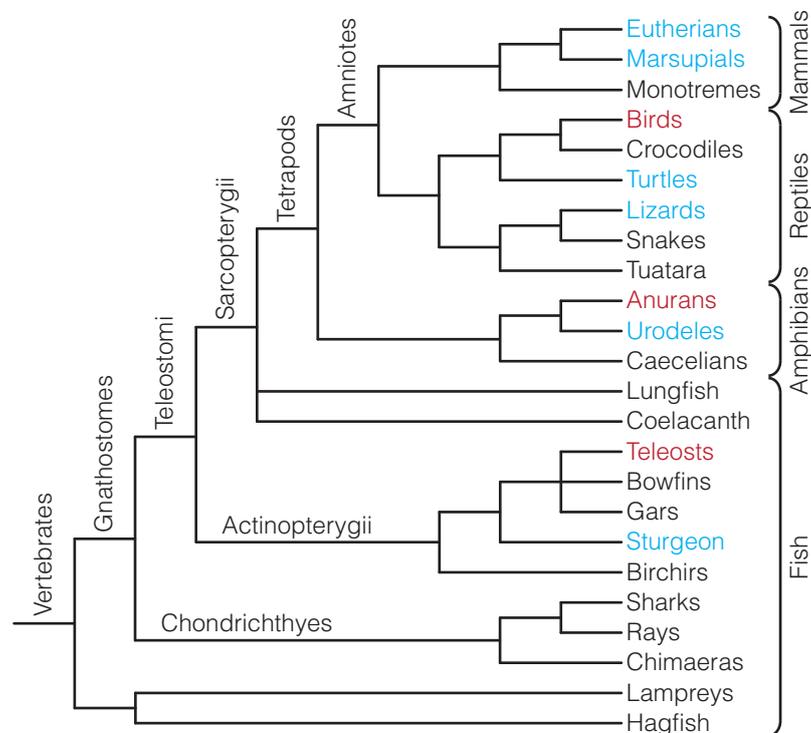


Figure 1.7: Modes of PGC specification across vertebrates. The species coloured in blue show evidence of epigenesis, those in red show evidence of preformation. The vertebrate phylogeny is taken from Section 1.1.

Recently, a paper claimed that sturgeons undergo preformation, based on teleost *Nanos* localising within sturgeon embryos (Saito et al., 2014). Notably, they did not use the endogenous RNA, and it is known that *Xenopus Dazl* and *Vasa* localises within axolotl embryos (Andrew Johnson, personal correspondence). This demonstrates that localisation of germ plasm markers is a characteristic of RNA from species utilising preformation and not the recipient embryo. We therefore continue to consider sturgeons as undergoing epigenesis.

Figure 1.7 clarifies that there are still many lineages where the mode of PGC specification has not been identified experimentally. It is however possible to infer what the likely mode of PGC specification is based on other developmental and morphological characteristics. Specifically, how these characteristics resemble their equivalent in species where the mode of PGC specification has been experimentally verified. One key example is in lungfish, where a mitochondrial cloud cannot be located: this resembles urodeles and not anurans, suggesting that lungfish undergo epigenesis (Johnson et al., 2003b).

As association between morphology and the mode of PGC specification has been observed (Johnson et al., 2003b). Figure 1.8 illustrates this within fish, all of the species with a derived morphology are assumed to be undergoing preformation, whereas the sturgeon and lungfish have primitive morphologies and are thought to be using epigenesis. Following this, it can be predicted that other taxa which have retained a primitive morphology such as sharks, coelacanth and crocodiles (Johnson et al., 2003a; Figure 1.8), have PGCs specified by epigenesis.

The situation within lepidosaurs (lizards, snakes and tuatara) is more complicated. There is still a lot of disagreement over the species phylogeny within the lepidosaurs and there is very little evidence for the mode of PGC specification. Figure 1.9 shows a consensus tree of the major lineages, there are still some trifurcating branches where conflicting results exist (Fry et al., 2006; Jones et al., 2013; Sites et al., 2011; Vidal and Hedges, 2009; Wiens et al., 2012). The main work on germ cell specification in lepidosaurs observed that Gekkonidae, Iguanidae and Lacertidae PGCs were located and migrated in a similar pattern to turtle PGCs. However in Scincidae, Chamaeleonidae, Agamidae, Crocodylidae, Anguillidae and Serpentes the location and migration resembled birds (Hubert, 1985). Since the molecular characterisation of *Vasa*, only one species has

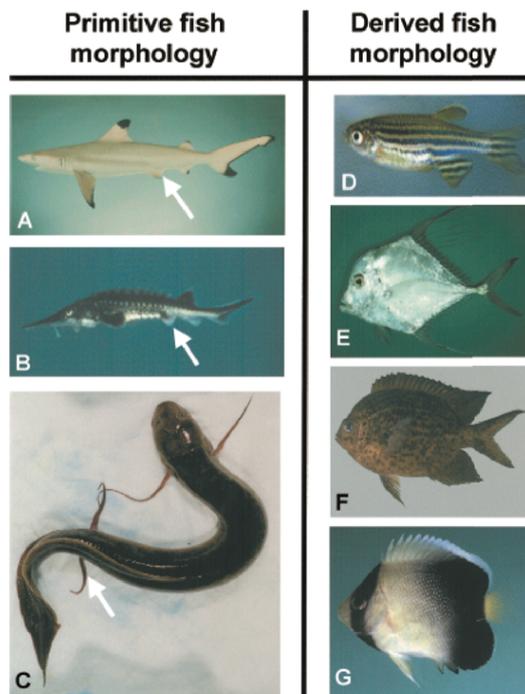


Figure 1.8: Morphology correlates with the mode of PGC specification. The left column shows fish with a primitive morphology, as they have retained pelvic appendages (arrows). The right column shows fish with a derived morphology with the absence of pelvic appendages. Those with a derived morphology are teleosts and so thought to be undergoing preformation, meanwhile the primitive morphology is associated with the epigenesis mechanism (see main text for references). The species shown are (A) black tip reef shark (*Carcharhinus melanopterus*), (B) Gulf sturgeon (*Acipenser oxyrinchus*), (C) African lungfish (*Protopterus annectans*), (D) zebrafish (*Danio rerio*), (E) Indian threadfish (*Alectis indicu*), (F) Damsel fish (*Acanthochromis sp.*) and (G) Cream Angelfish (*Apolemichthys xanthurus*). Figure from Johnson et al., 2003b.

been looked at, *Podarcis sicula*; this showed no localization of *Vasa* within the developing oocytes (Maurizii et al., 2009). *Podarcis sicula* is a member of the Lacertidae family, and therefore compliments the previous study. Looking at these family groups on a tree (Figure 1.9) shows that the species thought to be undergoing preformation do not form a monophyletic clade. This would either suggest that the mode of PGC specification has altered multiple times in lepidosaurs, or that the current information on topology or PGC specification is incorrect.

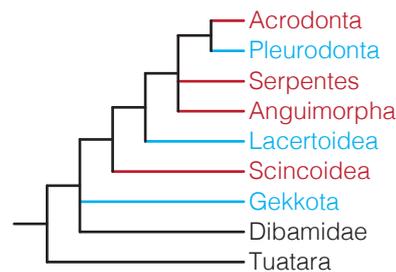


Figure 1.9: Evolution of lepidosaurs. The tree is a consensus phylogram of lepidosaur lineages (Fry et al., 2006; Jones et al., 2013; Sites et al., 2011; Vidal and Hedges, 2009; Wiens et al., 2012). The branches are coloured blue if that order is thought to be undergoing epigenesis, red for preformation (Hubert, 1985; Maurizii et al., 2009). The Acrodonta order is composed of the families Agamidae and Chamaeleonidae. The Pleurodonta order contains the Iguanidae family.

It is clear that epigenesis and preformation occur in a wide range of vertebrate species and that neither mode is used exclusively within a monophyletic clade. It is therefore important to know which is ancestral and which is derived, since the derived mechanism has evolved more than once.

The earliest studies on the mode of PGC specification occurred in the model species of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio* and *Xenopus tropicalis* all of which undergo preformation (Ikenishi, 1998; Illmensee and Mahowald, 1974; Strome and Wood, 1982; Section 1.2.1). It was therefore assumed that the lack of germ plasm in mice was a unique characteristic of mammals and that preformation was the ancestral mechanism (Saffman and Lasko, 1999). The earlier work on epigenesis in urodeles (Boterenbrood and Nieuwkoop, 1973) was largely ignored.

This assumption was questioned in 2003, when two independent studies concluded that epigenesis was the ancestral mechanism (Extavour and Akam,

2003; Johnson et al., 2003b). The first of these was a detailed study across vertebrates looking at many of the species described above (Johnson et al., 2003b). The correlation between preformation and a derived morphology; the differences in germ plasm localisation and segregation (Section 1.2.1); and the lack of germ plasm in basal species such as lungfish, all suggested that preformation had evolved independently multiple times within vertebrates. This proposal was later corroborated by a review of all literature on PGC specification in metazoa that demonstrated epigenesis was more widespread and frequently occurring than preformation (Extavour and Akam, 2003). It is now generally accepted that epigenesis is the ancestral mechanism of PGC specification (Gilbert, 2006; Saitou and Yamaji, 2012).

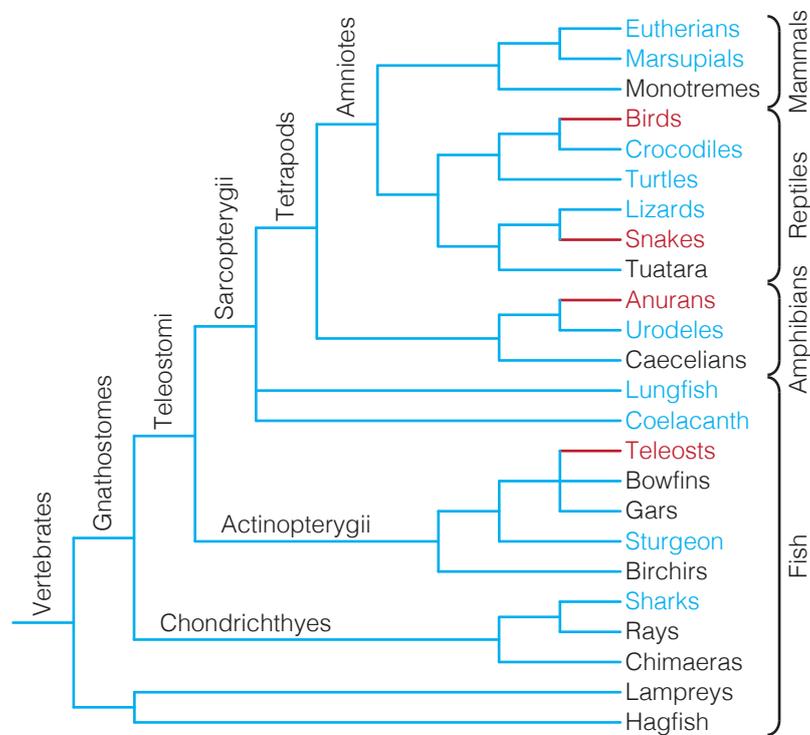


Figure 1.10: Epigenesis is the ancestral mechanism in vertebrates. Each lineage is coloured blue if they are undergoing epigenesis or red if they are utilising preformation. The branches for which we have no information are presumed to be utilising the ancestral mechanism, epigenesis, and as such are coloured blue.

In conjunction with evidence discussed previously, we can therefore predict that epigenesis is utilised in those species that are yet to be experimentally verified (Figure 1.10). This illustrates that preformation has evolved independently in birds, snakes, anurans and teleost fish. This suggests that the acquisition of preformation might provide an evolutionary advantage. Particularly since

all of these taxa converged on germ plasm as a mechanism to separate the somatic and germ cells at the inception of development. This advantage has been proposed to be a release from developmental constraint (Johnson et al., 2003b, 2011).

1.4 Developmental Constraint

Developmental constraint is a term used to describe a bias in the probability of certain developmental changes occurring during evolution. Constraint is a negative bias and so the opposite term is developmental drive, a positive bias towards specific developmental changes (Figure 1.11). The types of changes that can occur to developmental mechanisms can be classified into four groups (Arthur, 2011). For example a gene might be expressed earlier in development (heterochrony), in a different location (heterotopy), the level of expression could change (heterometry) or the expression might change to that of another gene (heterotypy).

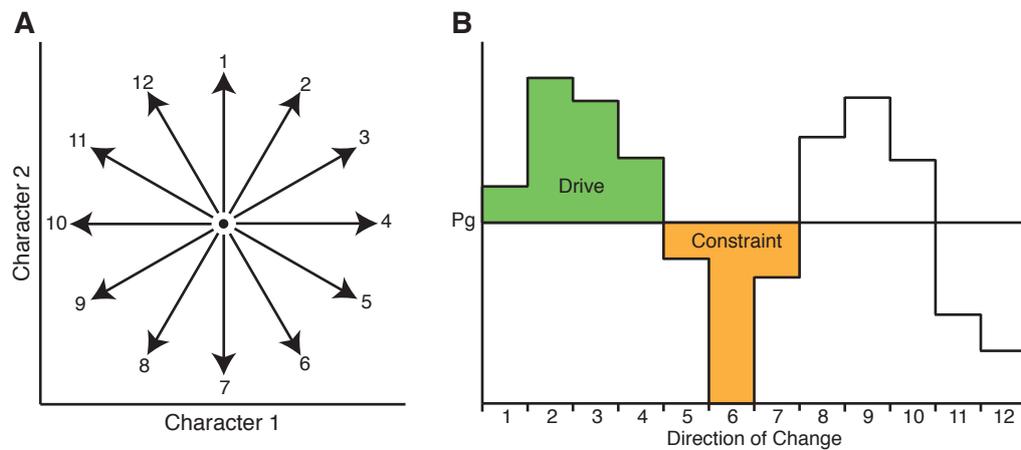


Figure 1.11: Developmental drive and constraint. (A) shows 12 directions in which the two character traits can change. (B) Each of those directions of change may have a different probability of occurring. The horizontal line represents equiprobability, or a lack of developmental bias, directions with a higher probability than that are said to be under developmental drive. Those with a smaller probability are under developmental constraint. Figure adapted from Arthur, 2001, 2011.

An example of constraint in developmental pathways can be observed in the neck vertebrae of mammals (Arthur, 2011). It can be assumed that natural selection towards a longer neck has occurred within both mammals and birds, and indeed there is a wide variety of neck length in both groups. However, the

two groups have acquired this variance by different methods. In birds many species have independently evolved a large number of neck vertebrae, for example swans have 23-25 and flamingos 18-20. However in mammals, almost all species have seven neck vertebrae (Galis, 1999). There is a range in size and length of the vertebrae, for example the giraffe, but the number barely changes. It can therefore be supposed that there is a developmental constraint within mammals that limits the number of vertebrae but not the size. Although this mechanism of constraint is unknown, there is evidence towards embryo fatality when cervical vertebrae are lost (Galis et al., 2006).

1.4.1 Developmental Constraint and the Germ Line

It has been proposed that the epigenesis mode of PGC specification imposes a developmental constraint that is released in species undergoing preformation (Johnson et al., 2003b, 2011). This theory suggests that constraint is enforced by the developmental pattern that induces the formation of PGCs. Any repatterning event that has a detrimental effect on the number or quality of PGCs will be under developmental constraint. The adult organism will have poorer quality or number of either sperm or egg cells (if any at all) and thus a lower evolutionary fitness.

An example of how this constraint might occur is shown in Figure 1.12. In this case a change in the developmental pattern of the mesoderm eliminates the PGCs in axolotl. However, in *Xenopus* which undergoes preformation, this same mesodermal change does not affect the germ line. This demonstrates how constraint might limit the potential for change in species undergoing epigenesis, and when released in species using preformation allows for an increase in evolvability.

It follows that if there has been a constraint release in one clade, but not in another, then we might expect to see a difference in the number of species (Crother et al., 2007). Indeed in classes with a known acquisition of germ plasm, there is an asymmetrical pattern (Figure 1.13). In archosaurs, amphibians and actinopterygians, the clade that has acquired preformation shows an increased rate of speciation. These data suggest a release of constraint in those species that have acquired preformation.

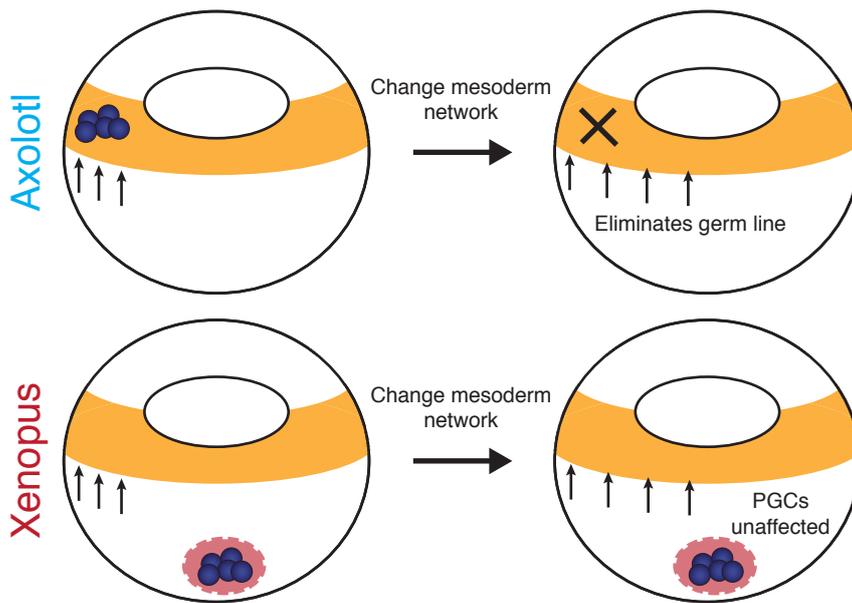


Figure 1.12: Developmental constraint and PGC specification. In this hypothetical situation a change in the mesoderm network occurs in both axolotl and Xenopus. In axolotl this change alters the inductive pathways of PGCs and so the germline is eliminated. However, the same change in Xenopus does not affect the germ line and so it is possible for the change to be inherited by the next generation. Figure adapted from Johnson et al., 2011.

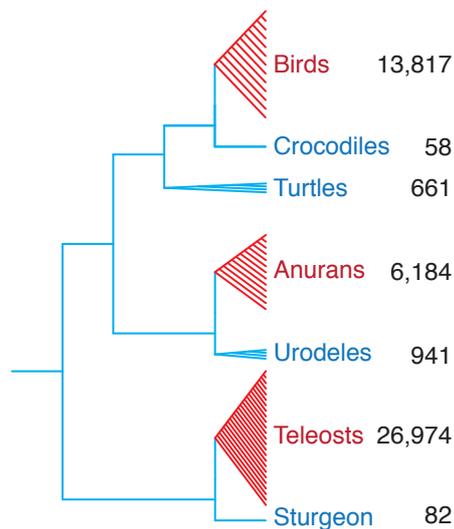


Figure 1.13: Speciation differs within clades undergoing preformation. The number of species are shown for the Archosauria, Amphibia and Actinopterygii. In all three cases the clade that has acquired preformation shows an increase in speciation. Number of species are approximate values from the NCBI taxonomy pages, downloaded 08/01/14.

Evidence towards this theory of constraint and constraint release has also been observed within the developmental networks themselves. For example the mesoderm network in axolotl is simpler than *Xenopus* and more closely resembles that of mouse, another species with the ancestral mechanism of PGC specification (Figure 1.14; Swiers et al., 2010). One particular difference is the mass duplication of genes in *Xenopus*; for example there are six Nodal genes, one of which has been extensively duplicated (Loose and Patient, 2004; Takahashi et al., 2006). There are also seven copies of the Mix gene in *Xenopus laevis* (Pereira et al., 2012). This compares to axolotl which has two Nodal genes and one Mix gene (Swiers et al., 2010) and mammals which have one copy of each (Robb et al., 2000; Zhou et al., 1993). This similarity between axolotls and mice is also seen in the pluripotency network, where the master regulator Nanog is missing in *Xenopus* but conserved between urodeles and mammals (Dixon et al., 2010; Hellsten et al., 2010).

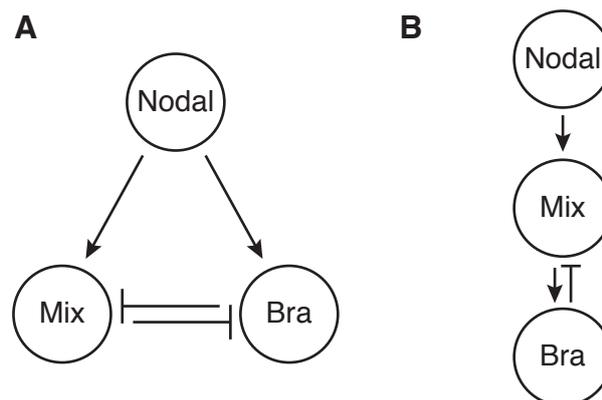


Figure 1.14: The mesoderm induction networks differ in *Xenopus* and axolotl. The interactions between Nodal, Mix and Brachyury differ in the mesoderm specification of *Xenopus* (A) and axolotl (B). The network in axolotls is conserved in mouse (Swiers et al., 2010). Figure adapted from Swiers et al., 2010.

Together these data suggest that there is a constraint on the developmental patterns in species undergoing epigenesis that has been released after the acquisition of germ plasm. However, whether evidence for constraint and constraint release is observable at a molecular level is unknown.

1.5 Genetic Evolution

Evolutionary relationships can be observed at the molecular level by comparing the DNA sequences between organisms. This is based on the logical assumption that the closer related species will share the most similar DNA sequences. To analyse these relationships between DNA sequences, phylogenetic trees are constructed.

1.5.1 Phylogenetic Trees

A phylogenetic tree is a graphical representation of the relationships between sequences. It consists of a root, branches, nodes and leaves. Nodes are the taxonomic units within a tree, they can either be leaf (terminal) nodes which represent the extant taxa or internal nodes which represent the hypothetical common ancestors. Branches connect these nodes together and the branching pattern is known as the topology. The length of each branch represents the evolutionary distance between two nodes. The root indicates the common ancestor of all nodes on the tree, it is often placed using an outgroup (a distantly related species).

There are many different methods of tree building but they can be classified into two groups; those that use distance matrices and those that use character data. The two better known distance based methods are UPGMA (Lemey et al., 2009) and Neighbor-joining (NJ; Saitou and Nei, 1987; Studier and Keppler, 1988). Character based methods include Maximum Parsimony (Farris, 1970; Fitch, 1971), Maximum Likelihood (Schmidt and von Haeseler, 2009) and Bayesian Inference (Ronquist et al., 2009). It is the latter two that we have used within this project.

The Maximum-Likelihood (ML) algorithm searches for a tree that maximises the probability of observing the character states, given the topology and evolutionary model. Each likelihood calculation involves summing over all possible character states in the internal nodes. The tree is then optimised to get the maximum probability using a combination of branch lengths and evolutionary models.

Bayesian inference (Ronquist et al., 2009) works quite differently, instead of searching for the best tree, this method instead searches for a probability distribution of trees. As such each internal node is given a confidence score based on this posterior probability distribution. The trees in this distribution are obtained using a Markov Chain Monte Carlo (MCMC) sampling technique. This is a step-by-step approach where from the initially parameters, each generation involves a small change which is either kept or rejected based on the likelihood and prior probabilities. After sufficient generations, the initial samples (burn-in) are discarded and the remaining trees are used to build a consensus tree.

There are many methods used to test how well the data supports the phylogenetic tree, including the likelihood tests discussed in Section 3.1.3. However, the most common method is to use a bootstrap analysis (Efron and Gong, 1983; Felsenstein, 1985). In this case a new alignment is created by randomly choosing columns from the original data. This new alignment is used to construct a tree and the process is repeated multiple times. Once all the required replicates have run, each internal node is scored based on the proportion of times it occurred within the bootstrap replicates. A node supported by less than 70% of the bootstrap replicates is considered unreliable (Lemey et al., 2009; Zharkikh and Li, 1992); this is approximately equivalent to a posterior probability of 0.8 (Erixon et al., 2003).

When building trees for individual genes, even when using the same species, each tree might have a different topology (Dávalos et al., 2012; Maddison, 1997; Rokas et al., 2003b). This conflict between gene trees is known as phylogenetic incongruence (Rokas et al., 2003b). This incongruence can be resolved by concatenating multiple alignments of individual genes, so long as the genes have a strong phylogenetic signal (Salichos and Rokas, 2013). Phylogenetic incongruence has been associated with incomplete lineage sorting, wherein the ancestral polymorphisms obscure the species phylogeny (Maddison and Knowles, 2006; Pamilo and Nei, 1988). However, this problem more commonly occurs between closely related species where the effective population sizes are relatively large compared to the time between divergences. Incongruence is also linked to heterogenous base composition, incorrect gene histories and limited taxon sampling (Foster and Hickey, 1999; Graybeal, 1998; Lockhart et al., 1994; Philippe et al., 2011; Rokas et al., 2003a; Sanderson and

Shaffer, 2002). Another well known cause of incongruence is divergent rates of sequence evolution (Bull et al., 1993; Chang, 1996; Philippe and Laurent, 1998).

If there is a significant increase in the rate of evolution within one or more branches of a tree it can affect the phylogeny by a mechanism known as 'Long Branch Attraction' (LBA; Anderson and Swofford, 2004; Felsenstein, 1978; Sanderson et al., 2000). Branches with a faster rate of evolution are likely to share more identical bases by chance than the inherited bases shared between true relatives. Therefore these long branches tend to cluster together, mostly at the base of the tree since the outgroup will inherently be a longer branch than the internal branches (Philippe and Laurent, 1998). It is therefore important to know whether there is rate heterogeneity in the dataset.

1.5.2 Rate of Molecular Evolution

Changes to DNA sequences accumulate over time, mostly due to errors during replication or exposure to environmental factors such as UV (Lemey et al., 2009). These mutations in protein-coding genes can be point substitutions, insertions or deletions. Point mutations that affect the encoded amino acid are known as non-synonymous, otherwise they are considered synonymous. The mutation rate is assumed to be relatively constant and can be calculated by the frequency of genetic disease (e.g. haemophilia) occurring spontaneously, or by comparing non-functional stretches of DNA (Haldane, 1949; Nachman and Crowell, 2000). The rate by which these mutations are maintained or lost within and between species is known as the rate of molecular evolution.

To calculate the rate of evolution, it is paramount to consider the effect of natural selection, i.e. whether a non-synonymous mutation is favourable (under positive selection) or deleterious (negative selection). The early models assumed that all mutations would affect fitness in either of these two ways (for review, see Bromham and Penny, 2003). However, the advent of the neutral theory suggested that the majority of inherited substitutions would have no influence on the fitness of an organism, they would therefore simply reflect the mutation rate (Kimura, 1968; Kimura and Ohta, 1971). This means that in general, genes evolve at a relatively constant rate, known as the molecular clock.

The molecular clock hypothesis does not hold true for all genes, nor indeed all sites within a gene (for review, see Bromham and Penny, 2003). Changes to

the strength of selection on particular sites, or whole genes (mutations that render a gene copy as non-functional, i.e. a pseudogene) will affect the rate of evolution. Population size is also known to affect the rate of evolution, with small populations showing greater rate of fixation of nearly-neutral alleles (Lanfear et al., 2013; Ohta, 1987; Ohta and Kimura, 1971). There can also be changes to the underlying mutation rate, particularly if there is a deleterious mutation in the repair enzymes or a change in environmental variables (Bromham and Penny, 2003).

Differences in rate of evolution between lineages can be attributed to any of the above alterations, and have also been linked to morphological characteristics such as body mass (Welch et al., 2008), generation time (Goetting-Minesky and Makova, 2006; Thomas et al., 2010), species longevity (Nabholz et al., 2008; Welch et al., 2008) and metabolic rate (Martin and Palumbi, 1993). These essentially alter the number of cell divisions within a time frame (generation time, and the associated changes to body mass and longevity), or alter the presence of oxygen radicals in the cell (metabolic rate) affecting the mutation rate directly.

To test whether there is a significant difference in the relative rate of evolution between two lineages, a third reference species is used to deduce the number of unique substitutions in each taxa. This is known as the Relative Rate Test (Kimura, 1980; Muse and Weir, 1992; Tajima, 1993). As when building phylogenetic trees, data characteristics such as compositional bias and limited taxon sampling can affect the results (Robinson et al., 1998; Tourasse and Li, 1999). To calculate the neutral rate of evolution (i.e. the mutation rate), sites with no influence on fitness must be selected, specifically four-fold degenerate sites where all possible changes to the DNA are synonymous (Britten, 1986).

An extreme change in the rate of evolution can lead to gene loss. Although an absent gene cannot be studied in the methods outlined above, it is important to note when these dramatic changes in gene conservation occurred. A large proportion of gene loss occurs after the duplication of a gene or whole genome (Brunet et al., 2006; Wolfe and Shields, 1997). When there are two redundant copies, one is likely to show an increase in the rate of evolution which can either lead to a novel function (neofunctionalization) or gene loss (nonfunctionalization) (Lynch and Conery, 2000; Zhang, 2003).

1.6 Hypothesis and Aims

The two modes of PGC specification, epigenesis and preformation, differ in how they segregate the germ cells from the somatic cells (Section 1.2). Preformation in vertebrates occurs through the maternally deposited germ plasm, which segregates these two cell types prior to fertilisation (Ikenishi, 1998). Epigenesis occurs later in development, when signals between embryonic cells induce the formation of the primordial germ cells (Ohinata et al., 2009). Preformation is the derived mechanism, and has evolved independently in birds, snakes, anurans and teleost fish (Section 1.3). This suggests that preformation provides an evolutionary advantage, and it has been proposed that this advantage is an increase in evolvability due to a release of developmental constraint (Section 1.4; Johnson et al., 2003b, 2011).

We have investigated whether vertebrate molecular evolution is associated with the mode of PGC specification. There are already known associations between preformation and changes to the developmental networks, as well as increased speciation (Section 1.4.1; Crother et al., 2007; Swiers et al., 2010). It therefore follows that differences between preformation and epigenesis might be observable at the molecular level, in either regulatory sequences, epigenetic marks or within protein coding genes. It is the last of these that we have studied.

Preliminary work in my masters thesis (Forey, 2010) showed that urodele sequences were more likely to have a mammalian top BLAST hit than anuran sequences were. We also showed that by building concatenated alignments, in almost all cases, the distance between urodeles and mammals was shorter than the distance between urodeles and anurans. However, these data were heavily flawed and included non-coding sequences as well as genes that were not true orthologs. It did however suggest that DNA sequences in urodeles and anurans showed very different patterns of conservation.

To investigate whether there is a known difference in urodele and anuran sequences, particularly in reference to how they compare to mammals, we reviewed the literature identifying 64 gene trees with the relevant taxa. Of the 56 unique genes, only 27 were able to consistently reflect the species phylogeny, grouping anurans and urodeles together (Appendix Table A.1, page

221). The majority of the remainder (20 gene trees) showed urodeles and mammals grouped together, an example of which is shown in Figure 1.15. Interestingly, very few of these incongruent trees were remarked upon and there was certainly no observation on the breadth of the bias.

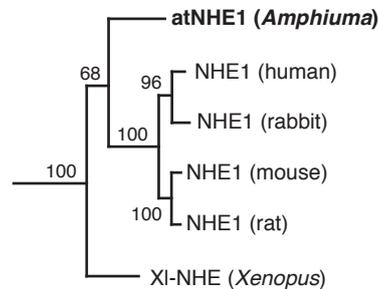


Figure 1.15: NHE1 gene topology. This gene shows the urodele *Amphiuma* grouping with mammals instead of anurans; figure is modified from McLean et al., 1999.

We have therefore investigated whether the mode of PGC specification associates with the phylogenetic incongruence observed above. We have done this using a global approach to build gene trees across amphibians, observing the patterns of incongruence. We have then investigated whether these patterns are unique to amphibians or whether they correlate with the mode of PGC specification throughout vertebrates. The causation of incongruence is then queried by looking for patterns of rate heterogeneity, and indeed whether this too correlates with the mode of PGC specification.

We have also analysed a few select genes in detail, particularly those involved in pluripotency. For these genes we have studied the phylogenetic relationships, rate of evolution as well as patterns of gene gain/loss.

Materials and Methods

The methods and programs utilised in Chapters 3-5 are freely available online at DataDryad (www.datadryad.org, doi:10.5061/dryad.rd70f) and our own website www.nottingham.ac.uk/~plzloose/phyloinc/. As well as all of the major scripts written during this project, there is also a step-by-step guide to recreating the data and results. The programs used to create the trees and synteny diagrams in Chapter 6 are not currently available to download.

2.1 Programming and data storage

All scripts were written in Perl (version 5.16.2) using modules such as BioPerl and Ensembl API tools (Flicek et al., 2014; Stajich et al., 2002). Custom scripts were used to pipeline and parse results from the programs listed in Table 2.1. Programs were written using the editing software Smultron and run on a 12 dual-core 32Gb Mac Pro 5.1.

Table 2.1: Programs used within the project

Program	Version	Reference(s)
BLAST+	2.2.24	Altschul et al., 1990
Cd-hit(-est)	4.0	Fu et al., 2012; Li and Godzik, 2006
CAP3	-	Huang and Madan, 1999
SeqClean	-	as used in Lee et al., 2005
MUSCLE	3.8.31	Edgar, 2004
Gblocks	0.91b	Castresana, 2000; Talavera and Castresana, 2007
PAUP*	4.0b10	Swofford, 2002
TreePuzzle	5.2	Schmidt et al., 2002
PhyML	3.0	Guindon et al., 2010
MrBayes	3.2.1	Ronquist et al., 2012
ModelTest	3.7	Posada and Crandall, 1998
ProfTest	3.2	Darriba et al., 2011
Consel	1.20	Shimodaira and Hasegawa, 2001
HyPhy	2.10b	Pond et al., 2005

Data were stored within a MySQL relational database, sequences were assigned an arbitrary number (seq_number) used as the primary key in all relevant tables. We stored as much information as possible from the various program outputs.

2.2 Sequence Library

During this project we created many sequence datasets, the first of which was comprised mostly of EST sequences. This dataset was expanded by adding our own transcriptomes, publically available transcriptome data and additional whole genomes as they became available.

2.2.1 ESTs, mRNAs and cDNAs

The initial dataset contained ESTs and mRNAs from NCBI and cDNAs from Ensembl. These sequences were downloaded for almost all available deuterostome species and processed to remove redundancy and low quality sequences. Each species was processed independently and the total number of sequences are shown in Table A.2 (Appendix, page 224).

The NCBI EST collection was downloaded (08/03/11) and divided by species. All deuterostome species with over 500 sequences were retained, except for the mammals where 7 species were selected representing eutherians, marsupials and monotremes (*Macaca nemestrina*, *Oryctolagus cuniculus*, *Rattus sp.*, *Isodon macrourus*, *Trichosurus vulpecula*, *Macropus eugenii* and *Ornithorhynchus anatinus*). The sequences were cleaned of vector using SeqClean and the NCBI UniVec-Core database (downloaded 08/03/11). SeqClean also removes poly-A/T tails, ends rich in 'N' and sequences with low complexity. However, we still found a limited number of vector contaminants and so used custom perl scripts to clean sequences using BLAST against the NCBI UniVec-Core database. This program removed regions with a BLAST alignment which had an e-value < 1e-10. After cleaning, all sequences less than 200bp were removed, leaving 10,897,805 EST sequences.

To remove redundancy we used cd-hit-est to cluster the sequences, and then ran CAP3 on each cluster. Upon completion we ran CAP3 on each species's concatenated outputs. The resultant 2,872,016 ESTs were blasted against our 'nr' database (see page 39), and any sequences unable to match a sequence from

another species with an e-value less than $1e-10$ were removed, leaving 1,608,276 ESTs. Within these ESTs there was still a degree of redundancy and therefore the top BLAST hit was used to cluster the sequences. From these clusters, regions able to locally align by BLAST were re-assembled with CAP3, thus ensuring that the process was not hampered by low quality regions. This process resulted in 949,257 non-redundant EST sequences.

The NCBI mRNA sequences were downloaded from the Entrez website for each of the species within our EST collection (723,718 sequences, 04/10/11). As before, sequences shorter than 200bp were removed. To reduce memory consumption in the various processing steps and exclude mitochondrial and other contaminating sequences an upper limit of 10,000bp was also imposed. Cd-hit-est was used to remove redundancy within each species. The Ensembl cDNA sequences were also downloaded for each species if available (355,965 sequences, 08/03/11) and processed by the same method as the mRNA sequences.

The EST, mRNA and cDNA sequences were combined for each species and processed using CAP3 and cd-hit-est to obtain the longest single reference sequence representing each transcript in our collection. The final dataset contained 1,344,819 sequences derived from 165 species across deuterostomes (Table A.2, Appendix page 224). This formed our 'query' dataset that was used throughout the project.

2.2.2 Transcriptome sequencing

The transcriptomes of *Ambystoma mexicanum* (axolotl), *Acipenser ruthenus* (sturgeon) and *Neoceratodus forsteri* (lungfish) were obtained by Illumina next generation sequencing. For the axolotl, total RNA was isolated from collagenase treated oocytes from a single adult female and from a range of early developmental stages from a single batch of axolotl embryos. For the sturgeon and lungfish, total RNA was isolated from whole ovary. RNA was checked for quantity and quality using the Nanodrop (Thermo-Fischer) and using the RNA Nano BioAnalyzer chip (Agilent). Libraries were prepared using the Illumina TruSeq RNA sample library preparation kit (Illumina, CA), samples were multiplexed to three of four samples and 12 pmol of pooled library was loaded per lane on a HiSeq 2000 (Illumina, CA). The resulting reads (approximately $1.5 \times$

10^{10} 76bp paired end reads for axolotl, 2.5×10^8 76bp paired end reads for sturgeon) were assembled over a range of word lengths using CLC Assembly Cell (CLC Bio). These assemblies were processed with cd-hit-est to retain the longest non-redundant transcript. The resulting assembled sequences were annotated by blasting against the vertebrate non-redundant protein collection and those sequences with an e-value greater than $1e-03$ were discarded. These steps were all performed by Matt Loose.

The *Ambystoma mexicanum* transcriptome required additional processing to remove redundancy between the query dataset and our new transcriptome sequences. Of the 2,493 EST based sequences for which we had found orthologs (see Section 2.3), 1,975 had longer sequences within the transcriptome and so were replaced.

2.2.3 Single Genomes for Mapping

For the single genome mapping (see Chapter 4) the Ensembl protein-coding transcripts and protein sequences were downloaded for *Mus musculus*, *Danio rerio* and *Xenopus tropicalis* (Table 2.2; Flicek et al., 2013, 2012). All known information on these genes was also downloaded from BioMart.

Table 2.2: Data downloaded from Ensembl.

Species	No. Transcripts	No. Genes	Date
<i>Mus musculus</i>	74,418	22,335	16/01/12
<i>Xenopus tropicalis</i>	22,075	18,429	23/04/13
<i>Danio rerio</i>	47,050	26,235	23/04/13

2.2.4 Additional Transcriptomes

In addition to our own novel transcriptomes we downloaded the following data (Table 2.3) from the NCBI Sequence Read Archive (SRA) database. Each species was assembled using CLC Assembly Cell and then blasted against the ‘prot-nr’ database. The sequences with a BLASTx result with an e-value $< 1e-03$ were retained.

Additionally the *Ambystoma mexicanum* transcriptome sequences available from www.axolomics.org (Stewart et al., 2013) were downloaded and assembled together with our own axolotl reads. This combined dataset was then processed in the same manner as the other transcriptomes.

Table 2.3: Transcriptomes downloaded and processed.

Species	Accession No.	Reference
<i>Notophthalmus viridescens</i>	SRP018244	Abdullayev et al., 2013
<i>Rana chensinensis</i>	SRP016636	Yang et al., 2012
<i>Rana kukunoris</i>	SRP016636	Yang et al., 2012
<i>Carlia rubrigularis</i>	SRP017492	Unpublished
<i>Lampropholis coggeri</i>	SRP017492	Unpublished
<i>Saproscincus basiliscus</i>	SRP017492	Unpublished
	ERR216304, ERR216306, ERR216315, ERR216316, ERR216322, ERR216325	Unpublished
<i>Eublepharis macularius</i>		
<i>Protopterus annectans</i>	SRP013624	Amemiya et al., 2013
<i>Leucoraja erinacea</i>	SRX036536	King et al., 2011
<i>Scyliorhinus canicula</i>	SRX036537	King et al., 2011
	SRR388692, SRR388693, SRR388694, SRR389308	Smith et al., 2013
<i>Petromyzon marinus</i>		

We also downloaded the known cDNAs from *Pelodiscus sinensis*, *Chrysemys picta bellii*, *Latimeria chalumnae* and *Lepisosteus oculatus* from Ensembl (Flicek et al., 2013). All were downloaded in May 2013, except for coelacanth (*Latimeria chalumnae*) which was downloaded in October 2013. Each species was run through the same process as the original cDNAs downloaded.

2.2.5 Whole Genomes

For analysing the synteny of genes, see Section 2.9 (page 53), we downloaded whole genomes from Ensembl (Table 2.4; Flicek et al., 2013). For each species the ‘top level’ sequences were downloaded, which contain the entire chromosomes and any unlocalised scaffolds. For mouse and human which contain multiple variant assemblies we downloaded the primary assembly files so each sequence was only represented once.

In addition to the Ensembl genomes we also downloaded the genomes from *Xenopus laevis* (v6.0 from Xenbase; Bowes et al., 2008; James-Zorn et al., 2013), *Leucoraja erinacea* (v1.0; King et al., 2011), *Python molurus bivittatus* (v5.0.2; Castoe et al., 2013), *Callorhinchus milii* (v6.1.3; Venkatesh et al., 2014) and *Branchiostoma floridae* (v1.0; Putnam et al., 2008).

Table 2.4: Whole Ensembl genomes downloaded

Species	Common name	Release
<i>Anolis carolinensis</i>	Anole Lizard	72
<i>Bos taurus</i>	Cow	73
<i>Canis familiaris</i>	Dog	73
<i>Chrysemys picta bellii</i>	Painted turtle	PreEnsembl
<i>Ciona savignyi</i>	Tunicate	73
<i>Danio rerio</i>	Zebrafish	72
<i>Drosophila melanogaster</i>	Fruitfly	72
<i>Equus caballus</i>	Horse	73
<i>Felis catus</i>	Cat	73
<i>Gallus gallus</i>	Chicken	74
<i>Gasterosteus aculeatus</i>	Stickleback	72
<i>Homo sapiens</i>	Human	72
<i>Latimeria chalumnae</i>	Coelacanth	72
<i>Lepisosteus oculatus</i>	Spotted Gar	74
<i>Loxodonta africana</i>	Elephant	73
<i>Macaca mulatta</i>	Macaque	73
<i>Macropus eugenii</i>	Wallaby	73
<i>Meleagris gallopavo</i>	Turkey	72
<i>Monodelphis domestica</i>	Opossum	72
<i>Mus musculus</i>	Mouse	72
<i>Oreochromis niloticus</i>	Tilapia	72
<i>Ornithorhynchus anatinus</i>	Platypus	72
<i>Oryctolagus cuniculus</i>	Rabbit	73
<i>Oryzias latipes</i>	Medaka	72
<i>Ovis aries</i>	Sheep	74
<i>Pan troglodytes</i>	Chimpanzee	73
<i>Pelodiscus sinensis</i>	Chinese softshell turtle	72
<i>Petromyzon marinus</i>	Lamprey	72
<i>Pongo abelii</i>	Orangutan	73
<i>Rattus norvegicus</i>	Rat	73
<i>Sarcophilus harrisii</i>	Tasmanian devil	72
<i>Strongylocentrotus purpuratus</i>	Sea Urchin	Genomes 20
<i>Sus scrofa</i>	Pig	72
<i>Taeniopygia guttata</i>	Zebra Finch	72
<i>Takifugu rubripes</i>	Fugu	72
<i>Xenopus tropicalis</i>	Xenopus	72

2.2.6 BLAST databases

All BLAST databases were created using the 'makeblastdb' program, part of the BLAST+ suite of tools, and indexed so that sequences could be retrieved using the 'blastdbcmd' program.

The NCBI Genbank vertebrate and invertebrate sequences (downloaded 23/02/2011) were filtered by length (>200bp). Sequences labelled as microsatellites or complete genomes were removed. For species with ≥ 20 sequences, cd-hit-est was used to remove redundancy. The resulting sequences were used to create the initial 'nr' BLAST database of 843,901 sequences.

The BLAST database created for the orthology finding consisted of the 'nr' database together with our query sequence dataset. Species represented in both datasets were combined using cd-hit-est. This database, the 'whole nucleotide database', was later separated into individual species and order (anuran, urodele etc.) BLAST databases. The query sequence dataset was also divided into species specific BLAST databases, as well as any additional transcriptomes and genomes.

To find the protein coding region of our query sequences we created a protein BLAST database, known as 'prot-nr'. This database consisted of the vertebrate and invertebrate Genbank protein sequences (downloaded 20/06/11) amalgamated using cd-hit on the entire collection. This database contained a total of 407,788 sequences.

2.3 Orthology finding

To test the evolution of individual genes it is vital to differentiate between gene copies created through a speciation event (orthologs) and those created through a duplication event (paralogs), see Figure 2.1. Only orthologs provide information on the relationship between species.

A small proportion of sequences appeared to be from an incorrectly labelled species, and so were removed from the query sequence dataset. These contaminants were identified by blasting against the 'whole nucleotide database' and having a result from a different taxonomic order with greater than 99% identity that matched to over 90% of the initial sequence. This removed 1,997 of our 1,344,819 starting sequences.

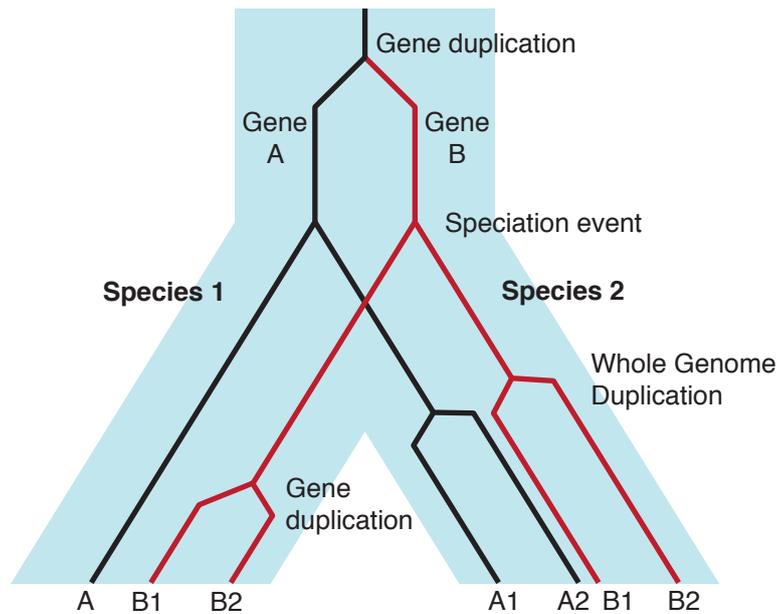


Figure 2.1: Orthologs and Paralogs. This figure shows the evolution of two genes. They originally divided from each other through a duplication event, meaning that they are paralogs to each other. In the following speciation event, species 1 retained both copies and duplicated gene B meanwhile species 2 underwent a whole genome duplication. This means that gene A is orthologous to genes A1 and A2, but that A1 and A2 are paralogs of one another. In the case of gene B, Spe1.B1 is orthologous to Spe2.B1 and Spe2.B2 but paralogous to Spe1.B2.

Reciprocal best BLAST Hit (RBH) methods were used to identify orthologs, this method is not as computationally intensive as tree building but is more error prone. However, most errors caused by RBH are false negatives, this is due to its propensity towards finding one-to-one relationships (Chen et al., 2007; Koonin, 2005). For example, in Figure 2.1 the many-to-many relationship of gene B means that in the initial BLAST from Species 1 to Species 2, if gene B1 matches B2 but then in the reciprocal BLAST back to Species 1 B2 matches B2 the gene will fail even though they are true orthologs (Figure 2.2). However, in situations such as this it is likely that one of the gene copies will have changed function (neofunctionalization), causing a change in the DNA sequence (for review, see Zhang, 2003). This would therefore increase the probability of getting a one-to-one reciprocal between the copies that retained a shared function. This is particularly important for those species with duplicated genomes such as teleost fish and the pseudo-tetraploid *Xenopus laevis* (Hellsten et al., 2010; Hoegg et al., 2004).



Figure 2.2: RBH identifies one-to-one orthologs. Gene B as shown in Figure 2.1, if species 1 is urodele and species 2 is anuran. There were two gene duplication events (red squares) leading to a many-to-many relationship between orthologs. RBH between these species might fail, if the original BLAST from urodele 1 matches anuran 2 but the reciprocal has urodele 2 as the top hit.

The biggest problem with reciprocal BLAST is when the datasets are incomplete, for example if species 1 only had gene A, but species 2 only had a copy of gene B then there would be no way to differentiate between orthologs and paralogs. To limit this error as much as possible, we require the orthologs to be RBH within at least three species. This increases the chance that one of the species will contain both copies of the paralogs, causing RBH to correctly fail. An overview of this method is shown in Figure 2.3.

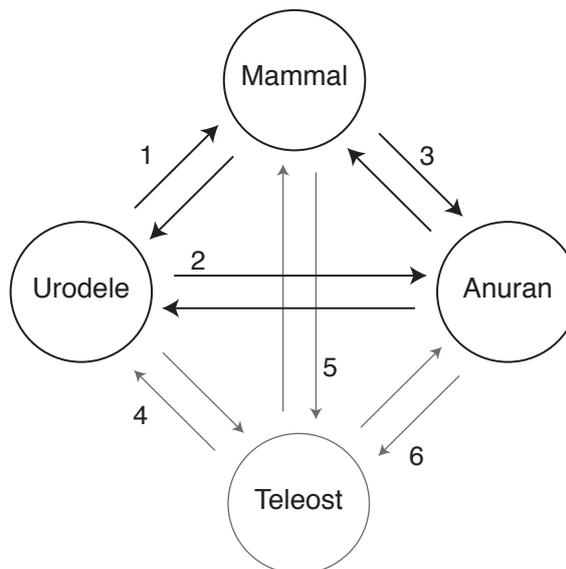


Figure 2.3: Reciprocal Best BLAST. Using Amphibians as an example, we required reciprocal best hits between a urodele, anuran and mammal. If this triplet of orthologs was identified we searched for other orthologs, in this case from teleost fish, each of which had to be the reciprocal best hit to each gene in the triplet. The reciprocal BLASTs are numbered in the order in which they occur.

Each sequence from our query dataset was first blasted against the mammalian database (Figure 2.3). This database contains multiple species with complete genome sequences, further limiting RBH false positives. The top hit was

then blasted back against the original species database, requiring this top hit to be the original sequence. This process was repeated against the sister taxon BLAST database. The reciprocal sister taxon gene was then blasted against the mammalian species identified previously, and again the reciprocal was performed. These last two BLASTs were species specific to control for potential species preference, for example if the two amphibian sequences were true orthologs but the urodele sequence matched human while the anuran sequence's top hit was from mouse. For the sequences which had all three required orthologs by RBH, the same methods were used to identify orthologs from other species and orders, for example a teleost ortholog (Figure 2.3). Each of these was required to have a one-to-one reciprocal to the query, sister taxon and mammalian sequences already identified.

After inspecting the results of this process we saw that there were more false negatives than necessary, most of which were caused by multiple BLAST results with equal bitscore and e-values. We therefore allowed all BLASTs, except for the initial ones against a new dataset (1, 2 and 4 in Figure 2.3), to look within the first 5 hits. The result was counted as a pass if the required sequence was within these results and had the same e-value and bitscore to the top hit. The BLAST parameters were also relaxed to encourage more, longer, alignments by reducing the gap opening penalty from 5 to 2.

When BLAST was limited to a specific e-value within the program parameters, it occasionally rejected sequences that actually passed this e-value threshold. This is caused by the sequences with an initial score less than the parameter being trashed before the alignments are refined and extended, and the score improved (Altschul et al., 1990). We therefore imposed no limit on e-value during the BLASTs and instead removed all sequences where the query-sister and query-mammal e-values were greater than $1e-10$ after the program had run. All sequences with a significant ($e\text{-value} < 1e-10$) BLAST hit to a mitochondrial genome were also removed.

Over 100 of the identified orthologs were individually tested by blasting them against the NCBI online non-redundant nucleotide database. In each case every sequence of the triplet matched the same genes, thereby suggesting our sequences were true orthologs with an error rate $< 1\%$.

One-to-One Orthologs

For any analyses where we compared just two databases against each other to find orthologs, for example with the single genome mapping, a single reciprocal BLAST was performed. For these BLASTs an e-value parameter of 0.001 was applied and there was no flexibility allowed for multiple hits sharing e-value and bitscore.

2.4 Locating the open reading frame

The majority of our sequences were derived from ESTs which are often low quality and frequently contain sequencing errors resulting in frame shifts. We therefore created a program which would locate the open reading frame (ORF) whilst also correcting for frame shifts. This used the top BLASTx result against the 'prot-nr' database, locating frame shifts using the local alignment regions (high-scoring segment pairs; HSPs) and multiple alignments; as shown in Figure 2.4.

If the BLASTx result contained multiple HSPs in different frames (on the same strand) then the HSPs were ordered according to their starting position. Each HSP was then analysed in turn, based on how it related to the following HSP. If there was no overlap between them then each HSP was retained separately. However, if there was an overlap less than 6bp the entire overlap was removed, retaining the frame (Figure 2.5).

For HSPs which overlapped by more than 6bp we used the information from building multiple alignments for the raw DNA sequences, see Section 2.5.1 (page 46). If the overlap position was within the alignment, then we searched for a single insert/deletion (in/del). If such a position existed then we used that location to trim each HSP, retaining the frame as in Figure 2.5. However if the alignment was more complicated, or did not show any frame changes then we removed the entire overlap region.

In a minority of cases, a short HSP was completely overlapped by a longer HSP (Figure 2.6). This was due to two frame shifts close to each other, which BLASTx was able to align in the wrong frame. To deal with these situations we tried to divide the longer HSP into two regions either side of the short HSP. We used a sliding window of 10bp to analyse the percent identity within the

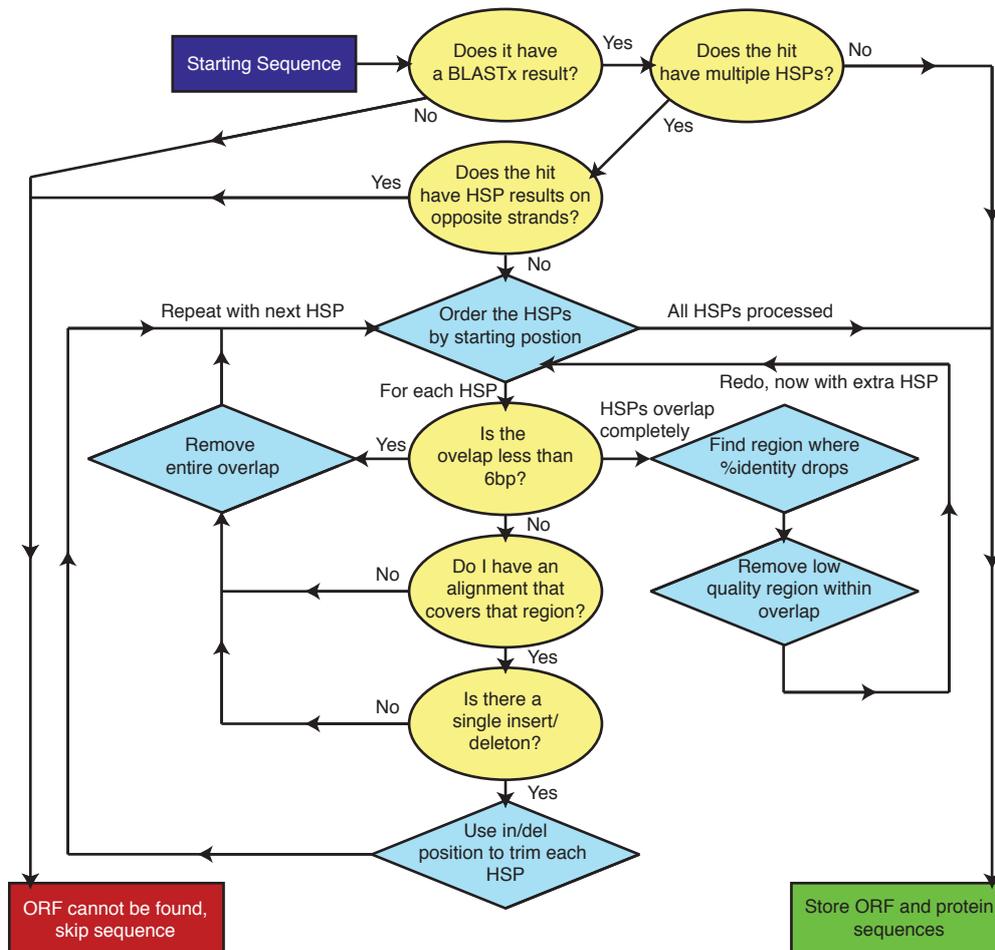


Figure 2.4: Flowchart of final ORF finder program. The diagram outlines the process of finding the ORF using the top BLASTx hit; multiple alignment information is used to help locate the frame shift.



Figure 2.5: Removing an overlap. In this example the HSP to the left is in the second frame, the one to the right is in the first frame. The sequences retained are indicated by the arrows.

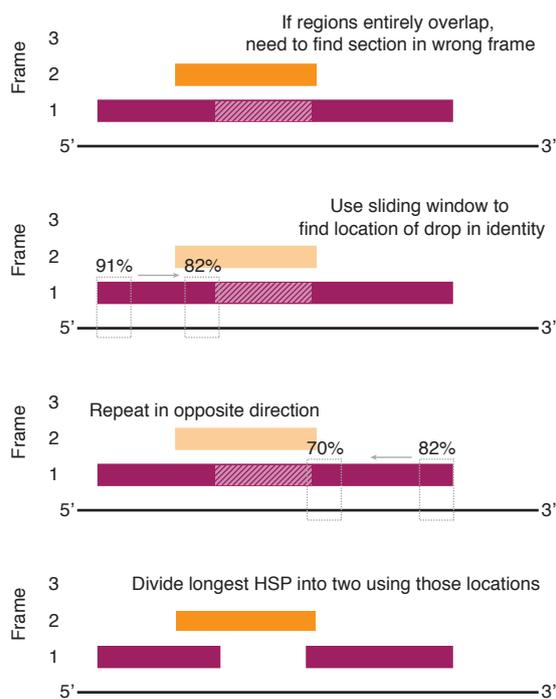


Figure 2.6: Completely overlapping HSPs. When one HSP sits entirely within another a sliding window is used to find the positions where the identity drops. This region is then removed from the longer HSP, dividing it into two.

longest HSP. If the starting percent identity was greater than 90% we observed the position where it dropped by 8%. However, if the starting percent identity was greater than 80% the program looked for a drop of 10%. This process was carried out in both directions and the region between the markers was removed. If the starting percent identity was less than 80% or no significant drop in identity was located than the entire region overlapping with the shorter HSP was removed. This provided us with three HSPs instead of two, which could be processed as before.

This program removes all ambiguous codons and can therefore be used to translate ESTs with high confidence. However, it inserts gaps within the sequence, and regions that are poorly conserved are unlikely to be retained. To confirm that our protein sequences were valid, we blasted the protein sequence of each ortholog against our protein query sequence. Those where the mammal and sister taxon sequences had an e-value greater than $1e-10$ were not analysed any further.

2.5 Four-taxon Phylogenies

We built 4-taxon phylogenies as part of the global analyses in Chapters 3 and 5. Each one consisted of a query, sister taxon, mammal and outgroup sequence, as identified in Section 2.3.

2.5.1 Alignments

We employed two different methods to create multiple alignments, the first used the original cDNA sequences. We then developed a new method which was used for the protein-coding regions. These same methods were used for building the 3-taxon alignments used within the relative rate test (Section 2.7).

cDNA alignment

When aligning unprocessed cDNAs we discovered that it was important to remove regions that are incapable of aligning prior to running a multiple alignment program. To remove these regions and to ensure the sequences were in the same orientation, the information from the original RBH BLAST against the query sequence was used. Each sequence was orientated according to the query sequence, and trimmed to the HSP. The query sequence itself was trimmed to the longest region able to align to all orthologs (Figure 2.7A). The decrease in BLAST gap penalty during the RBH (Section 2.3) lengthened these alignments.

However, when constructing the alignment using the sister taxon as the starting point, even if the same mammal and outgroup sequences were used, the alignment differed. Therefore, the sister taxon was blasted against the remaining orthologs. This information (Figure 2.7B) was used to locate the maximum regions able to align for each ortholog against the query and sister sequences. The sister taxon sequence was trimmed in the same manner as the query. In many cases this led to a greater alignment length, as well as ensuring that the alignment was independent of the starting species (Figure 2.7C).

Once the sequences were trimmed to alignable regions and in the same orientation, the program MUSCLE built the multiple alignment. The alignment was then passed to the program Gblocks to remove gaps and poor quality regions. The minimum length of a block was set to 20bp, ensuring that our alignment was of high quality with no ambiguously aligned regions.

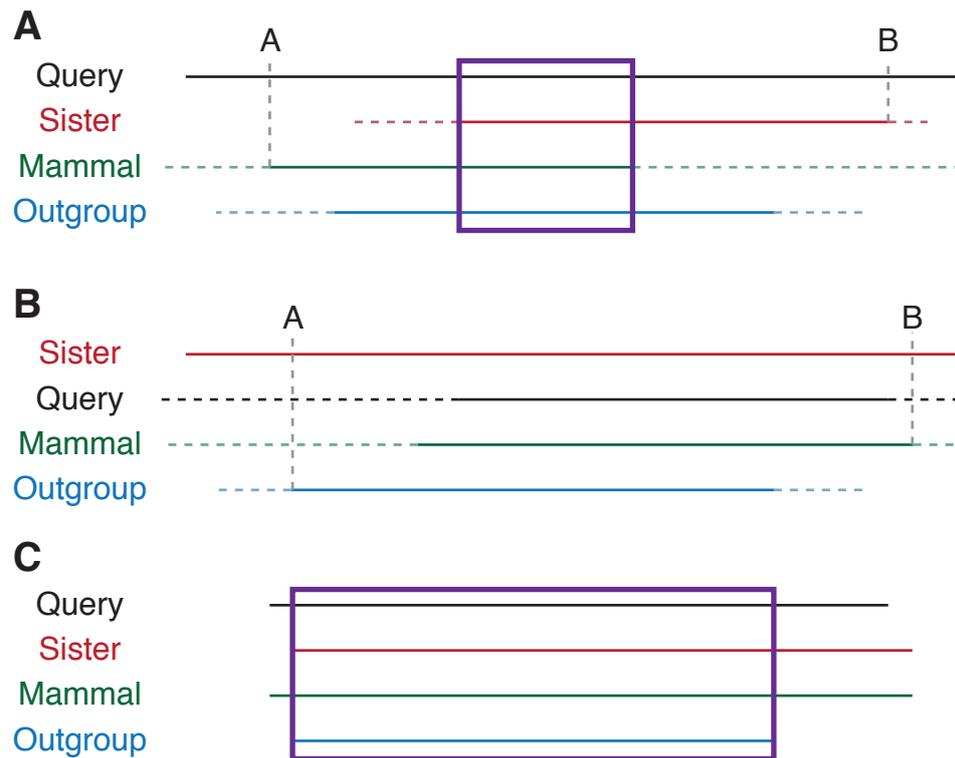


Figure 2.7: Utilising the BLAST alignments to trim the orthologs. (A) The BLASTs for each ortholog against the query sequence, the dashed lines show the overhang regions unable to align and positions A and B show the maximum region of the query sequence. The original region used in the multiple alignment is shown by the purple box. (B) The BLAST results from each sequence against the sister taxon. (C) By using the information in (A) and (B) the sequences can be trimmed to the positions shown, this increases the multiple alignment.

Protein based alignment

When using the protein coding sequences identified in Section 2.4 (page 43) the sequences had already undergone a large degree of trimming to the regions able to match a protein sequence. We therefore deemed it unnecessary to trim them any further.

The protein sequences were obtained for all orthologs and any stop positions were substituted for an 'X'. This was important as MUSCLE removes all stop positions from the sequences before alignment, which makes it difficult to recreate the alignment using DNA sequences. After aligning the sequences using MUSCLE, all alignment positions which contained an 'X' were removed, thereby removing all stop positions and unknown bases from all sequences evenly (Figure 2.8).

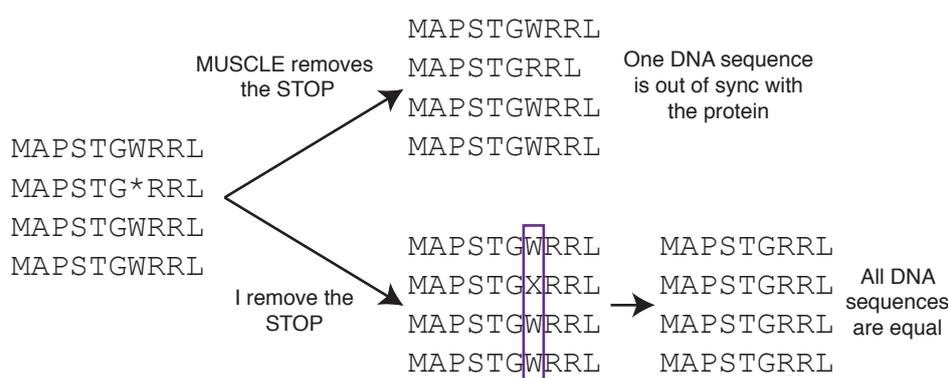


Figure 2.8: STOP positions in MUSCLE. By default any STOP positions entering MUSCLE are removed before the alignment, leaving a discrepancy between the protein alignment and DNA sequence. This was solved by changing all STOPs to an 'X' and then removing the position from the whole alignment, thereby affecting all DNA sequences equivalently.

After removing these positions the alignment was ran through Gblocks, setting the maximum length of a block to 20 amino acids and the minimum number of conserved sequences for a flank position to 4 (or 3 if building a 3 sequence alignment). This high level of stringency meant that any small regions surrounded by gaps were excluded and that all of the sequences had to be conserved in the regions flanking gaps. This created our final protein alignment.

To align the coding DNA to this protein alignment we first went through the original MUSCLE alignment identifying all positions that contained an 'X'. We also looked through the Gblocks output files to find the location of all blocks.

Using these data we recreated the alignment using the DNA sequences, removing the codons in an 'X' position as well as those outside of the Gblocks regions. This alignment of all codon positions was either retained as is, or further modified to exclude the third codon position, or the first two codon positions.

2.5.2 Building Trees

We developed two methods to build 4-taxon trees, depending on whether the alignment contained DNA or protein sequences. Both methods were designed to be computationally un-intensive and yet still build the most accurate phylogeny possible.

DNA Trees

To build DNA 4-taxon trees we began by selecting the best DNA substitution model for the data using the program ModelTest. This uses PAUP* to find the likelihood of each model and then tests them using Akaike Information Criterion (AIC; Akaike, 1974). The model with the best score was then used by PAUP* to build a Maximum Likelihood (ML) tree and to run 1000 bootstrap replicates. This same model was then used in TreePuzzle to test the three possible topologies using the SH (Shimodaira and Hasegawa, 1999) and KH (Kishino and Hasegawa, 1989) likelihood tests; this program also calculates the estimated likelihood weights (ELW; Strimmer and Rambaut, 2002). Finally, the program Consel was run, using the output from TreePuzzle, which performed the Approximately Unbiased likelihood test (AU; Shimodaira, 2002).

Protein Trees

To build the protein trees we performed the same tests as above but used different programs since some, particularly PAUP*, are DNA specific. We selected the best model using ProtTest, limiting the choice to the Dayhoff, Blosum62, JonesTT, LG and WAG models, with or without gamma correction (Dayhoff et al., 1978; Henikoff and Henikoff, 1992; Jones et al., 1992; Le and Gascuel, 2008; Whelan and Goldman, 2001). ProtTest tests the models according to the Bayesian Information Criterion (BIC) as well as AIC (Darriba et al., 2011; Schwarz, 1978). We selected the model with the highest score in both tests. We

then ran PhyML to estimate the proportion of invariant sites and gamma parameters on the ML tree. After fixing these parameters we use PhyML to run 100 bootstrap replicates, this number was significantly smaller than when using DNA sequences because of the increased time to compute. As before, we used TreePuzzle and Consel to run the likelihood tests on all three topologies.

2.6 Whole Gene Phylogenies

To create a complete phylogenetic tree (as in Chapter 6) we first needed to build a reliable alignment with all available orthologs and paralogs. We therefore searched the NCBI entrez database for all sequences with the name or synonym of the gene of interest in the title. We downloaded all of these sequences in genbank format so that we could select the protein-coding sequence. In addition to this we selected the gene of interest in either mouse or human in Ensembl and using the perl API, downloaded all known orthologs and paralogs (and paralogs of orthologs) from the entire EnsemblGenomes collection. For each of these genes we selected the coding sequence from the longest transcript. We searched our own MySQL databases for the query sequences that had been mapped to the mouse gene (as in Chapter 4). We also blasted the mouse gene of interest against various assembled transcriptomes (*Ambystoma mexicanum*, *Notophthalmus viridescens*, *Rana chensinensis*, *Rana kukunoris*, *Neoceratodus forsteri*, *Protopterus annectans*, *Acipenser ruthenus*, *Scyliorhinus canicula* and *Leucoraja erinacea*); these sequences were then blasted against a protein database to identify the ORF. We also ran this process on the known Ensembl cDNAs of *Pelodiscus sinensis*, *Chrysemys picta bellii* and *Lepisosteus oculatus*. Finally we blasted the gene against our own novel transcriptomes from *Ambystoma mexicanum*, *Acipenser ruthenus* and *Neoceratodus forsteri* and used the coding sequence identified previously in Section 2.4 (page 43). After assembling the coding sequences from all of the above datasets, each sequence was translated into protein.

We ran the protein sequences through MUSCLE, and began working manually to remove the sequences that were from the wrong gene or were unable to align. Well supported clades of paralogous genes were removed, as well as taxa from clades that were over-sampled, particularly the mammals and teleost fish.

After each step of removing sequences MUSCLE was re-run to keep the alignment up-to-date. Once we were left with just the gene of interest, and perhaps one or two paralogs depending on the gene, we began reviewing the alignment by hand. Changes to the alignment were made only if it was clear that MUSCLE had incorrectly aligned the sequences; this was particularly common for short EST sequences. Once we were satisfied with the quality of the alignment, we saved multiple versions depending on how many non-overlapping sequences there were. We also saved a version containing only the full length sequences.

On each alignment Gblocks was run using the default settings to remove all gap and low quality positions. We then used PhyML to create a maximum likelihood tree using the LG model with estimated gamma correction (Le and Gascuel, 2008; Yang, 1994). The tree was tested by 100 bootstrap replicates. We also created a Bayesian tree using the program Mr.Bayes which used the JTT model, again with estimated gamma correction (Jones et al., 1992). We set Mr.Bayes to use 4 runs and continued expanding the generations until the split deviation reached less than 0.011 or we reached 10M generations.

We also aligned the DNA sequences according to the protein alignment that was selected by Gblocks and as before used PhyML to create a ML tree. We selected the GTR+G model, with estimated matrix and gamma parameters and as before 100 bootstrap replicates were run. The same model was used in Mr.Bayes to create the Bayesian tree, using the same procedure as for the proteins.

Each tree was manually rooted on the most basal taxon and re-ordered according to the 'balanced shape' option in MEGA5 (Tamura et al., 2011) before exporting as a PDF. The final alignment files are included in the accompanying CD-ROM.

2.7 Relative Rate Test

To compare the relative rate of molecular evolution between sister taxa we ran the relative rate test (RRT) on three taxon alignments. Each alignment contained the two sister taxa as well as a reference species.

The original RRTs were run using 3 taxon alignments built using the same methods as previously (Section 2.5.1). Each test was run using the program HyPhy with either the GTR+G or JTT+G model depending on whether the alignment was in DNA or protein. HyPhy works by creating two ML trees, one of

which allows a complete estimation of parameters. The second is constrained so the two sister taxa branches are evolving at equal rates. The program then tests the likelihood of both trees and asks whether they are significantly different (Muse and Weir, 1992; Pond et al., 2005). If they are evolving at significantly different rates, the branch lengths calculated in the initial tree are used to determine which of the sister taxa is significantly slower/faster.

The same methods were applied in Chapter 6 to assess the rate differences within a whole gene family. In this case the alignment from a phylogenetic tree (Section 2.6) was used and each pair of sequences were tested in turn using the same outgroup reference. The results from each comparison between a taxon using epigenesis and one that has acquired preformation were stored. For each of these comparisons the difference in branch length was noted as well as the probability of a significantly different rate of evolution.

2.8 Gene Ontology

The Gene Ontology (Ashburner et al., 2000) is a list of terms (GO-terms) that describe the molecular function, biological process and cellular component of a gene. The terms are related to each other, mostly in a hierarchal sense of increasing specificity (e.g. 'GO:0005667 transcription factor complex' is the child of 'GO:0043234 protein complex'). This unified language allows for a bioinformatic comparison of genes between multiple species. However, since there are currently 398,424 GO-terms (as of 17/02/2014), many of which are very similar (e.g. 'GO:0001071 nucleic acid binding transcription factor activity' and 'GO:0003700 sequence-specific DNA binding transcription factor activity') there is still a level of human-bias/error in the choice of terms allocated to a gene.

The site GOrilla (Eden et al., 2009) was used to compare the Gene Ontology terms associated with a target list of genes against a background list. GO-terms that are significantly enriched in the target gene list compared to the total are identified. The p-value is corrected to account for multiple testing, resulting in the False Discovery Rate (FDR) q-value.

2.9 Synteny

To assess the synteny between different genes, the best source of information is the Ensembl genome browser. To access this information automatically, we first blasted each of the sequences in a fasta file against a range of vertebrate genomes (Section 2.2.5); each region that had a hit was then blasted against the 'prot-nr' database. If the resulting hits contained the name of the gene within their description then the gene was likely an ortholog. We then mined the Ensembl API to locate the genomic region and print the names of the 10 neighbouring protein-coding genes both upstream and downstream. This allowed us to locate the syntenic regions around each gene, even if it was not annotated in Ensembl.

This information was used to populate a table which contained all of the species queried, the gene and its key neighbours. For any genes that we have been unable to locate using the above program, we manually searched the databases. This involved manually querying Ensembl and blasting the initial gene against the mRNA, EST and genome sequences for that species in NCBI (if available).

We also searched for synteny in orthologs beyond vertebrates, and for this we used the known ortholog information stored in the EnsemblGenomes database. As before, once we knew the location of the ortholog, we were able to print the known neighbouring genes. Once we had the syntenic information for vertebrates and invertebrates it was possible to identify the key patterns and deduce the relationship within each gene family.

Global Analysis of Vertebrate Protein-Coding Genes

To test whether molecular evolution associates with the mode of PGC specification we have analysed protein-coding gene conservation in vertebrates. We have performed a global analysis wherein all available genes have been compared between sister taxa where one has retained epigenesis and the other has acquired preformation. This approach has been used in each clade where preformation has evolved; amphibians, actinopterygian fish and sauropsids. We have analysed these genes using phylogenetic trees and distance matrices, as well as testing for rate heterogeneity.

As discussed in Section 1.5.1 (page 26) not all phylogenetic trees agree on the same topology between species. This conflict is known as phylogenetic incongruence and has been attributed both to biological processes as well as analytical errors (Maddison, 1997; Rokas et al., 2003b). In our analyses we know the true evolutionary relationships among the species (Section 1.1), and we therefore use the term incongruence to specify those topologies that do not recapitulate the species phylogeny.

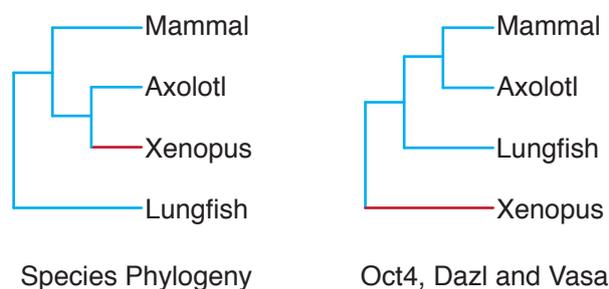


Figure 3.1: Oct4, Dazl and Vasa topologies. The Oct4, Dazl and Vasa topologies showed an incongruent topology according to Johnson et al., 2003a. Dazl did not show Lungfish in the above position but all three showed Mammal grouped with Axolotl. The branches are coloured according to the mode of PGC specification; epigenesis in blue, preformation in red.

Incongruence within amphibians is widely reported in the literature (Table A.1, Appendix page 221), and has been suggested to associate with the mode of PGC specification (Johnson et al., 2003a). This proposal was based on the incongruent topologies for Oct4, Dazl and Vasa; three genes involved in PGC specification (Figure 3.1). In these three trees, urodele and mammal, both of which use epigenesis, were grouped together, to the exclusion of anurans which have acquired preformation. It was therefore suggested that this topology reflected the mechanistic relationship of PGC specification. However, it has since been shown that the Oct4 incongruent tree was due to paralogy (see Section 6.1; Frankenberg et al., 2010; Niwa et al., 2008; Tapia et al., 2012).

We investigated the extent of incongruence within vertebrates, and whether it associates with the mode of PGC specification. Specifically whether species that have retained epigenesis are grouped together contrary to the species phylogeny. For this experiment we built four-taxon trees for as many orthologous genes as possible (see Sections 2.3 and 2.5). We also directly compared the DNA sequence similarity between species with the construction of distance matrices.

One of the potential causes of incongruence is Long Branch Attraction, due to differences in the rate of molecular evolution (Section 1.5; Anderson and Swofford, 2004; Felsenstein, 1978; Philippe and Laurent, 1998; Sanderson et al., 2000). We therefore analysed rate heterogeneity by performing the relative rate test, which compares the rate of evolution between two sister taxa (Section 2.7; Kimura, 1980; Muse and Weir, 1992; Tajima, 1993).

3.1 Data Evaluation

Before analysing the results of the phylogenetic trees and relative rate tests it was important to evaluate the data and develop a method for ensuring high quality alignments. The following process was carried out on all sequences but for clarity of presentation I will describe it using the amphibian four-taxon trees only.

The amphibian four-taxon trees consisted of an anuran, urodele, mammal and teleost sequence. There were only three possible topologies when rooted on the teleost outgroup, either the species phylogeny (anuran-urodele), or an incongruent topology with mammals grouping with either urodeles or anurans (Figure 3.2).

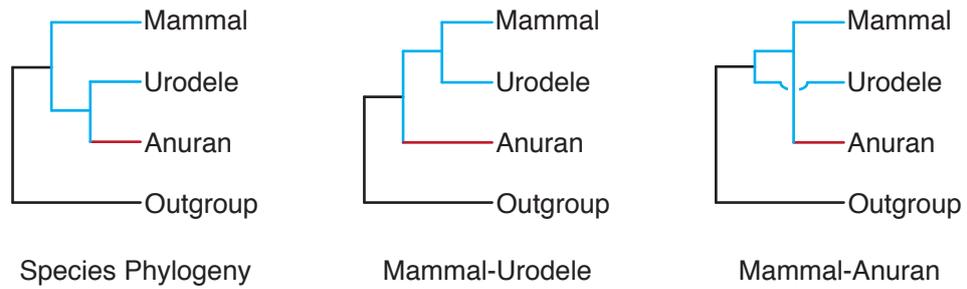


Figure 3.2: Amphibian 4-taxon topologies. The three possible topologies are shown, the species phylogeny, Mammal-Urodele and Mammal-Anuran. The branches are coloured according to the mode of PGC specification, either epi-genesis (blue) or preformation (red).

Each tree was tested for significance using 1000 bootstrap replicates and was considered significant if supported by more than 70%. Figure 3.3 shows the proportion of tree topologies obtained for the amphibian species. Approximately 40% of significant tree topologies show the species phylogeny, the remainder are incongruent. Within the incongruent trees there is a bias towards the Mammal-Urodele topology.

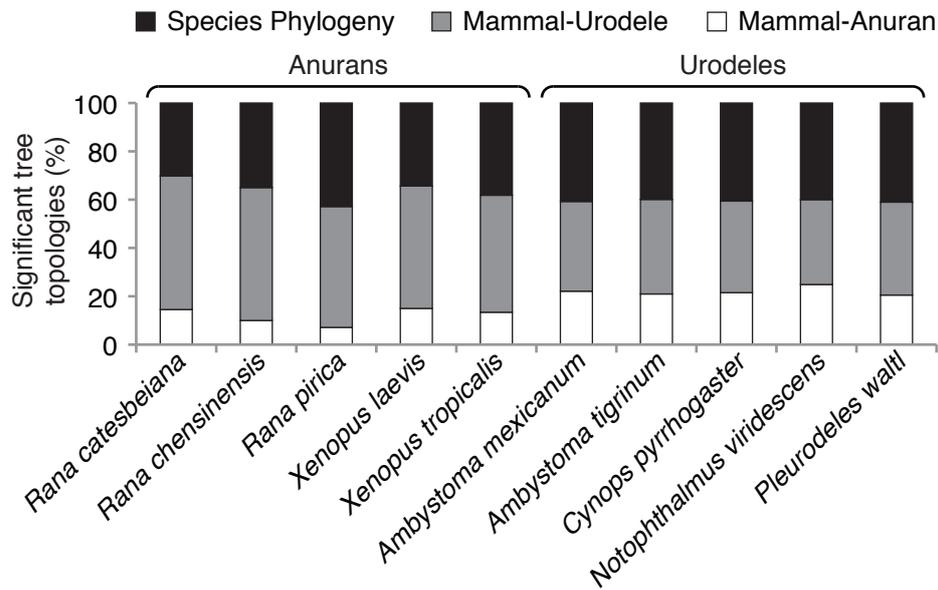


Figure 3.3: Initial tree topologies. The proportion of trees showing the species phylogeny are shown in black, the proportion showing the Mammal-Urodele topology in grey and the remaining trees in white had the Mammal-Anuran topology. Only those species with >20 significant trees are shown, trees were considered significant if supported by >70% of the bootstrap replicates.

This bias is further illustrated in Figure 3.4, where each species shows the likelihood of it grouping with mammals when the tree is incongruent. All of the anuran species are less likely to group with mammals while all the urodeles

species are more likely to do so. The number of tree topologies for all amphibian species are shown in Table 3.1.

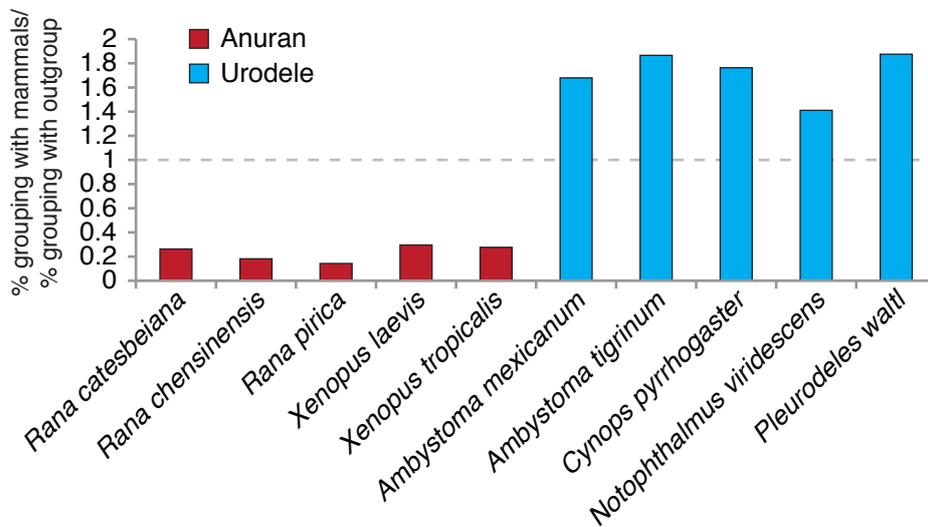


Figure 3.4: Initial bias within the incongruent topologies. The likelihood of each species grouping with mammals within the incongruent trees is shown. The anurans are shown in red, the urodeles in blue. Only those species with >20 significant trees are shown, trees were considered significant if supported by >70% of the bootstrap replicates.

These results show a surprisingly large number of trees are unable to recapitulate the species phylogeny, in many species the majority of trees have a Mammal-Urodele topology. This contradicts our assumption that the gene trees should reflect the species phylogeny. However it is possible that there are poor quality alignments, or other analytical artefacts affecting the result.

3.1.1 Refining the Quality Parameters

To assess the quality of the alignments we looked at the information stored within the MySQL tables. We noticed that for a small number of trees the corrected distance reached very high values. In fact 29 (out of 12,793) trees had a maximum distance set to 999.999, the highest value we were able to store in our MySQL database. Although there appeared to be nothing wrong with the alignments, in each of these cases the rate matrix had estimated at least one extremely large value (Table A.3, Appendix page 224). We therefore decided to remove trees with an uncommonly large corrected distance. To select the parameter value used to discard trees we looked at both the number of trees as well as their appearance (Table 3.2 and Figure B.1, Appendix page 247).

Table 3.1: Number of initial amphibian trees.

Species	Total	Species Phylogeny	Mammal- Urodele	Mammal- Anuran
<i>Rana catesbeiana</i>	948	137	251	66
<i>Rana chensinensis</i>	210	35	55	10
<i>Rana pirica</i>	63	12	14	2
<i>Xenopus laevis</i>	2759	463	682	202
<i>Xenopus tropicalis</i>	3535	672	853	236
<i>Ambystoma mexicanum</i>	1667	293	267	159
<i>Ambystoma tigrinum</i>	1409	255	250	134
<i>Andrias davidianus</i>	34	10	6	4
<i>Cynops pyrrhogaster</i>	1628	319	298	169
<i>Desmognathus ocoee</i>	8	0	0	3
<i>Notophthalmus viridescens</i>	459	90	79	56
<i>Pleurodeles waltl</i>	73	16	15	8
Total	12 793	2302	2770	1049

Table 3.2: Number of trees by maximum corrected distance.

Maximum distance	Number of Trees
≥ 0.5	10 455
≥ 1	1486
≥ 2	189
≥ 3	128
≥ 4	111
≥ 5	106
≥ 10	87
≥ 50	52
≥ 100	39

Table 3.2 shows that the majority of the 12,793 Amphibian trees have a maximum corrected distance between 0.5 and 1. However there are a small number with values greater than this, and by observing which trees had a distorted appearance (Figure B.1, Appendix page 247) we decided to remove all trees with a maximum corrected distance greater than or equal to four. We also looked at the difference between the minimum and maximum corrected distances using the same approach and decided to remove all of those with a difference ≥ 2 . Combining these two parameter selection criteria removed a total of 117 trees.

We next looked at the problem of small distances and short branch lengths. We noticed that a few trees had uncorrected distances that were extremely small (Table 3.3). All of the trees had a minimum uncorrected distance less than 0.5, but few had a value smaller than 0.1. Some of the trees with problematic branch lengths are shown in Figure B.2 (Appendix, page 248).

Table 3.3: Number of trees by minimum uncorrected distance.

Minimum distance	Number of Trees
≤ 0.5	12 672
≤ 0.2	4654
≤ 0.1	136
≤ 0.08	107
≤ 0.06	71
≤ 0.04	41
≤ 0.02	27
≤ 0.01	20

Based on this information we removed all trees with a minimum uncorrected distance less than or equal to 0.02. We also looked at the value for the overall percentage identity and decided to remove those with a value $\geq 90\%$ as these alignments showed almost no differences among the four sequences. By combining these two parameter criteria we had removed an additional 28 trees, leaving 12,644. Although the number of trees had not changed a great deal, the quality had increased by removing these outliers.

Finally we looked at alignment length; this was problematic as most of our sequences are ESTs and therefore the alignments tended to be short (Figure 3.5). We decided to exclude all alignments with a length less than 400bp, removing a total of 3,464 trees. Although this is a large proportion of trees, we felt that including any shorter alignments would decrease our confidence in the results.

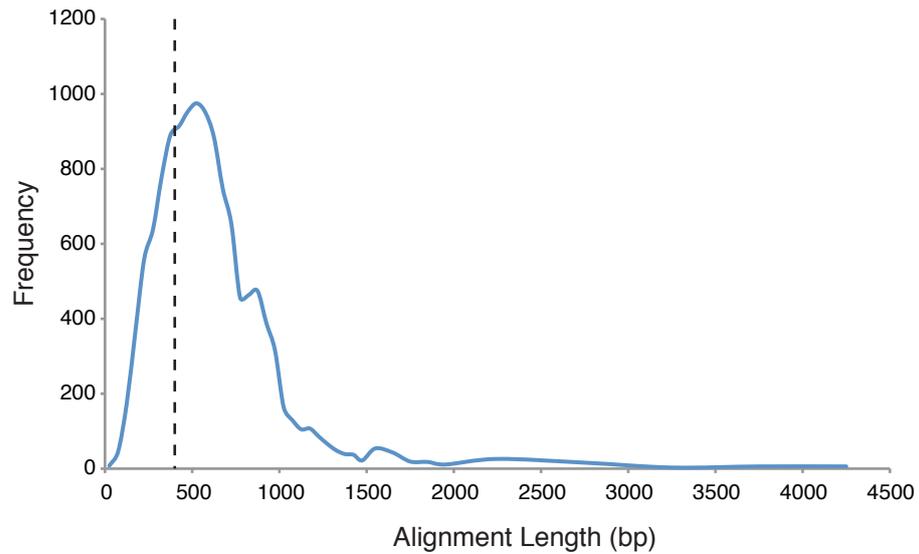


Figure 3.5: Alignment Lengths. This histogram shows the frequency of each alignment length, the mean length was 598.2. The dashed line marks the 400bp limit we applied to the alignments.

By incorporating the parameter values discussed above I was now able to produce the following graph of tree topology proportions and likelihood of grouping with mammals (Figure 3.6 and Table A.4, Appendix page 225). This shows the same result as before; approximately 40% of gene trees recapitulate the species phylogeny. Within the remainder there is a strong and consistent bias towards urodeles grouping with mammals. This suggests that the incongruent trees are not due to low quality or short alignments with insufficient information.

3.1.2 Protein coding results

Using the methods described in Section 2.4 (page 43), we identified the protein-coding regions for each sequence. For the 133,095 amphibian query sequences, we were able to find protein coding regions for 89,918 of them. Protein alignments were created and trees built using the same orthologs as for the DNA. The results of all amphibian protein gene trees are shown in Figure 3.7. The majority of trees now show the species phylogeny, and although there is still a bias within the incongruent trees it is not as clear as before. The original DNA comparisons showed approximately twice as many Mammal-Urodele trees as Mammal-Anuran trees; this difference has decreased using protein alignments.

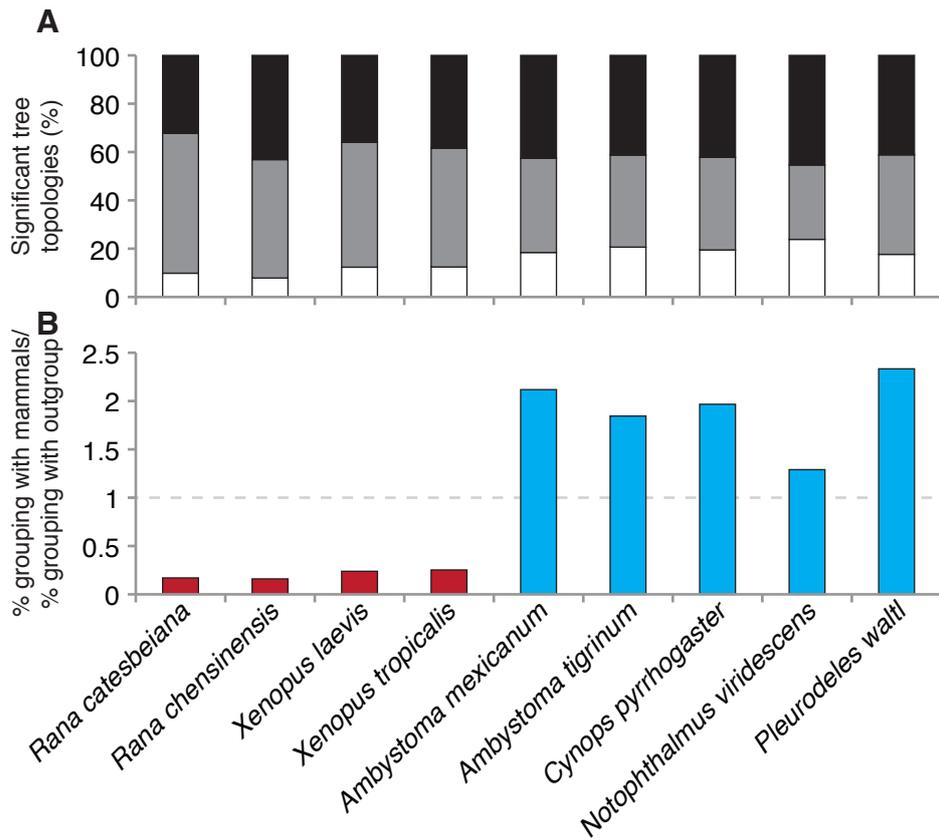


Figure 3.6: Amphibian results with refined quality parameters. (A) The proportions of significant topologies: the trees reflecting the species phylogeny are shown in black, the Mammal-Urodele trees are in grey and the Mammal-Anuran trees in white. (B) The likelihood of each species grouping with mammals when the tree is incongruent: the anurans are shown in red, the urodeles in blue. Only species with >20 significant trees are shown.

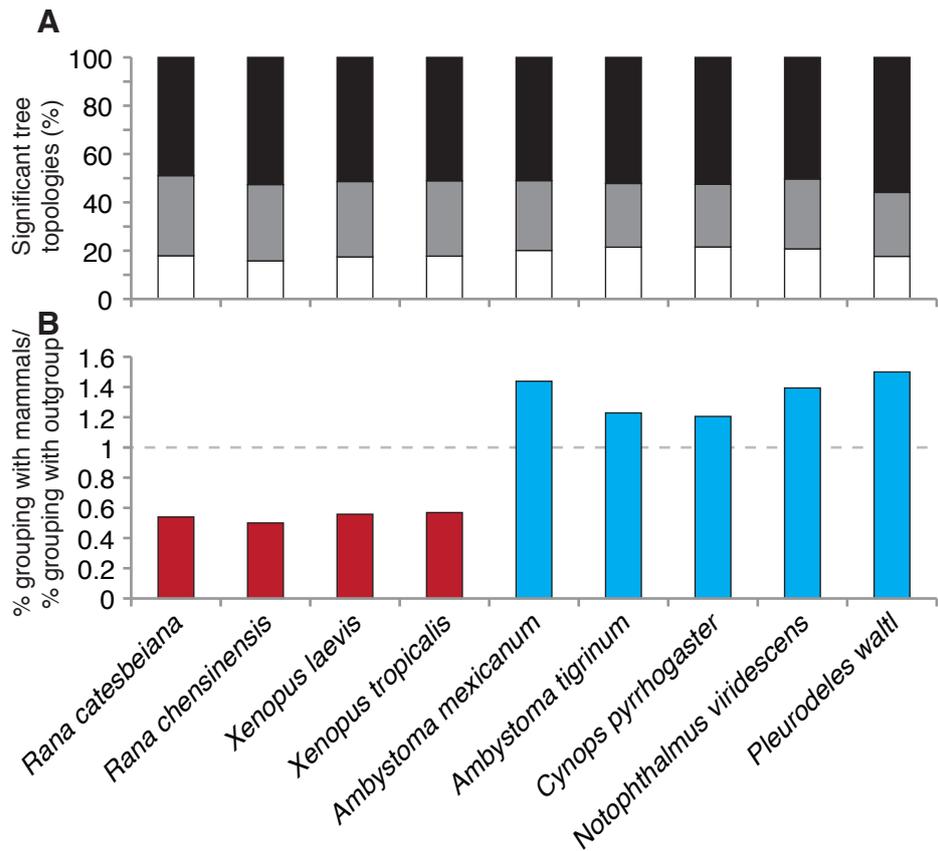


Figure 3.7: Protein tree results. (A) The proportions of significant topologies: the trees reflecting the species phylogeny are shown in black, the Mammal-Urodele trees are in grey and the Mammal-Anuran trees in white. (B) The likelihood of each species grouping with mammals when the tree is incongruent: the anurans are shown in red, the urodeles in blue. Only species with >20 significant trees are shown.

To explore why the result changes I calculated the results using DNA, protein and various codon position alignments (Figure 3.8 and Table 3.4). The proportion of Mammal-Anuran trees does not change between the different alignment types, however the numbers of species phylogeny and Mammal-Urodele trees do change. Using the protein sequences increases the number of Species-Phylogeny trees, and decreases the number of Mammal-Urodele trees. This change is not due to aligning new regions, as the codon alignments including all three positions (ORF 1+2+3) show the same pattern as the original DNA alignments. The difference can be understood however when excluding the third position of each codon, in this case the ORF(1+2) alignments have a similar result to the protein. This suggests that it is an affect of the third codon position that is causing an increase in Mammal-Urodele tree topologies. Indeed, the third codon position shows the highest proportion of Mammal-Urodele trees.

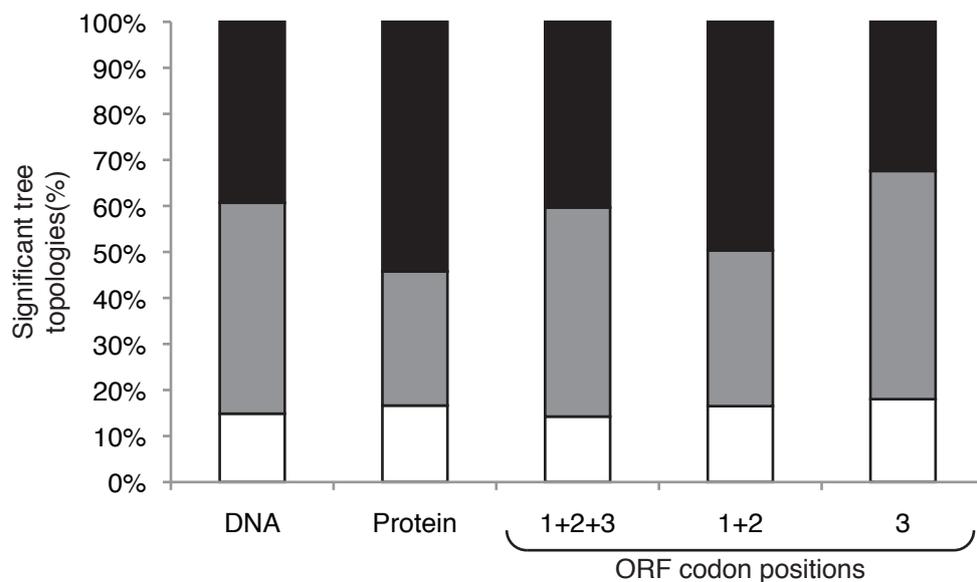


Figure 3.8: Tree topologies using different alignments. As before the species phylogeny trees are shown in black, the Mammal-Urodele in grey and the Mammal-Anuran in white. The proportion of significant tree topologies are shown for the original DNA alignments, the protein alignments and using the ORF sequences.

The third codon position is the most variable position since most mutations are synonymous, it is therefore likely to be saturated over the distances we are analysing. However, saturation plots for the third codon position cannot be built using 4-taxon trees; there would be too few points for a reliable result. Instead we looked at the alignment properties at the third codon position and noticed a large variance in the %GC. The first two codon positions all share

Table 3.4: Number of significant trees using each alignment type.

Type	Significant	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
DNA	3793	1491	1738	564
Protein	3530	1914	1029	587
ORF(1+2+3)	3835	1547	1742	546
ORF(1+2)	3656	1816	1237	603
ORF(3)	3336	1081	1654	601

very similar %GC values, however the third codon position has a large variance (Figure B.3, Appendix page 249). It is well understood that tree topologies are affected by biases within the %GC (Foster and Hickey, 1999), and we therefore reasoned that this variance was the cause for the increased proportion of Mammal-Urodele gene trees.

To investigate this, we examined the %GC for each sequence in our trees, using the different codon positions (Figure 3.9). In the third codon position, when the tree recapitulates the Species Phylogeny, the two amphibians tend to have a similar %GC to each other, while the mammal and outgroup both have higher %GC values. The trees with a Mammal-Urodele topology show a higher %GC in the outgroup than in the other three species. In the 546 trees with a Mammal-Anuran topology, the mammals and anurans both have lower %GCs than the urodele and outgroup species. Since these differences generally reflect the tree topology it suggests that the tree building process is being affected by the base composition of the third codon position. Although we did not observe an increase in Mammal-Urodele trees using protein sequences, it is known that even protein trees can be affected by an underlying GC bias (Foster and Hickey, 1999). We have therefore presented the results using the first two codon positions from here on, as there was no variance in the %GC.

Exploring the bias in %GC

The high variance of %GC observed in the third codon position was not unique to amphibians, as is exemplified by looking at both the Amphibia and Actinopterygii (Figure 3.10).

All four taxa have a wide variance in %GC at the third codon position, but there does not appear to be any kind of pattern within this variance; there is no noticeable correlation between the base composition and the mode of PGC

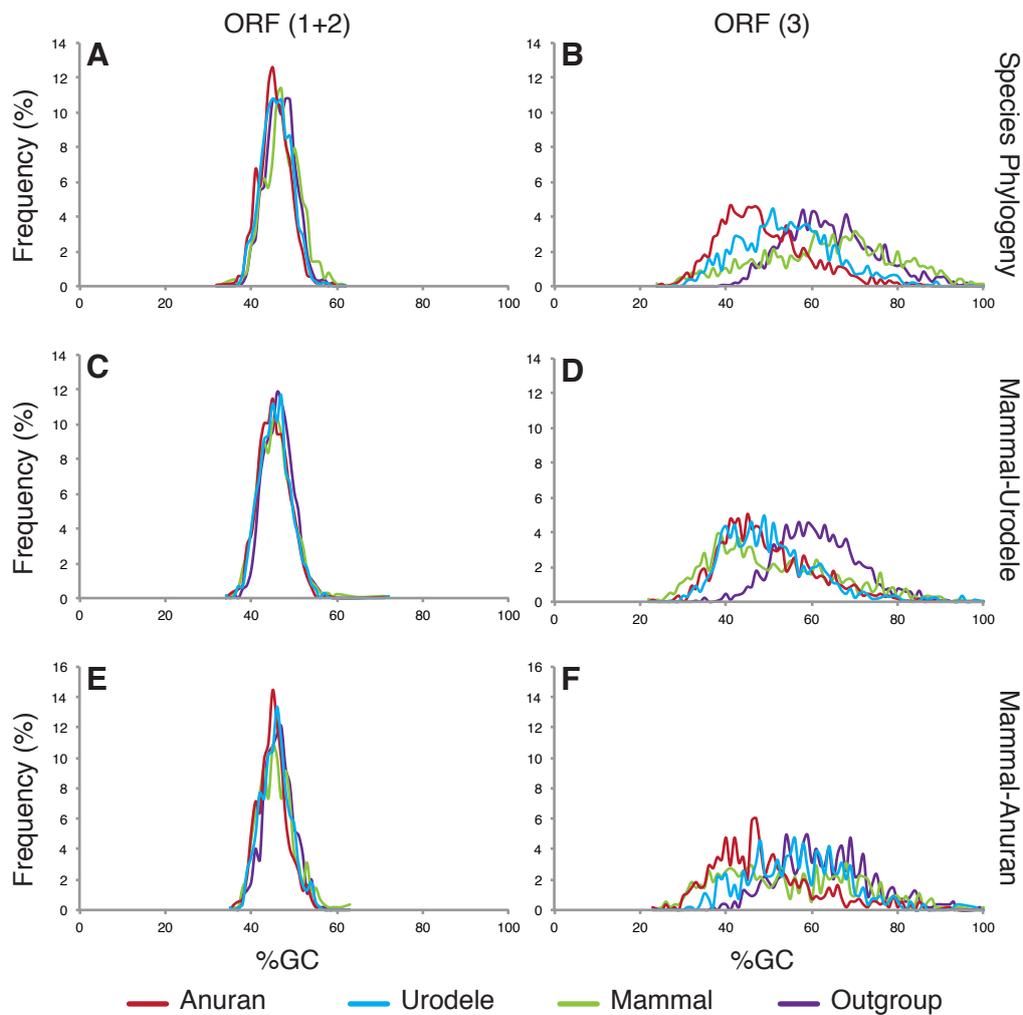


Figure 3.9: The %GC frequencies for the different tree topologies. The percentage of trees with each %GC are shown using either the first two codon positions (A, C and E) or the third codon position (B, D and F). The tree topology is calculated using the whole ORF. (A and B) show the Species Phylogeny, (C and D) Mammal-Urodele, (E and F) Mammal-Anuran.

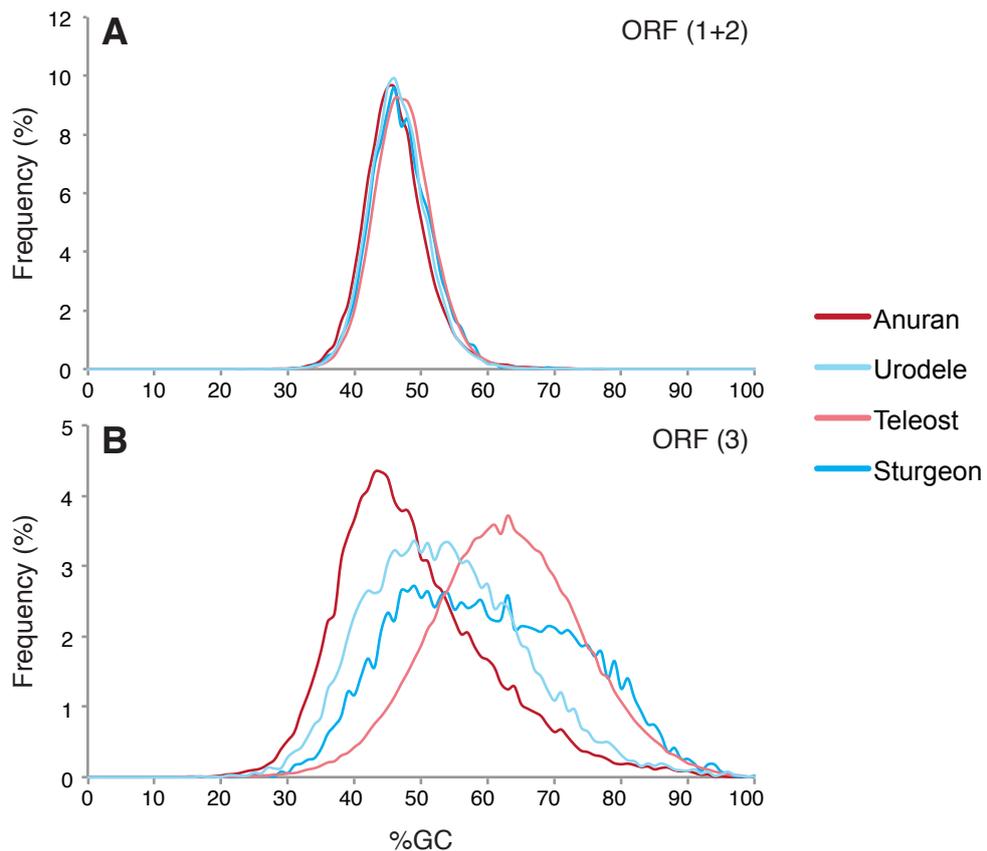


Figure 3.10: The %GC for Amphibia and Actinopterygii. The %GC is shown for all of the sequences from anurans, urodeles, teleost and sturgeon species. The results using the first two codon positions (**A**) show no difference between the orders, however using the third codon position (**B**) shows a wide variance.

specification. The anuran %GC at the third codon position is low compared to urodeles, whereas teleosts have a higher %GC than sturgeon. There is therefore no generalised shift in %GC associated with the evolution of a preformed germ line. Interestingly, there does appear to be a close similarity between the %GC of urodeles and sturgeon.

3.1.3 Likelihood tests of the tree topology

Previously all tree topologies have been deemed significant using bootstrapping. An alternative is to use a likelihood ratio test, where the null and alternative hypotheses are tested using maximum-likelihood (for review, see Goldman et al., 2000; Huelsenbeck and Crandall, 1997; Schmidt and von Haeseler, 2009). We have used the one-sided Kishino-Hasegawa (KH; Goldman et al., 2000; Kishino and Hasegawa, 1989), Shimodaira-Hasegawa (SH; Shimodaira and Hasegawa, 1999) and approximately unbiased (AU; (Shimodaira, 2002))

likelihood ratio tests. In each case the null hypothesis is that all tree topologies are equally good explanations of the data. We have also calculated the expected likelihood weights (ELW; Strimmer and Rambaut, 2002).

Each likelihood test was run on all three possible topologies, and the ML tree was deemed significant if both alternative topologies had probability values less than 0.05. In other words, their probability of being equally good explanations of the data was negligible. The same values were used for the ELW tests, the two alternative topologies had to have weights less than 0.05.

Figure 3.11 shows the tree topologies using these different tests on the first two codon position amphibian alignments. All four of the tests have very similar results, in each case there are approximately 60% of the trees reflecting the species phylogeny. Within the remaining incongruent trees, there is still a bias towards the Mammal-Urodele topologies.

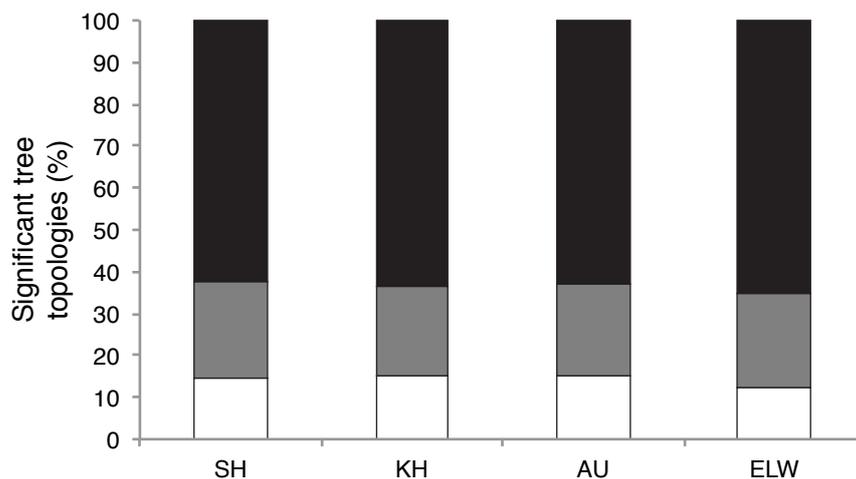


Figure 3.11: The amphibian tree topologies using different likelihood tests. As before the species phylogeny trees are shown in black, the Mammal-Urodele topologies in grey and the Mammal-Anuran in white.

Table 3.5: Number of significant trees using each likelihood test.

Test	Significant	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
SH	210	131	48	31
KH	209	133	44	32
AU	287	181	62	44
ELW	331	215	75	41

The likelihood ratio tests differ to the bootstrap test in the number of trees identified as significant. Table 3.5 shows that the likelihood tests deem between

200 and 300 trees as being significant, this contrasts to the 3,656 significant trees by bootstrapping. This exemplifies the difference in stringency between the likelihood ratio tests and the bootstrap replicates.

Interestingly, we noticed an unusual result when using the protein alignments to build trees, in these cases the AU test behaved unlike the other three tests. Figure 3.12 shows the tree topologies when using the amphibian protein alignments. The AU test has a greater proportion of incongruent trees, as well as almost a complete loss of the bias towards Mammal-Urodele topologies.

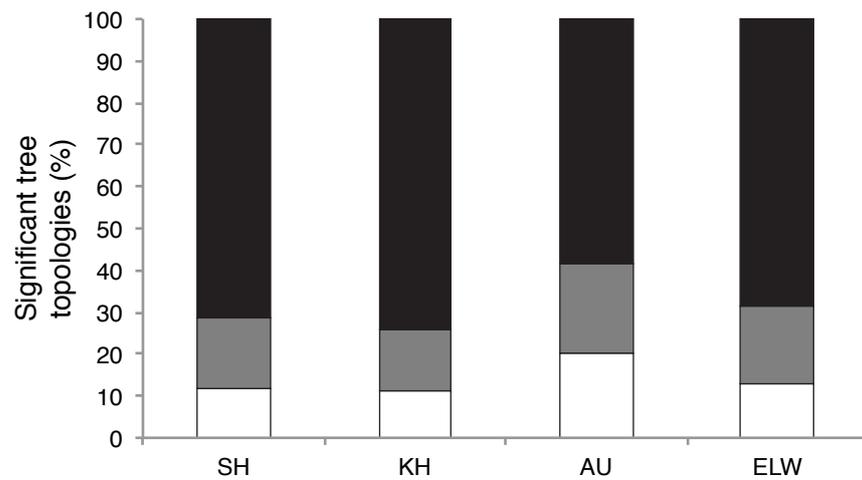


Figure 3.12: The protein trees using different likelihood tests. As before the species phylogeny trees are shown in black, the Mammal-Urodele topologies in grey and the Mammal-Anuran in white.

Table 3.6: Number of significant trees using each likelihood test and the protein alignments.

Test	Significant	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
SH	180	128	31	21
KH	173	128	26	19
AU	528	308	112	108
ELW	331	226	63	42

Looking at the absolute numbers (Table 3.6) shows that the AU test identified many more significant trees than the other likelihood tests. I have been unable to find any published records of the AU test behaving inappropriately, but since we only see this result using the protein sequences it certainly suggests an error within the program. We therefore decided to use only the results from the SH test from here on.

3.1.4 Pipeline Summary

We have evaluated the data and developed a pipeline which removes low quality alignments and uses only the first two codon positions (Figure 3.13). Thereby meaning the phylogenetic trees are not products of poor quality, short alignments or artefacts due to unequal base compositions. These phylogenies will be tested for their significance using both 1,000 bootstrap replicates and the SH likelihood ratio test.

This same pipeline will also be applied to the three-taxon alignments used for the distance matrices and relative rate tests.

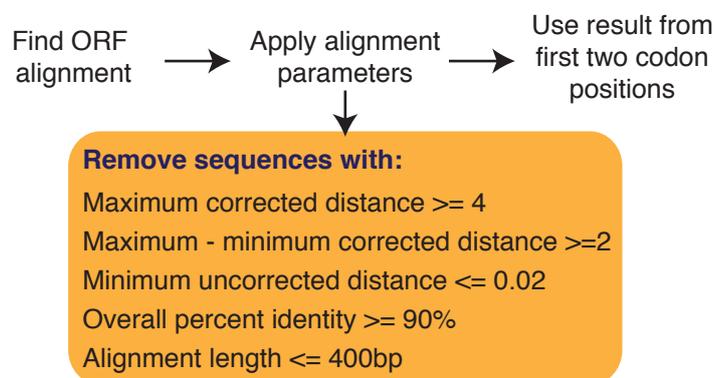


Figure 3.13: Analysis pipeline. For all future work shown in Chapters 3, 4 and 5 the data will have first gone through this pipeline to remove unreliable results.

3.2 Four-Taxon Phylogenetic Trees

Using the process described in the previous section we built high-quality alignments using the first two codon positions. The four-taxon phylogenetic trees from these alignments were assessed as to whether they produced incongruent trees, and if these trees were biased. This allowed to us to test whether the mode of PGC specification correlated with incongruent tree topologies. We began this process in amphibians but then widened the analysis to include all vertebrate groups where preformation has evolved.

3.2.1 Amphibians

Within amphibians we compared anurans, using preformation, against urodeles which utilise epigenesis. There are three possible topologies; the species

phylogeny, in which anurans and urodeles group together and the two incongruent topologies, where either anurans or urodeles group with mammals (Figure 3.2, page 57).

After applying the pipeline developed in the previous section, approximately 50% of the trees reflected the species phylogeny (Figure 3.14 and Table 3.7). The remainder showed a bias towards the Mammal-Urodele topology, which was consistent across all of the amphibian species studied.

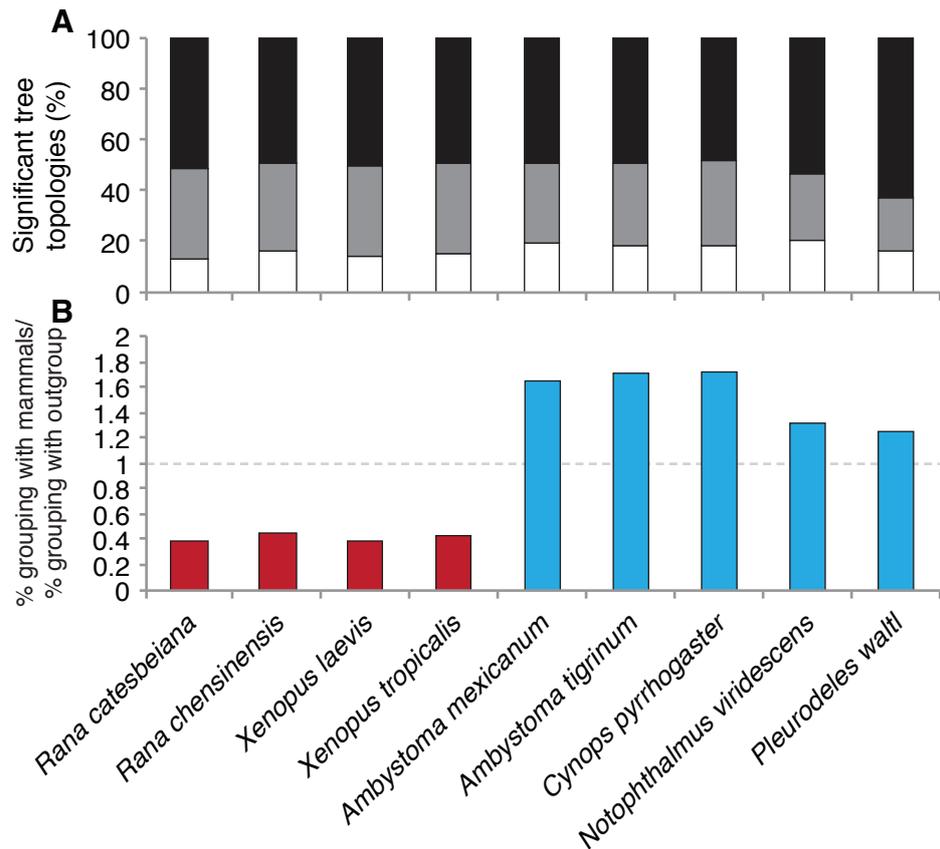


Figure 3.14: The amphibian bootstrap results using the first two codon positions. (A) The proportion of significant topologies for either the species phylogeny (black), Mammal-Urodele (grey) or Mammal-Anuran (white). (B) The likelihood of the incongruent trees to group the species with mammals, the anurans undergoing preformation are shown in red, the urodeles in blue.

Incorporating the Axolotl novel transcriptome

Using the data above we were able to analyse 7,678 trees, of which 3,656 had significant bootstrap support. Considering we had two species with near complete genomes (*Xenopus tropicalis* and *Xenopus laevis*) this was a surprisingly small number of trees. This was due to the lack of urodele sequences which are publicly available. To counter this we sequenced the transcriptome of the urodele

Table 3.7: Amphibian results using the first two codon positions.

Species	Total	Species Phylogeny	Mammal- Urodele	Mammal- Anuran
<i>Rana catesbeiana</i>	456	110	75	29
<i>Rana chensinensis</i>	77	18	13	6
<i>Rana pirica</i>	17	3	4	6
<i>Xenopus laevis</i>	1762	424	298	119
<i>Xenopus tropicalis</i>	2299	553	396	169
<i>Ambystoma mexicanum</i>	849	182	117	71
<i>Ambystoma tigrinum</i>	891	189	123	72
<i>Andrias davidianus</i>	21	6	4	3
<i>Cynops pyrrhogaster</i>	1068	265	177	103
<i>Desmognathus ocoee</i>	2	0	0	2
<i>Notophthalmus viridescens</i>	184	51	25	19
<i>Pleurodeles waltl</i>	52	15	5	4
Total	7678	1816	1237	603

Ambystoma mexicanum, as described in Section 2.2.2 (page 35). To combine the two datasets, for the anurans we blasted each sequence against a database containing the original and new *Ambystoma mexicanum* sequences. We then selected the tree result to use based on whether the original or transcriptome sequence had the better e-value. For the original axolotl sequences we removed all of those which had a reciprocal ortholog within the new transcriptome. After combining this novel transcriptome we had a total of 16,698 trees. This not only doubled the number of trees available to analyse but also increased the average alignment length (Figure 3.15).

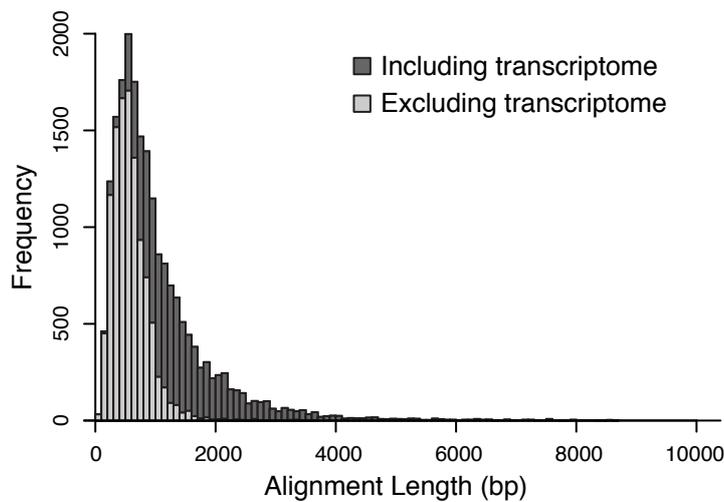


Figure 3.15: Open Reading Frame alignment lengths. The dark grey shows the alignment lengths including our novel *Ambystoma mexicanum* transcriptome. The pale grey shows the alignment lengths excluding the transcriptome.

Figure 3.16 and Table 3.8 show the tree results including the axolotl transcriptome, the other urodele species are included even though these results have not changed from Figure 3.14. The remaining amphibians all show the same general result as previously, approximately 50% of trees reflect the species phylogeny while there is a bias towards Mammal-Urodele topologies within the incongruent trees.

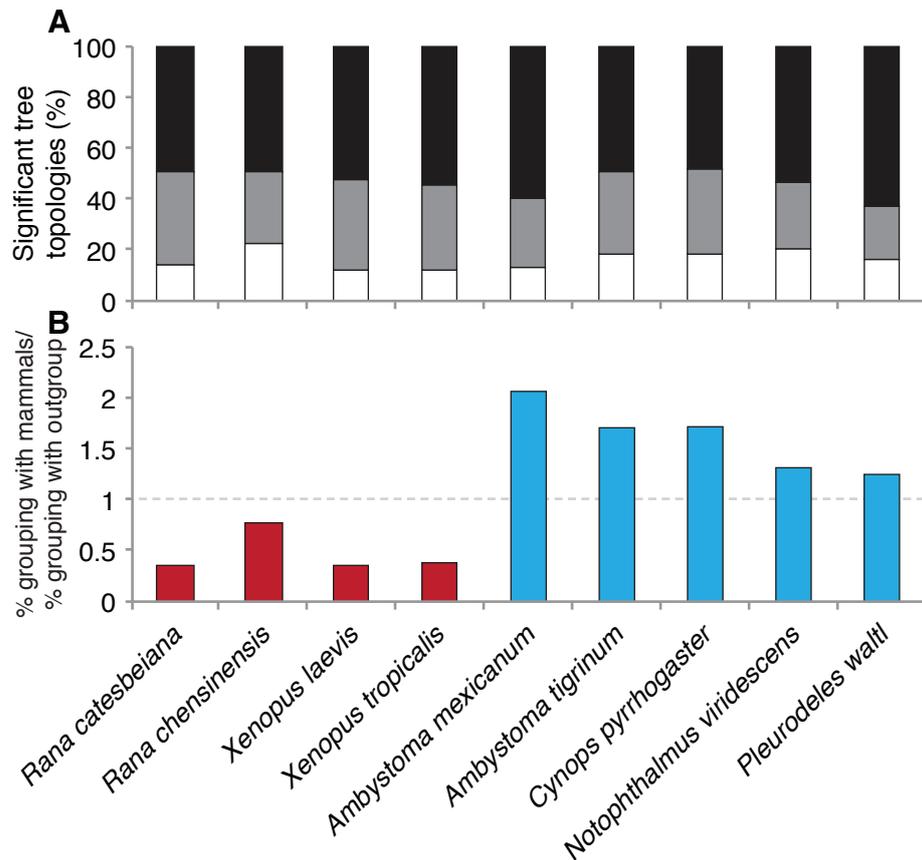


Figure 3.16: Amphibian bootstrap results including the axolotl transcriptome. The data is presented as in Figure 3.14.

There are some noticeable differences, particularly the result for *Rana chensinensis*. This species shows no change to the number of Mammal-Urodele trees while the number of other topologies increases (Table 3.8). This species now stands out amongst the other anurans. However, since there are only 45 significant trees this could still be an artefact caused by a limited analysis.

Within *Ambystoma mexicanum* itself, we have built an additional 2,732 trees, and we see a larger proportion of significant trees reflecting the species phylogeny. However, when incongruent the tree is twice as likely to group *Ambystoma mexicanum* with mammals than the outgroup (Figure 3.16B). Therefore

Table 3.8: Amphibian bootstrap results including the axolotl transcriptome.

Species	Total	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
<i>Rana catesbeiana</i>	942	206	158	58
<i>Rana chensinensis</i>	90	22	13	10
<i>Rana pirica</i>	25	4	5	4
<i>Xenopus laevis</i>	4178	1066	699	254
<i>Xenopus tropicalis</i>	5664	1504	904	349
<i>Ambystoma mexicanum</i>	3581	1027	471	228
<i>Ambystoma tigrinum</i>	891	189	123	72
<i>Andrias davidianus</i>	21	6	4	3
<i>Cynops pyrrhogaster</i>	1068	265	177	103
<i>Desmognathus ocoee</i>	2	0	0	2
<i>Notophthalmus viridescens</i>	184	51	25	19
<i>Pleurodeles waltl</i>	52	15	5	4
Total	16 698	4355	2584	1106

adding in the axolotl transcriptome increases the number of trees analysed but does not alter the result. In fact, for all affected species except *Rana chensinensis* the bias within the incongruent trees has intensified.

The likelihood ratio test

We next used the SH test to determine topology significance (Figure 3.17 and Table 3.9). In this more stringent test there are fewer significant trees, and the proportion of species phylogeny trees has increased (60%) compared to the bootstrap results. Within the incongruent trees the bias towards the Mammal-Urodele topology has also increased; *Ambystoma mexicanum* is now 3 times more likely to group with mammals when incongruent.

Summary

To conclude the amphibian four-taxon trees, Figure 3.18 shows the total obtained for each topology. We analysed 3,656 significant trees (bootstrap support >70%) before including the transcriptome, afterwards we built a total of 8,045 significant trees. Interestingly, although the total number of trees has increased, the proportion of each topology does not change dramatically. Approximately 50% of trees show the species phylogeny, 35% show mammals grouped with

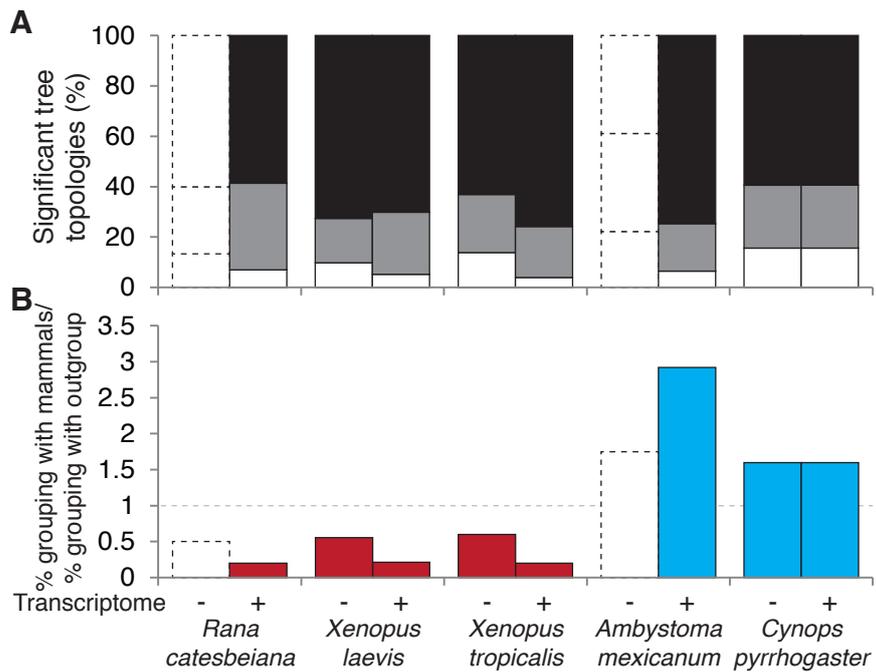


Figure 3.17: SH test results with and without the axolotl transcriptome. The top panel (A) shows the proportion of significant species phylogeny (black), Mammal-Urodele (grey) and Mammal-Anuran (white) topologies. The second panel (B) shows the likelihood of each species grouping with mammals when the tree is incongruent, the anurans are shown in red, the urodeles in blue. The results based on fewer than 20 significant trees are shown with dashed lines.

Table 3.9: Amphibian SH test results including the axolotl transcriptome.

Species	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
<i>Rana catesbeiana</i>	17	10	2
<i>Rana chensinensis</i>	3	1	0
<i>Rana pirica</i>	1	1	0
<i>Xenopus laevis</i>	134	47	10
<i>Xenopus tropicalis</i>	192	51	10
<i>Ambystoma mexicanum</i>	139	35	12
<i>Ambystoma tigrinum</i>	11	3	4
<i>Andrias davidianus</i>	1	0	0
<i>Cynops pyrrhogaster</i>	19	8	5
<i>Desmognathus ocoee</i>	0	0	1
<i>Notophthalmus viridescens</i>	2	2	1
<i>Pleurodeles waltl</i>	1	0	0
Total	520	158	45

urodeles and 15% show the Mammal-Anuran topology. Adding the novel axolotl transcriptome has not altered the outcome, but instead increased our confidence in the results by allowing the production of more trees.

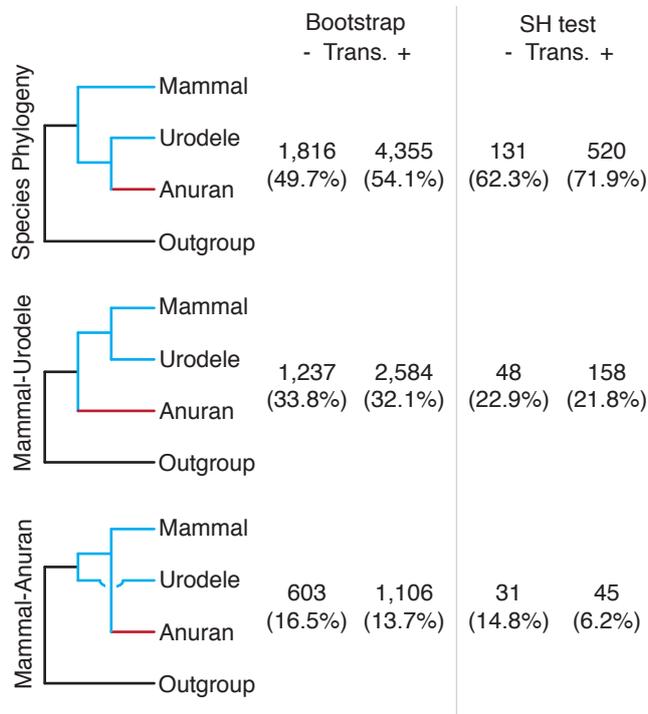


Figure 3.18: Summary of amphibian trees. This shows the total number of significant trees obtained for all three topologies. The values are shown for the bootstrap and SH tests, and also with and without the axolotl transcriptome.

The SH test results continue to show the bias within incongruent topologies. There is an increase in the number of species phylogeny trees, which we would expect, however there is also an increase in the bias within incongruent trees. This suggests that the bias observed previously is not an artefact caused by a poorly supported tree, but is instead a fundamental aspect of the data. Finally our results show that when an amphibian 4-taxon tree is unable to show the species phylogeny, there is a bias towards urodeles grouping with mammals. Therefore the two species that have retained epigenesis are incorrectly grouping together. As such, amphibians show a correlation between the mode of PGC specification and phylogenetic incongruence.

3.2.2 Actinopterygii

To determine whether this correlation is unique to amphibians we also built four-taxon phylogenies for the actinopterygian fish. We compared teleost fish against the sturgeon and paddlefish group, Acipenseriformes. Teleosts are

known to be undergoing preformation whilst sturgeons have likely retained the epigenesis mechanism of PGC specification (Section 1.3; Hashimoto et al., 2004; Johnson et al., 2011; Zelazowska et al., 2007). All trees were built with an additional mammal sequence and were rooted on an amphioxus outgroup. The three possible tree topologies are shown in Figure 3.19; the species phylogeny, where sturgeon and teleost group together, Mammal-Sturgeon and Mammal-Teleost.

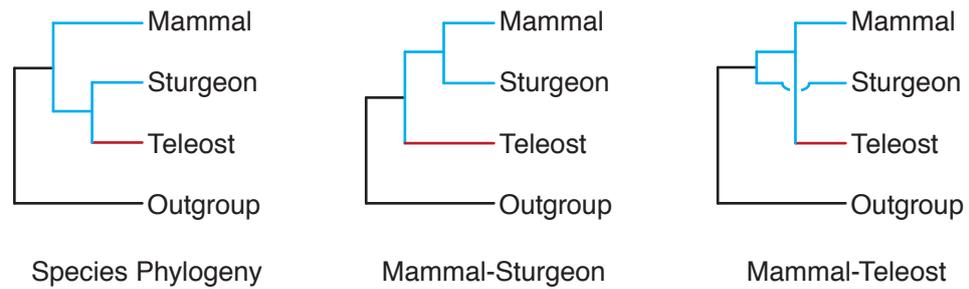


Figure 3.19: The three possible tree topologies for Actinopterygii. By building 4-taxon trees we can either have the Species Phylogeny, Mammal-Sturgeon or Mammal-Teleost topologies.

We downloaded sequences from 80 teleost species, although for clarity of presentation, only the results from 7 species are shown in the main text. The same results including all teleost species are shown in the appendix. We were initially able to build only 2,908 four-taxon trees, on average 35 trees per species. This very low number was due to the lack of *Acipenseriforme* sequences available. After the addition of the *Acipenser ruthenus* novel transcriptome we were able to build 36,338 trees.

The four-taxon tree results, significant by >70% of the bootstrap replicates, are shown in Figure 3.20. Approximately 70% of trees show the species phylogeny, a higher value than observed in amphibians. Within the incongruent trees there is a bias towards the Mammal-Sturgeon topology. This is particularly clear in Figure 3.20B, *Acipenser ruthenus* is 5 times more likely to group with mammals when the tree is incongruent. All of the teleost species are more likely to group with amphioxus than mammals (Appendix, Figure B.4 and Table A.5).

This consistent result is repeated in the trees significant by the SH test (Figure 3.21). Excluding the *Acipenser ruthenus* transcriptome, none of the species have more than 20 significant trees (Table A.6, Appendix page 225), therefore only the results that include the transcriptome are presented. *Acipenser ruthenus*

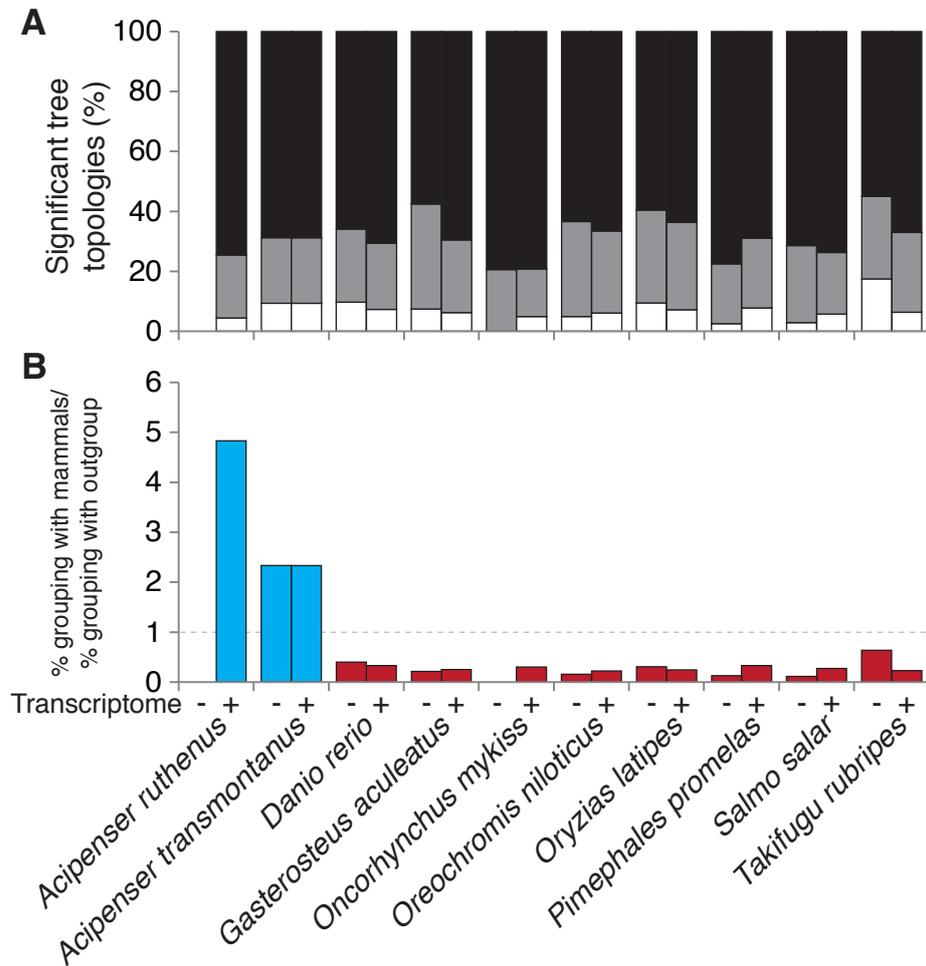


Figure 3.20: Bootstrap results for Actinopterygii. (A) shows the proportion of significant tree topologies for 7 species of teleost and the sturgeons; species phylogeny in black, Mammal-Sturgeon in grey and Mammal-Anuran in white. (B) shows the likelihood of each species grouping with mammals when the tree is incongruent. Sturgeons are shown in blue, teleosts in red. Each species is shown excluding and including the *Acipenser ruthenus* transcriptome. The results for all species are shown in Figure B.4 (Appendix, page 250).

is 21 times more likely to group with mammals than amphioxus when the tree is incongruent. This bias within the incongruent trees is far stronger than what was observed in amphibians.

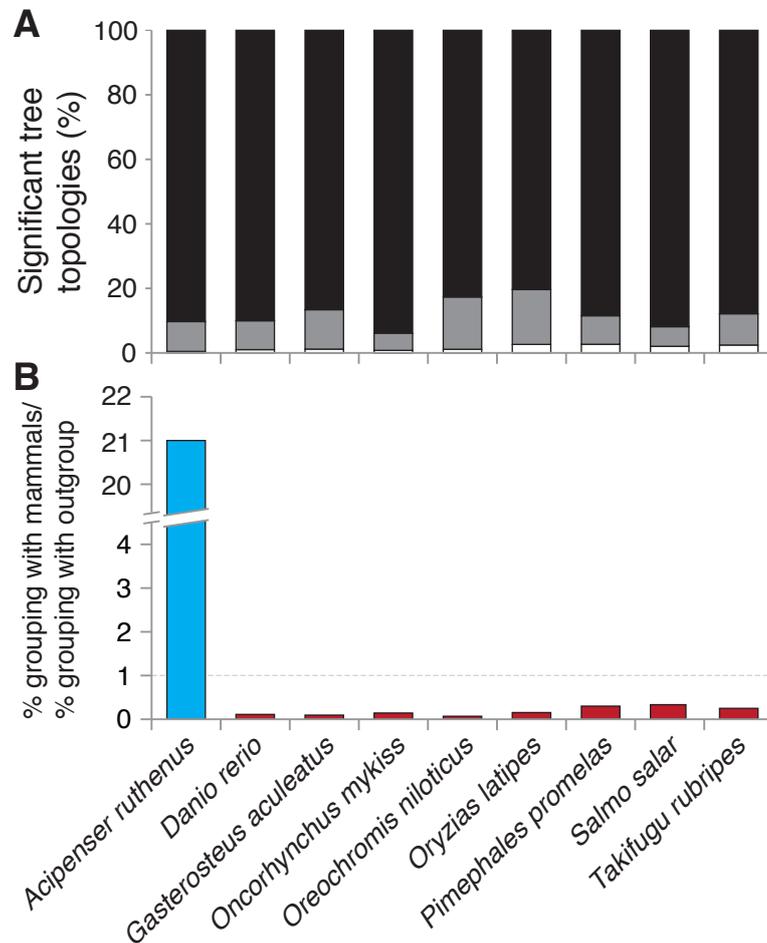


Figure 3.21: SH test results for Actinopterygii. The results are shown in the same format as Figure 3.20, including the sturgeon transcriptome. All teleost species are shown in Figure B.5 (Appendix, page 251).

Summarising the actinopterygian 4-taxon phylogenies (Figure 3.22), shows that adding the novel *Acipenser ruthenus* transcriptome has a major impact on the number of trees, but does not affect the result. Approximately 65% of the trees significant by bootstrapping recapitulate the species phylogeny. Within the incongruent trees, there is a strong bias towards the Mammal-Sturgeon topology compared to the Mammal-Teleost topology. This bias is even stronger when using the stringent SH test. As observed in the amphibians, the incongruent topology that occurs most often groups the two species undergoing epigenesis together.

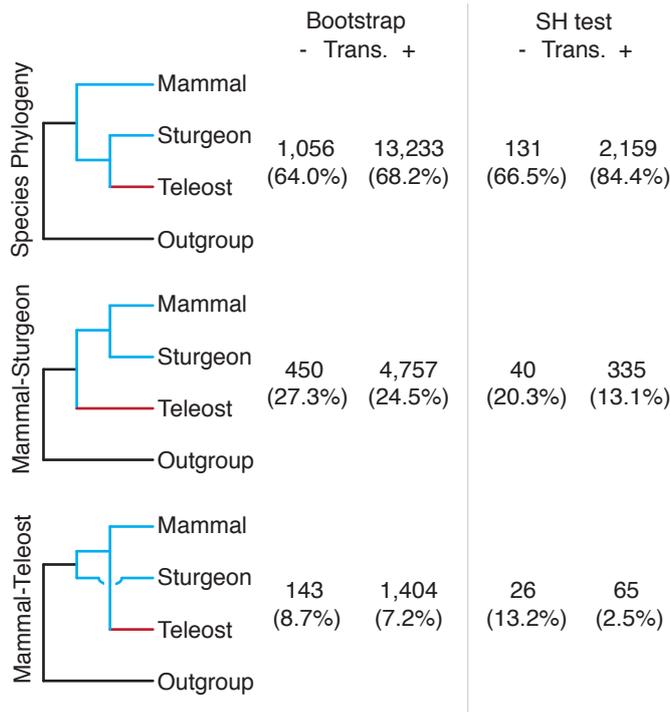


Figure 3.22: Summary of actinopterygian trees. This shows the total number of significant trees produced with and without the sturgeon transcriptome for both the bootstrap and SH tests.

3.2.3 Sauropsids

To investigate whether this correlation between incongruence and the mode of PGC specification occurs throughout vertebrates, we next investigated the sauropsids. To do this we compared two groups separately, the archosaurs and testudines, and the lepidosaurs. We separated these since there is clear evidence towards the mode of PGC specification in archosaurs but not in lepidosaurs (Section 1.3 (page 17)). There has also been a known change in the rate of evolution at the base of the Lepidosaurs, which would likely affect our tree topologies if included (Hughes and Mouchiroud, 2001). We had no additional novel transcriptome in either group so our results are based on the original sequences only.

Archosaurs and Testudines

Within this group we compared crocodiles and turtles (epigenesis) against birds (Preformation). Anuran sequences were used to root the trees, and when combined with mammal left three possible tree topologies (Figure 3.23).

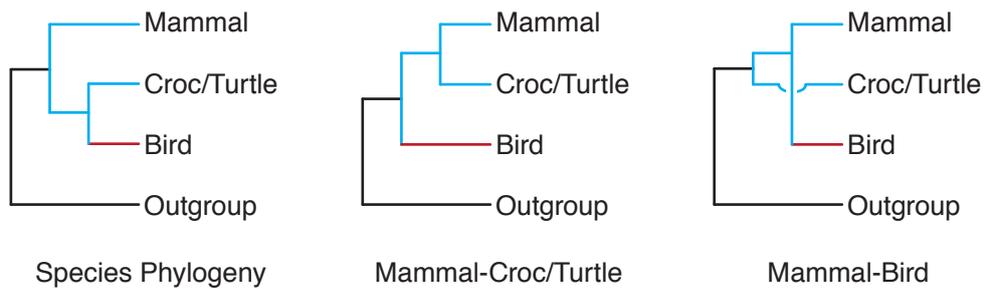


Figure 3.23: The three possible tree topologies for archosaurs. The phylogenetic tree will either show the species phylogeny where crocodile/turtle groups with bird, or a Mammal-Crocodile/Turtle or Mammal-Bird topology.

The results for the bootstrapped trees are shown in Figure 3.24 and Table A.7 (Appendix, page 225), approximately 80% of the trees reflect the species phylogeny. Interestingly, within the incongruent trees there is no consistent bias. Several birds show a high likelihood of grouping with mammals, two birds show an equal distribution between the topologies and one species shows a low likelihood of grouping with mammals. This is different to the amphibians and actinopterygian results which were consistent across whole orders. Archosaurs and Testudines also show the highest proportion of species phylogeny trees so far.

When using the SH test to measure tree confidence, only three species have over 20 significant results, all of which are birds (Table A.8, Appendix page 226). Across these species only 6 trees are incongruent, the remaining 67 reflect the species phylogeny. There are therefore not enough data available to display the SH test results on a per species basis.

Lepidosaurs

The available lepidosaur species consisted of two Iguanidae, one species of Gekkonidae and 16 snakes. The three species of lizard are all thought to be undergoing epigenesis, while the snakes are thought to be undergoing preformation (Hubert, 1985). We therefore built trees using a mammal, lizard, snake and anuran outgroup (Figure 3.25).

The bootstrap results (Figure 3.26 and Table A.9, Appendix page 226), show that almost all of the significant trees are correctly showing the species phylogeny. The few trees that are incongruent show no consistent bias. For the snake species that have incongruent trees, some group only with anurans while

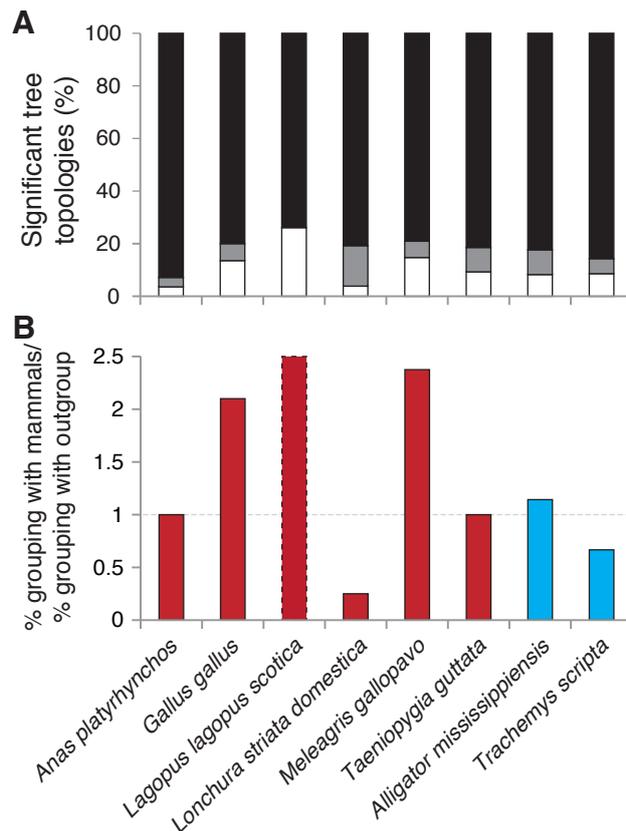


Figure 3.24: The bootstrap results for archosaurs. (A) shows the proportion of species phylogeny (black), Mammal-Croc/Turtle (grey) and Mammal-Bird (white) topologies significant by bootstrapping. (B) shows the likelihood of each species grouping with mammals; birds coloured red, crocodile and turtle in blue. *Lagopus lagopus scotica* only groups with mammal in the incongruent trees, this is represented by a full bar with a dashed outline.

other species only group with mammals. The two species of lizard that have more than 20 significant trees show an equal number of Mammal-Lizard and Mammal-Snake trees.

As in Archosaurs, the SH test had too few significant trees for us to analyse on a species by species basis (Table A.10, Appendix page 227). In fact there are only 3 incongruent trees across all Lepidosaurians when using the SH test to measure significance.

However it is possible to analyse the SH test results when combined across all species, as we have done for lepidosaurs and archosaurs in Figure 3.27. These results show that for archosaurs there is a tendency towards Mammal-Bird topologies when the tree is incongruent, although as we saw in Figure 3.24 this is not consistent across the species. Lepidosaurians show an unprecedented number of species phylogeny trees, and within those that are incongruent no bias towards either topology. Overall, the sauropsid trees differ from those

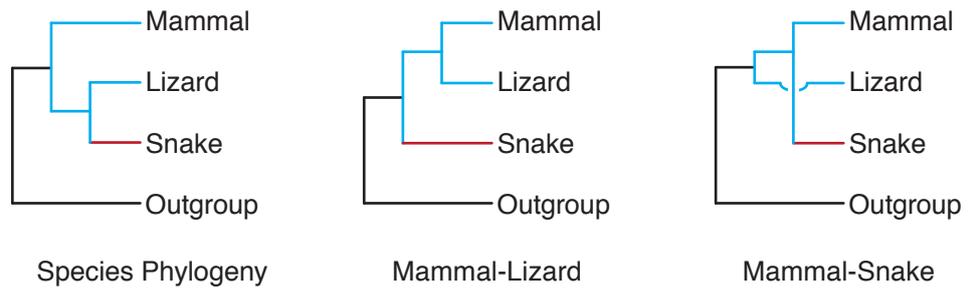


Figure 3.25: The three possible topologies for Lepidosauria. The tree topology will either show the species phylogeny, Mammal-Lizard or Mammal-Snake.

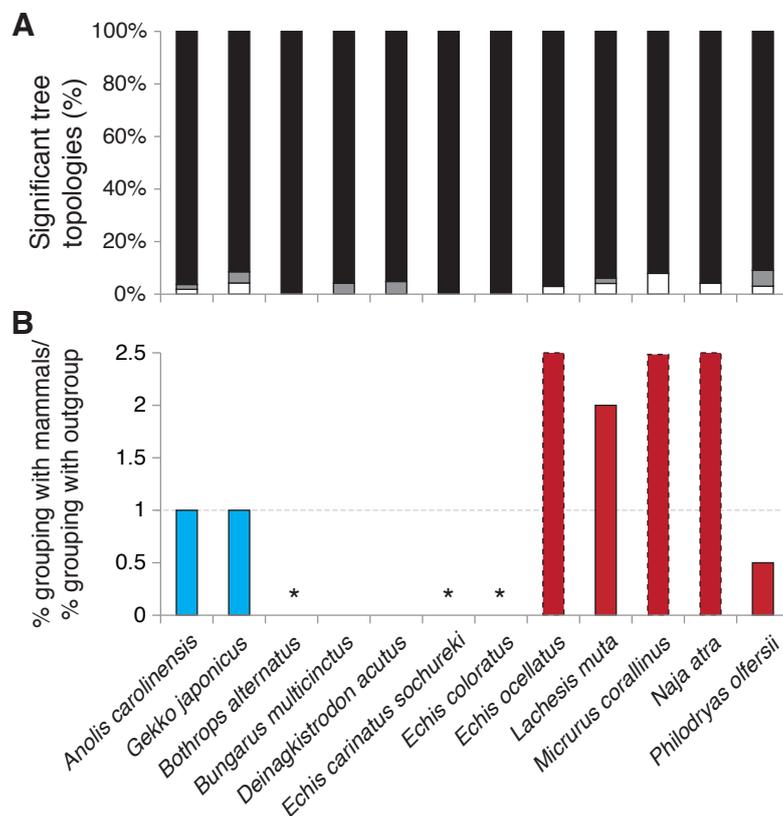


Figure 3.26: Bootstrap tree results for Lepidosauria. (A) shows the proportions of each significant topology, in black are the species phylogeny trees, grey shows the Mammal-Lizard topologies and white Mammal-Snake. The second panel (B) shows the likelihood of each species grouping with mammals when the tree is incongruent. The 2 lizard species are shown in blue, the snakes in red. *Echis ocellatus*, *Micrurus corallinus* and *Naja atra* only group with mammal when the trees are incongruent and as such have dashed outlines. * For these species there were no incongruent topologies.

from amphibians and actinopterygii: there are fewer incongruent trees and no consistent bias within them.

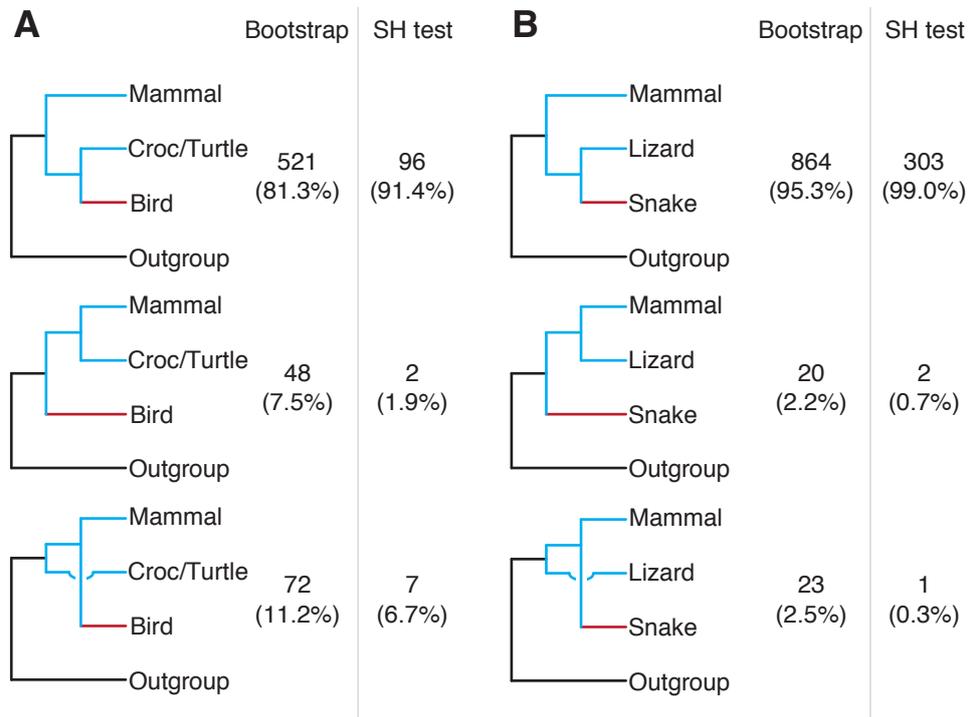


Figure 3.27: Summary of sauropsid trees. (A) shows the results for archosaurs, (B) lepidosaurs. The total number of significant trees, using either the bootstraps or SH test, are shown for each of the possible topologies.

3.2.4 Conclusion

Using a four-taxon methodology we have been able to build 55,305 trees, including our two novel transcriptomes. Of these, 28,987 were significantly supported by over 70% of bootstrap replicates, and 3,693 were significant using the SH likelihood test. In total 10,014 (34.5%) of the bootstrap supported trees were unable to recapitulate the species phylogeny. Within these we observed a bias in the amphibian and actinopterygian trees but not within the sauropsids. The bias that we did see tended to group the mammals with the species undergoing epigenesis, therefore grouping those that share this mechanism together. The taxon utilising preformation was commonly grouped with the outgroup in these incongruent trees. This suggests that there is a correlation between the mode of PGC specification and the incongruent tree topologies.

3.3 Distance Matrix

It follows that if the four-taxon trees are unable to recapitulate the species phylogeny, it may be due to differences in the sequence similarity between species. To test this we constructed distances matrices using the two sister taxa and the mammalian sequence. Figure 3.28 shows the three distances measured in the distance matrix. We then investigated which of the two sister taxa had the shortest distance to mammals, and which of the three distances was the smallest overall. We were able to construct more 3-taxon alignments than 4-taxon alignments, and therefore investigated the distances in a greater number of sequences.

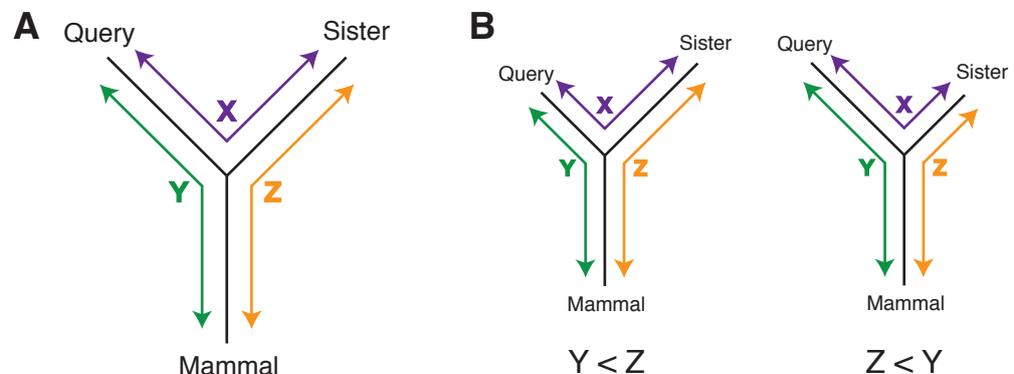


Figure 3.28: The three distances measured. (A) shows the three distances measured in a distance matrix, from the query to its sister taxon (X), the query to Mammals (Y) and the sister taxon to Mammals (Z). (B) shows two examples of different distances, where either $Y < Z$ or $Z < Y$.

The results for amphibians, actinopterygii and sauropsids all showed very similar results (Figure 3.29 and Figure B.6, Appendix page 252). In each case the taxon undergoing epigenesis showed a smaller distance to mammal than its sister taxon to mammal. The species which acquired preformation obviously show the opposite result.

Of the three distances we expect distance X to be the smallest overall as this is the distance between the most closely related species. However, in a surprisingly large number of distance matrices this is not the case. For those where distance X is not the shortest, it is more common for the distance from the epigenesis species to mammal to be the shortest overall. This is particularly clear in the anuran and teleost species (Figure 3.29B and D), where the short epigenesis-mammal distance is represented by a clear bar.

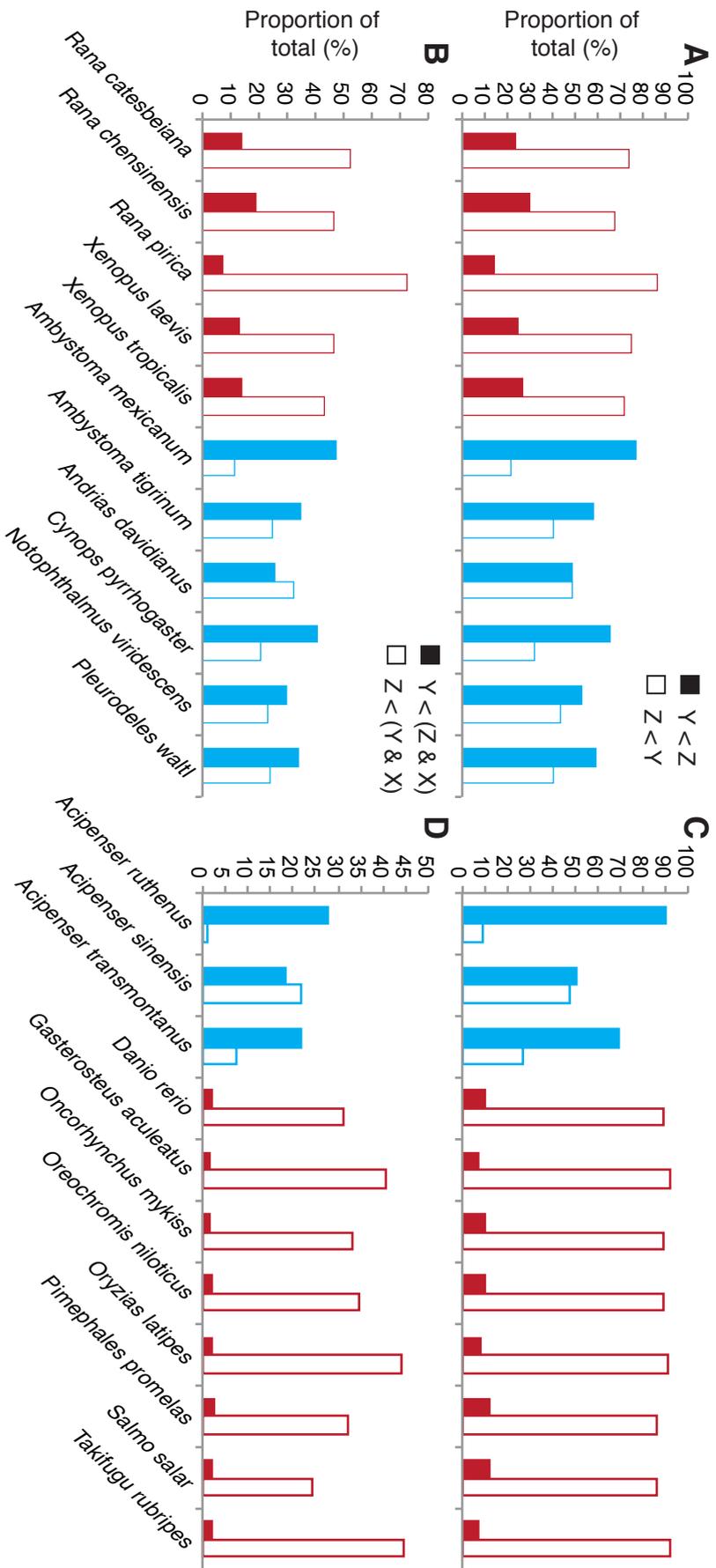


Figure 3.29: The amphibian and actinopterygian distance matrix results. The amphibians (**A** and **B**) and actinopterygian (**C** and **D**) species are coloured according to the mode of PGC specification, blue for epigenesis (urodeles and sturgeon), red for preformation (anurans and teleosts). The top panels (**A** and **C**) show the proportion of sequences where the query-mammal distance is less than the sister-mammal distance in filled bars ($Y < Z$), and the opposite in clear bars ($Z < Y$). (**B** and **D**) show the proportion of results where the query-mammal distance is the smallest overall in filled bars ($Y < (Z \& X)$), and the proportion where the sister-mammal distance is smallest overall in clear bars ($Z < (Y \& X)$).

There is clearly a dramatic difference in the distance to mammals between the sister taxa, much of which contradicts what we would expect from the species phylogeny. However, considering that the differences between the sister taxa have arisen from a fixed time point, this suggests that the distance matrix results are due to a change in the rate of evolution.

3.4 Relative Rate Test

We investigated whether a change in the rate of evolution occurred between sister taxa by performing a relative rate test (Kimura, 1980; Muse and Weir, 1992; Tajima, 1993). This compares the rate of molecular evolution between two sister taxa using a reference species (Figure 3.30). If the rates are determined to be significantly different, we can use the calculated branch lengths to deduce which species has a slower rate, and therefore which has a faster rate. To perform this test on each available sequence, we used the three-taxon alignments created earlier. Therefore each relative rate test used a mammalian sequence as the reference.

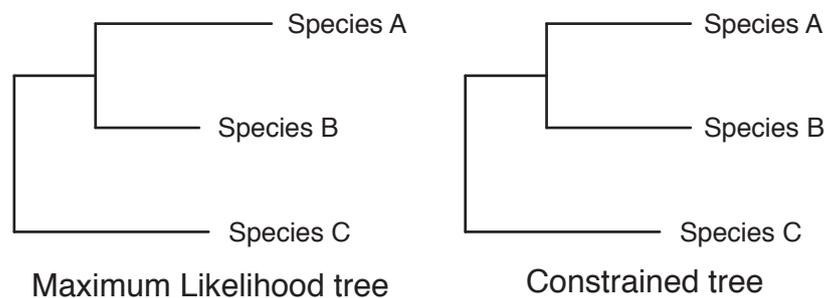


Figure 3.30: The relative rate test. The likelihood ratio version of the relative rate test (RRT) works by building a maximum likelihood tree of the three taxa; as well as a constrained tree where the two sister taxa (A and B) are evolving at equal rates. These two phylogenies are compared using a likelihood ratio test to determine whether the ML tree is a significantly better explanation of the data (Muse and Weir, 1992).

3.4.1 Amphibians and Actinopterygii

As before, within amphibians we are comparing anurans undergoing preformation against urodeles which retained epigenesis. In the actinopterygian analyses, we are comparing sturgeons, using epigenesis, against teleosts, which have

acquired germ plasm. For each species the number of sequences evolving at either a significantly faster or slower rate was calculated. The amphibian results, including the axolotl transcriptome, can be seen in Figure 3.31.

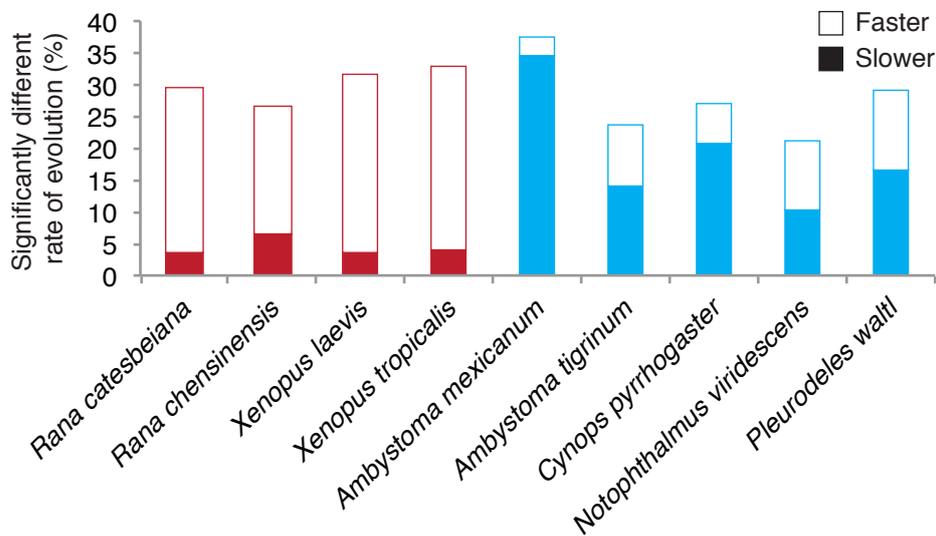


Figure 3.31: Amphibian relative rate test results. This figure shows for each species which had over 20 significant results, the proportion of those with a significantly faster rate of evolution in clear bars and the proportion with a significantly slower rate of evolution in filled bars. The anurans are coloured in red, the urodeles in blue.

For each species, approximately 30% of the sequences analysed were evolving at a significantly different rate to their sister taxon. Within the anurans, the vast majority of these sequences were evolving at a faster rate than in urodeles. Conversely, in urodeles the majority of sequences were evolving slower than in anurans. For the three urodele species with the fewest sequences, *Ambystoma tigrinum*, *Notophthalmus viridescens* and *Pleurodeles waltl*, this bias towards a slower rate was less obvious. In fact, for *Notophthalmus viridescens* there was one more sequence with a significantly faster rate than there was with a significantly slower rate (Table A.11, Appendix page 227). However the result for *Ambystoma mexicanum*, with the highest number of sequences tested, showed that 92.1% of those evolving at significantly different rates were slower than in anurans.

The RRT results from the actinopterygian comparison are shown in Figure 3.32, including the sturgeon transcriptome. In this case, the bias within those that are evolving at significantly different rates is dramatic. Almost all of the sequences show a significantly faster rate in the teleost ortholog than in the

sturgeon. This result is far more striking than what was observed in amphibians. There are almost no sequences which are evolving at a faster rate in sturgeons than in teleosts. This bias is consistent across all species with more than 20 significant results (Table A.12, Appendix page 227; Figure B.7, Appendix page 253).

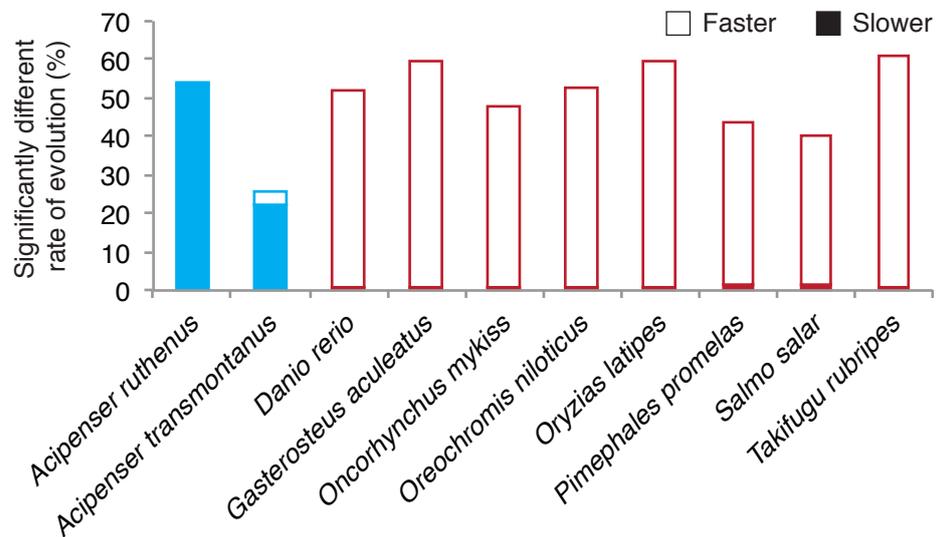


Figure 3.32: Actinopterygii RRT results. The results are shown in the same format as Figure 3.31. Sturgeon species are shown in blue, teleosts in red.

Summarising the data from both analyses (Figure 3.33) shows that in amphibians the majority of sequences (67.7%) have no significant difference in rate between urodeles and anurans. However, for those that do have a significant difference in rate there is a sharp bias towards the anuran evolving at a faster rate than the urodele. There are roughly seven times the number of sequences with a significantly faster rate in the anuran, compared to those with a faster rate in the urodele.

The actinopterygian sequences also show that the majority of sequences (53.1%) show no significant difference in the rate of evolution (Figure 3.33). However, those with a significant difference in rate show 93.4% are evolving faster in teleosts than in sturgeon. There are only 2,851 out of 91,650 sequences which show a significantly faster rate in the taxa that has retained epigenesis. Therefore the rates of evolution in both amphibians and actinopterygii correlate with the mode of PGC specification.

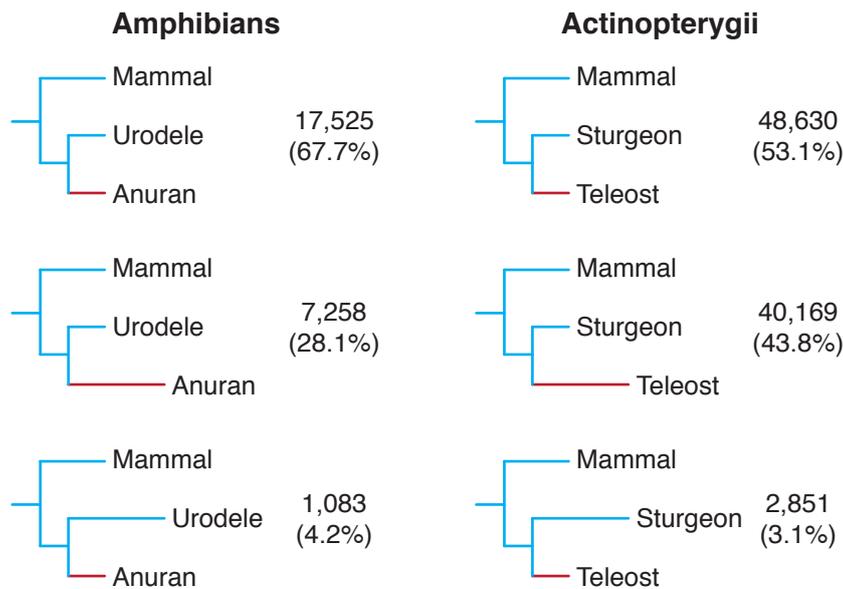


Figure 3.33: Summary of the amphibian and actinopterygian RRT results. The total number of sequences with no significant difference in rate are shown, as well as those with a significantly faster rate in with anuran, urodele, teleost or sturgeon.

3.4.2 Sauropsids

Although there was no bias within the incongruent sauropsid trees (see Section 3.2.3), a small bias was observed within the distance matrices. It was therefore particularly interesting to investigate whether there was any difference in the rate of evolution within sauropsids. As before the sauropsids were divided into the archosaurs, comparing birds against crocodiles and turtles, and the lepidosaurs where we compared lizards and snakes.

Figure 3.34 shows the results for the archosaurs and testudines. Although not as clear as in amphibians and actinopterygii, there is a bias within those that are evolving at significantly different rates. For each species, there are more sequences that are evolving at a faster rate in birds (and therefore slower in crocodiles and turtles) than the opposite way around. This bias appears clearer in the turtle (*Trachemys scripta*) than in the crocodile (*Alligator mississippiensis*). As in the 4-taxon trees, there are far fewer archosaur and testudine sequences available to analyse than in amphibians and actinopterygians (Table A.13, Appendix page 228).

The lepidosaur results are shown in Figure 3.35: as in all other vertebrate groups there is a bias within those sequences that are evolving at significantly different rates. Only three of the 16 snake species have more than 20 significant

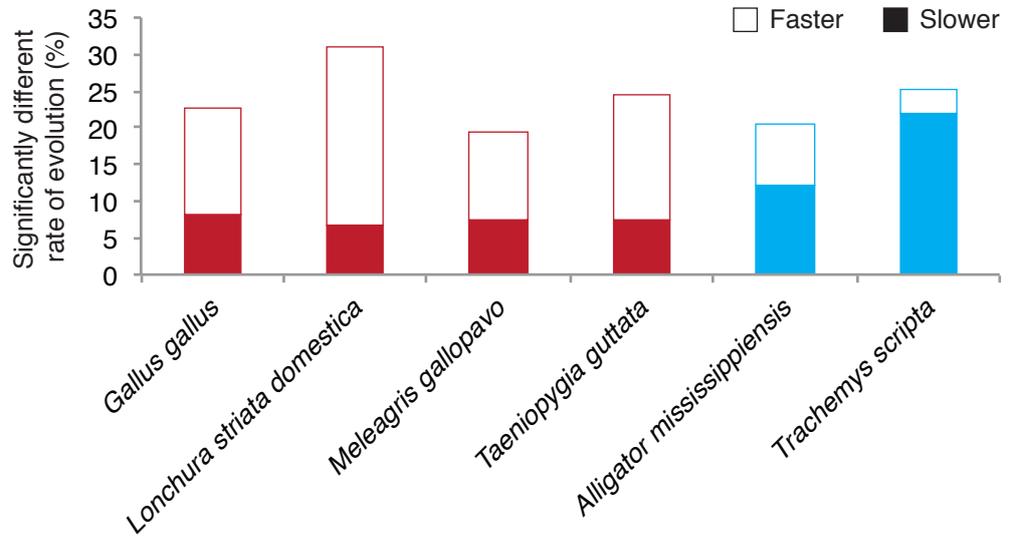


Figure 3.34: The archosaur and testudine RRT results. The results are shown in the same format as Figure 3.31. The birds are shown in red, and the turtle and crocodile species in blue.

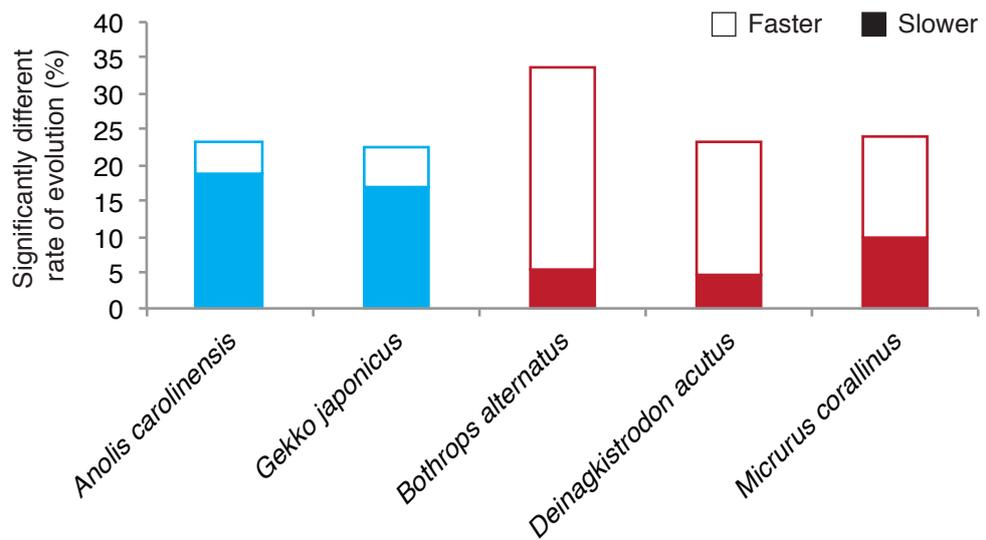


Figure 3.35: The lepidosaur RRT results. The results are shown in the same format as Figure 3.31. The lizards are shown in blue, snakes in red.

results (Table A.14, Appendix page 228) although the average proportion that are evolving significantly differently is not that dissimilar to amphibians. However, all of the species in Figure 3.35 show that there are more sequences with a significantly faster rate in snakes than lizards.

3.4.3 Conclusion

In total, across all the vertebrate species analysed, we ran 121,382 relative rate tests. Of these 52,217 (43.0%) showed a significant difference in rate of evolution between the two sister taxa. Of those with a significant difference in rate, 95.6% (49,898) showed a faster rate of evolution in the species that has acquired preformation. This left only 2,319 sequences that showed a faster rate in the species that had retained epigenesis (Figure 3.36).

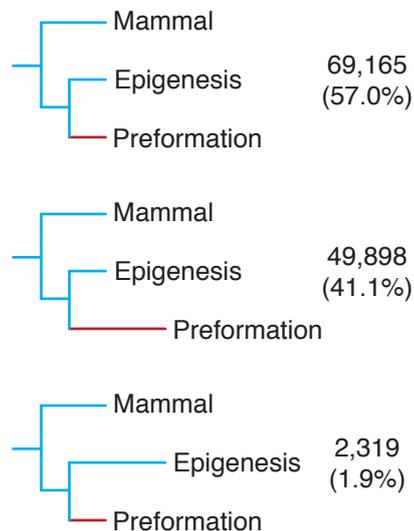
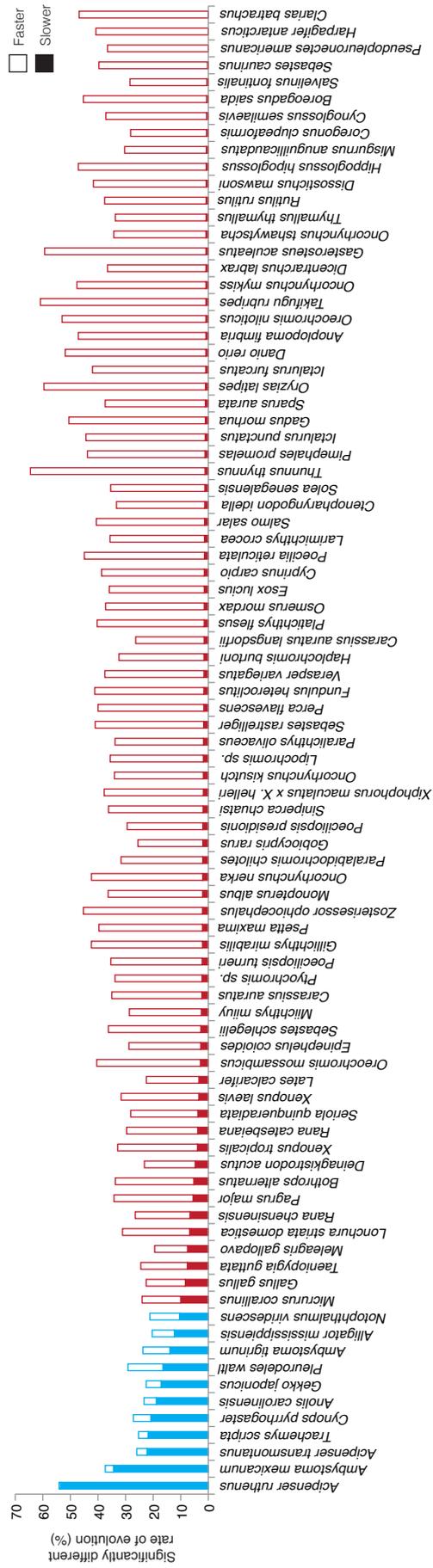


Figure 3.36: Summary of the Relative Rate Test results. This shows the total number of sequences from across vertebrates that had shown the three possible results from the relative rate test, either no significant difference or either a significantly faster or slower rate in the species undergoing preformation.

Not only was this result evident in the combined vertebrate results but it was consistent across all of the orders analysed, unlike the 4-taxon tree results. In fact, when we order all vertebrate species according to the proportion of sequences evolving significantly slowly, we see a perfect division between those that have retained epigenesis and those that have acquired preformation (Figure 3.37). There is therefore a strong correlation between the mode of PGC specification and the rate of molecular evolution across many genes.



3.5 Rate of evolution and tree topology

Summarising the relative rate test results onto a vertebrate tree reveals significantly longer branches for taxa using preformation (Figure 3.38). These long branches may be affecting phylogenetic incongruence through Long Branch Attraction (LBA; Anderson and Swofford, 2004; Felsenstein, 1978; Sanderson et al., 2000). This is a tree building artefact that results in long branches incorrectly grouping together.

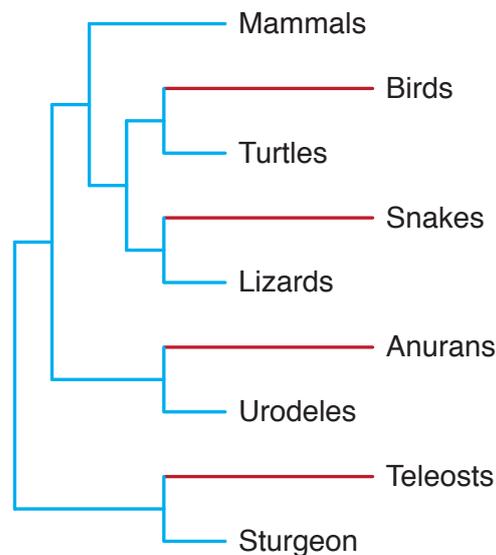


Figure 3.38: The relative rate test revealed long branches across vertebrates. We saw an increase in the rate of evolution for many sequences in those species that had acquired preformation, namely birds, snakes, anurans and teleost fish.

If this is the case then we might expect to see an increase in the incongruent tree topologies when the sequences are evolving at different rates. Figure 3.39 shows that this is true in amphibians. Not only is there an increase in the proportion of incongruent tree topologies but they are biased according to which species is evolving significantly faster. If the anuran species is evolving faster then there is an increase in Mammal-Epigenesis trees, i.e. where the anuran is grouped with the other long branch, the outgroup. Conversely, in the few cases when the urodele is evolving significantly faster there is an increase in the proportion of Mammal-Anuran tree topologies.

Furthermore, it follows that if the rate of evolution is affecting the tree topologies then if we replace a long branch with a shorter one it might shift the topologies towards the species phylogeny (Figure 3.40). Since all amphibian

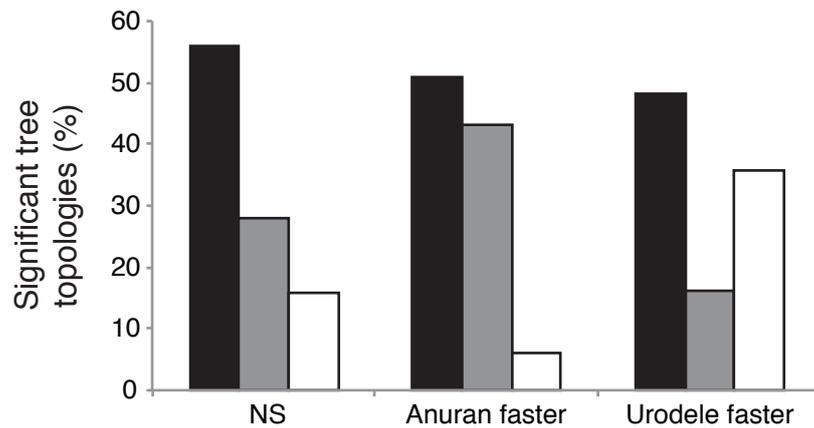


Figure 3.39: Amphibian tree topologies according to the RRT results. The black bars show the proportion of species phylogeny trees, the grey bars show Mammal-Urodele topologies and the white shows Mammal-Anuran. The results are divided according to the RRT results, either no significant difference in rate, or either anurans or urodeles evolving significantly faster. All tree topologies are considered significant using the bootstrap test.

4-taxon trees were built using a teleost outgroup, we can replace this sequence with a slower evolving sturgeon sequence. This will leave the tree with only one long branch and the lack of LBA might allow the species phylogeny to be ‘rescued’.

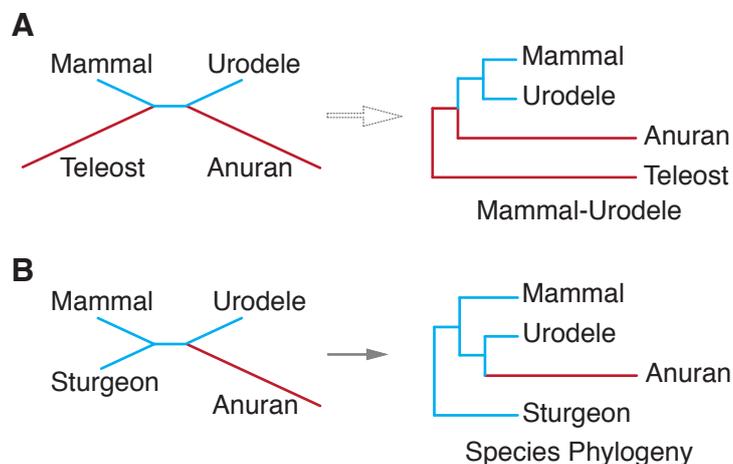


Figure 3.40: Changing the outgroup may ‘rescue’ the tree topology. If the rate of evolution is affecting the tree topologies than we may be able to reduce the incongruent tree topologies by replacing the teleost outgroup with a sturgeon sequence. If both the anuran and teleost sequences are evolving significantly faster (**A**), replacing one of those long branches with a shorter one (**B**) might remove the LBA and recover the species phylogeny.

We tested this theory by rebuilding all the amphibian 4-taxon trees with their sturgeon ortholog, if available. We also used the information from the

actinopterygian RRT to determine whether the relative rate of evolution between teleost and sturgeon was significantly different. The four main classes of rate difference, in both the amphibian and actinopterygii, are shown in Figure 3.41. For each of these we have plotted the proportion of tree topologies, found to be significant using bootstrapping, using either the teleost or sturgeon outgroup.

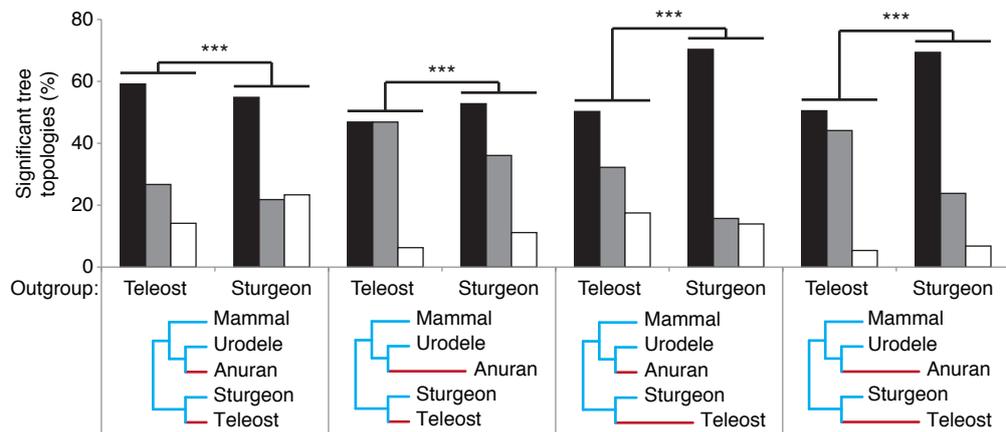


Figure 3.41: The tree topologies according to the RRT results in Amphibians and Actinopterygii. The 4-taxon tree results are shown as before, species phylogeny in black, Mammal-Urodele in grey and Mammal-Anuran in white. The results are divided according to the RRT, either no significant difference, a faster rate in anurans, a faster rate in teleosts or the latter two combined. The bootstrap results are shown using both the teleost and sturgeon outgroup. The number of trees in each category are shown in Table A.15 (Appendix, page 229). The difference between the sturgeon and teleost outgroup trees was tested using the Chi-squared test (** $p < 0.001$; 2 d.f.).

Figure 3.41 shows that when there is no significant difference in rate between the teleost and sturgeon sequences then changing the outgroup does not obviously affect the proportion of incongruent tree topologies, although the overall proportions are significantly different by the Chi-squared test. When the rate difference between teleost and sturgeon is significant then changing the outgroup to sturgeon leads to an increase in species phylogeny trees and a reduction in Mammal-Urodele topologies. This suggests that it is the long branches in anuran and teleosts that are resulting in the large proportion of Mammal-Urodele topologies.

The tree topologies significant by the SH test (Figure 3.42) still show a correlation between the faster rate in anurans and the Mammal-Urodele topology. For the sequences with a significantly faster rate in both anurans and teleosts,

there is a clear change in the tree topology after swapping outgroup. The majority of incongruent trees are no longer significant and there is an increase in the proportion of species phylogeny trees.

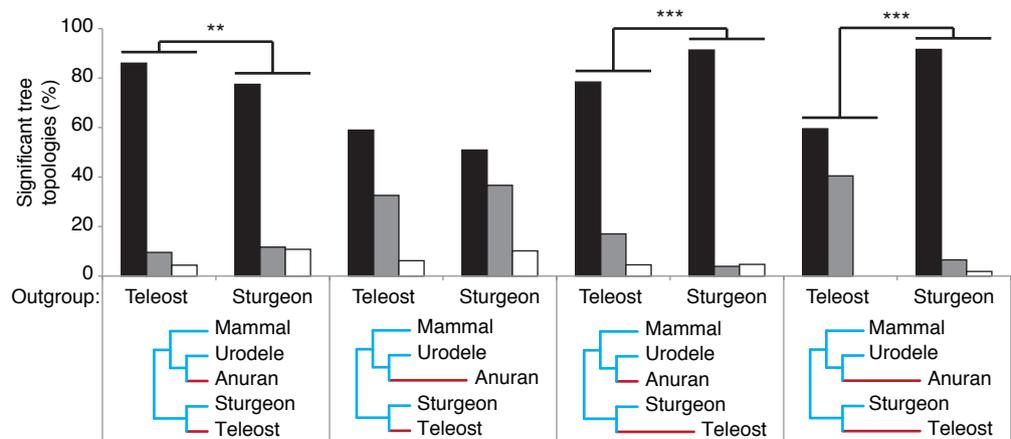


Figure 3.42: The SH test results according to the RRT results in amphibians and actinopterygii. The results are shown in the same format as Figure 3.41. The number of trees in each category are shown in Table A.16 (Appendix, page 230). Once again, the differences between the choice of outgroup were tested using the Chi-squared test (** $p < 0.01$, *** $p < 0.001$; 2 d.f.).

Overall these results suggest that the incongruent tree topologies observed in Section 3.2 are an artefact of the increased rate of evolution in species that have acquired preformation. It is the change in rate of evolution that is the key difference between sister taxa that have diverged in their mode of PGC specification.

3.6 Conclusion

We have compiled a transcriptome dataset from across vertebrates and identified orthologs. Using these data we have devised methods to test sequence evolution on a large scale. For each orthologous group we have built 4-taxon trees, distance matrices and tested the rate of evolution. In each of these cases we have specifically compared two sister taxa, one that has retained epigenesis and one that has acquired a preformation mode of PGC specification. These sister taxa consist of species from sauropsids, amphibians and actinopterygian fish.

Our first task was to assess the quality of the alignments and remove any that were unreliable, to this end we devised five parameters to remove poor

quality alignments. These parameters were based on alignments that were either too similar, too dissimilar or that were too short. We went on to identify the protein coding regions in all of our sequences. In doing this it emerged that the third codon position was having a large effect on our analyses, specifically in terms of increasing the number of incongruent tree topologies. It appeared as though this was caused by differences in codon usage, specifically a GC bias in the third codon position. We therefore removed the third codon position from our alignments prior to analysis.

The first analysis of sequence evolution was to create 4-taxon trees and assess whether they reflected the species phylogeny or were incongruent. We began this because of a bias within trees in the literature that showed a large proportion of Mammal-Urodele trees, in other words trees that grouped the two species undergoing epigenesis together. In our large scale analysis across amphibians and actinopterygian fish we saw that this bias was present in all the species we analysed. In each case the majority of trees showed the species phylogeny; but the incongruent trees were biased towards grouping the two species undergoing epigenesis together. In sauropsids we saw a different pattern, starting with an increase in the proportion of species phylogeny trees. Within the few incongruent sauropsid trees we were unable to see any bias that was consistent in all of the species.

We then tested the rates of evolution by running the relative rate test on 3-taxon alignments. This showed that when there was a significant difference in rate, there was a strong bias towards the preformation species evolving faster than its sister taxon that had retained epigenesis. This bias was consistent across all the species analysed in amphibians, actinopterygian fish and sauropsids. In fact, when the species were sorted according to the proportion evolving significantly slower, it perfectly differentiated between the two modes of PGC specification.

This strong correlation between the acquisition of preformation and a faster rate of sequence evolution was affecting the 4-taxon tree topologies. When there was a significant difference in rate between the sister taxa, it increased the proportion of incongruent tree topologies. We were able to recover more species phylogeny trees when we replaced the outgroup sequence for one undergoing epigenesis, i.e. with a shorter branch length. Therefore our initial 4-taxon trees

were an artefact, caused by long branch attraction resulting from the change in rate of sequence evolution. The results from this body of work were recently published (Evans et al., 2014).

Characterising Genes with a Change in Molecular Evolution

We have identified a number of sequences with incongruent phylogenies which appear to be driven by differences in the rate of evolution, but do not know what these genes are. The incongruent trees and differences in rate both correlate with the mode of PGC specification and so we might expect an overabundance of genes which regulate germ cell specification. Conversely, the hypothesis of constraint and constraint release suggests that genes which regulate somatic development will undergo constraint release (Johnson et al., 2011). We might therefore observe an association between somatic regulatory functions and the results from the 4-taxon trees and relative rate test.

To characterise the genes with an incongruent phylogeny or significantly different rate of evolution, we sought to annotate our sequences with their function, expression profile and evolutionary history. However, assigning this information onto novel sequences, many of which are derived from un-annotated ESTs, is nontrivial. It is possible to assign functional GO-terms using software such as Blast2Go (Conesa et al., 2005), but this is based entirely on similarity to annotated proteins and not experimental verification. Even if we were able to annotate all of our sequence dataset, identifying shared functions between 50,000 genes would be near to impossible.

We attempted to solve this problem by mapping our sequences onto a single, well annotated genome. We performed this mapping procedure three times, to the genomes of mouse, *Xenopus tropicalis* and zebrafish. This information was then used to investigate whether genes with a Mammal-Epigenesis topology or Preformation-faster relative rate result were associated with a particular function, expression profile or gene age.

4.1 Mapping to a Single Genome

Of our 1,210,525 query sequences in amphibians, actinopterygii and sauropsids we were able to find one-to-one orthologs for 484,754 (40%). The number of mouse, *Xenopus* and zebrafish genes that we were able to map to are shown in Table 4.1. These data show that we were able to map our sequences onto almost all of the genes in the respective genomes (see Table 2.2, page 36). Approximately half this number had mapped orthologs analysed in the 4-taxon trees or relative rate test (RRT).

Table 4.1: Genes with a one-to-one ortholog in our dataset.

Species	Genes mapped	w/Tree result	w/RRT result
<i>Mus musculus</i>	18 841	7928	11 072
<i>Xenopus tropicalis</i>	17 212	7402	10 559
<i>Danio rerio</i>	24 621	7929	12 110

An additional benefit of mapping to a single genome is that we are able to summarise our results across vertebrates without counting genes multiple times. For example, in Chapter 3 a gene might have been analysed in multiple species of amphibian and sauropsid. When mapped to a single genome we can count the results for this gene once, instead of in each ortholog studied. However this implies that all orthologs showed the same result in the trees or RRT, which was not always the case. We therefore developed methods which assign a result to each gene based on the mapped orthologs.

The first of these was based on finding all genes which in at least one ortholog had an incongruent tree or a significantly different relative rate of evolution. If a gene showed opposing incongruent topologies (Mammal-Epigenesis in one ortholog, Mammal-Preformation in another) or contradictory significant differences in rate, the gene was marked as ambiguous and excluded from further analysis. Using the relative rate test results we excluded 735 mouse, 706 *Xenopus* and 646 zebrafish ambiguous genes. The second method involved choosing a single result from within our mapped sequences, in this case we used the sequence with the longest ORF alignment as the representative for each gene.

The bootstrap results when mapped to the genomes are shown in Figure 4.1. They show that when the tree is incongruent there is still a bias towards the

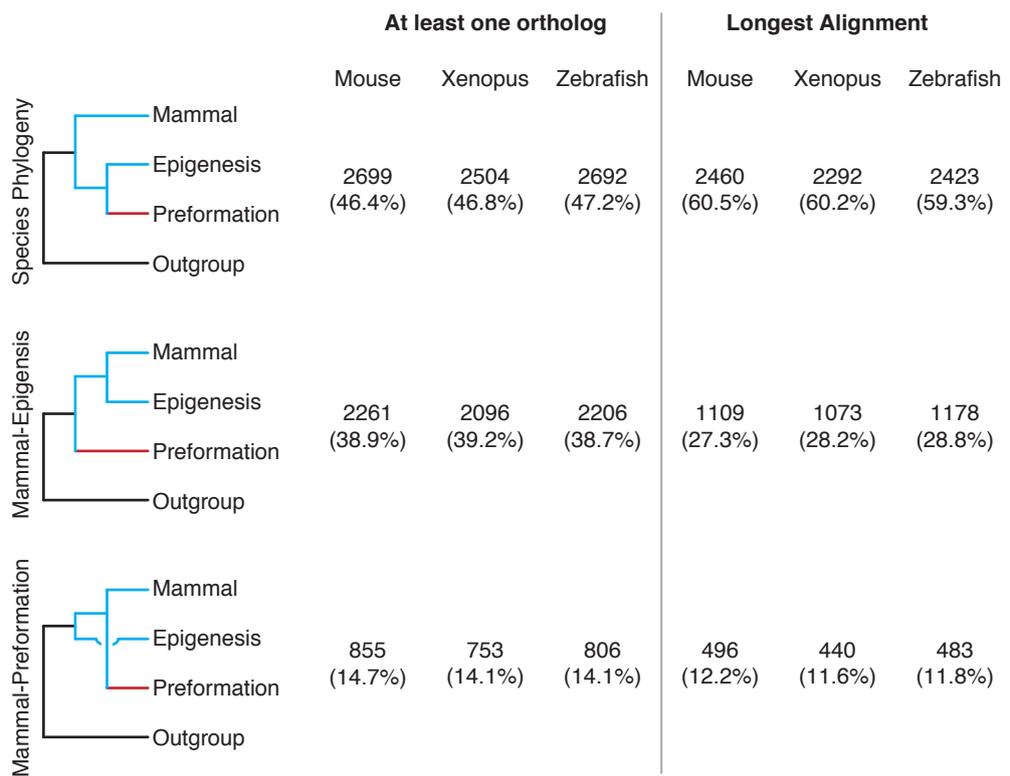


Figure 4.1: Mapped Bootstrap results. The vertebrate bootstrap results are mapped to the genomes of mouse, *Xenopus* and zebrafish. The results are allocated to the genes using two approaches, the first selects all genes where at least one ortholog has an incongruent tree. The second method uses the result from the vertebrate ortholog with the longest alignment.

Mammal-Epigenesis topology, as we saw in Chapter 3. This suggests that our previous results were not due to a few over-represented genes. The proportions of significant trees do not differ to any real extent between the three different species. As expected, there are higher numbers of incongruent trees using the first mapping method than the second. When assigning a result using the ‘at least one ortholog’ approach approximately 55% of the genes have an incongruent topology, this drops to 40% when using the ‘longest alignment’ method.

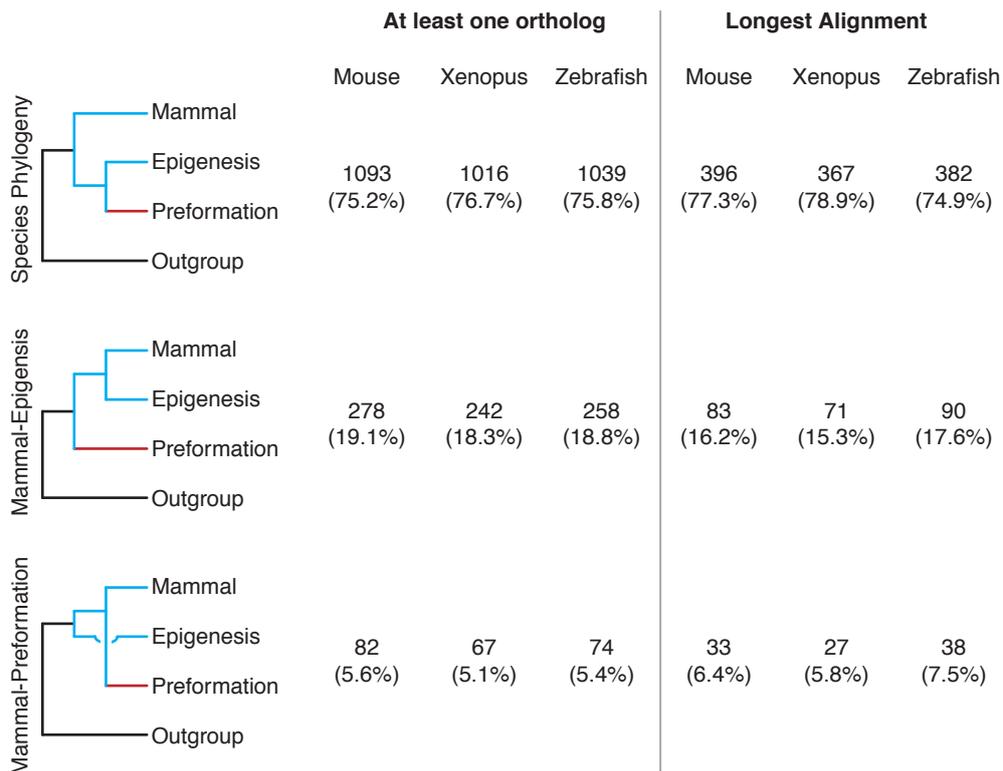


Figure 4.2: Mapped SH-test results. As in Figure 4.1, the tree topology results significant by the SH-test are mapped to the genomes of mouse, Xenopus and zebrafish using the two methods.

The SH-test (Figure 4.2) identifies fewer genes with a significant result than the bootstrap test. This drop in number is particularly evident when we use the ‘Longest Alignment’ method. This suggests that many of the trees with the longest alignments have non-significant topologies according to the SH-test. Even so, the bias within the incongruent topologies is still evident with approximately 17% of significant trees grouping the species undergoing epigenesis together.

Figure 4.3 shows the relative rate test (RRT) results. When mapped to the mouse genome, 7,555 genes have at least one vertebrate ortholog with a significantly faster rate of evolution in the organism utilising preformation compared

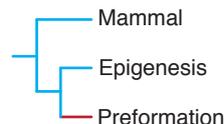
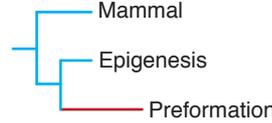
	At least one ortholog			Longest Alignment		
	Mouse	Xenopus	Zebrafish	Mouse	Xenopus	Zebrafish
	2375 (23.0%)	2306 (23.4%)	2672 (23.3%)	6335 (57.2%)	6003 (56.9%)	6543 (54.1%)
	7555 (73.1%)	7178 (72.8%)	8344 (72.8%)	4397 (39.7%)	4230 (40.1%)	5189 (42.9%)
	407 (3.9%)	369 (3.7%)	448 (3.9%)	337 (3.0%)	323 (3.1%)	372 (3.1%)

Figure 4.3: Mapped RRT results. As in Figure 4.1, the relative rate test results are mapped to the genomes of mouse, *Xenopus* and zebrafish using the two methods. The results are divided according to no significant difference in rate or a significantly faster rate in the ortholog utilising either preformation or epigenesis.

to its sister taxon. This drops to 4,397 using the alternative method of mapping, suggesting that a large proportion of the longest 3-taxon alignments show no significant difference in their rate of evolution. Even so, there is still a strong bias (approximately 12 fold) towards the preformation ortholog evolving significantly faster than its sister taxon.

To simplify the following analyses on the function and expression of these genes, as well as maximising the number of genes we can study, I will only present the data for those identified using the ‘at least one ortholog’ method. This permits us to analyse all genes which at some point have a significant incongruent phylogeny or difference in rate. I will also only show the data using mouse and zebrafish from here on as they have the highest number of annotated genes and represent a species utilising epigenesis and preformation respectively.

4.2 Gene Function

We asked whether genes with incongruent tree topologies or differing rates of evolution were functionally associated with the specification of germ cells or

early somatic development. To test this we used the known Gene Ontology (GO) terms for each gene to look at global patterns of function (Section 2.8, page 52). We also used a direct approach to assess the reliability of these results.

4.2.1 Gene Ontology

Using the program GOrilla, we compared the mouse genes with a Mammal-Epigenesis tree topology ortholog against those that only showed the species phylogeny, Table A.17 (Appendix, page 231). This showed only one over-represented molecular function GO-term ('GO:0003723 RNA binding'). There were 19 biological process terms with a significant p-value, of which 6 had a significant corrected FDR. These terms were all associated with metabolic processes. Eleven cellular component terms were significantly over-represented although none were specific (e.g. 'GO:0043226 organelle' and 'GO:0044424 intracellular part').

The GO-term results comparing the mouse genes with a preformation-faster RRT result against a background list of genes with no significant difference in rate are shown in Table A.18 (Appendix, page 231). Thirty nine GO-terms were significantly over-represented, most of which are involved in nucleoside binding and catalytic activity. There were 65 over-represented biological process GO-terms, the most significant one being 'GO:0044267 cellular protein metabolic process'. Of the 26 cellular component terms over-represented, most of the terms related to organelles.

We also looked for GO-terms which are significantly under-represented in mouse genes with preformation-faster results by swapping the target and background lists around. Table A.19 (Appendix, page 231) shows these results for molecular function, biological process and cellular component GO-terms. There were 48 significant molecular function terms, including the two transcription factor terms 'GO:0001071 nucleic acid binding transcription factor activity' and 'GO:0003700 sequence-specific DNA binding transcription factor activity'. There were also a lot of terms associated with transport and channel activities. Of the 75 biological process terms, many were associated with transport as well as 'GO:0007275 multicellular organismal development'. The cellular component terms were mostly associated with extracellular regions and channel

complexes, linking to the transport molecular function and biological process GO-terms.

These data suggest that some of the developmental regulator functions predicted to be associated with genes with a significant difference in rate are actually under-represented. Conversely there are no particular GO terms over-represented in the genes with a Mammal-Epigenesis topology or Preformation-faster RRT result. Instead, these genes appear to have a wide breadth of functions.

4.2.2 Transcription Factors

Many developmental regulators are transcription factors, and so we were surprised to see these GO terms under-represented in our genes of interest. We therefore used a different method to assess the reliability of these results, namely identifying transcription factors using BLAST. To create the BLAST database, we downloaded the known transcription factors from the transcription factor encyclopedia (accessed 10/05/2012; Yusuf et al., 2012). This resulted in 2,791 transcription factors from human, mouse, chicken, *Xenopus* and zebrafish. We blasted our query sequences against this database and mapped the results onto the single genomes, identifying 1,341 mouse and 1,877 zebrafish transcription factors.

Figure 4.4 shows the results for the 4-taxon trees and RRT results mapped onto the mouse and zebrafish transcription factors. As we saw previously there is a bias towards Mammal-Epigenesis tree topologies and preformation-faster relative rate test results. However, comparing these results to Figures 4.1-3 shows that the transcription factor trees have an increased proportion of Mammal-Preformation topologies. There is also a decrease in the number of Preformation-faster genes. The differences between the transcription factors and all genes are significantly different in Mouse for all tests, and significantly different for Zebrafish using the bootstraps and RRT (Chi-squared test; $p < 0.05$; 2 d.f.).

These data therefore substantiate what we saw using the GO-terms; transcription factors are significantly under-represented with Mammal-Epigenesis or Preformation-faster results relative to the whole population of genes.

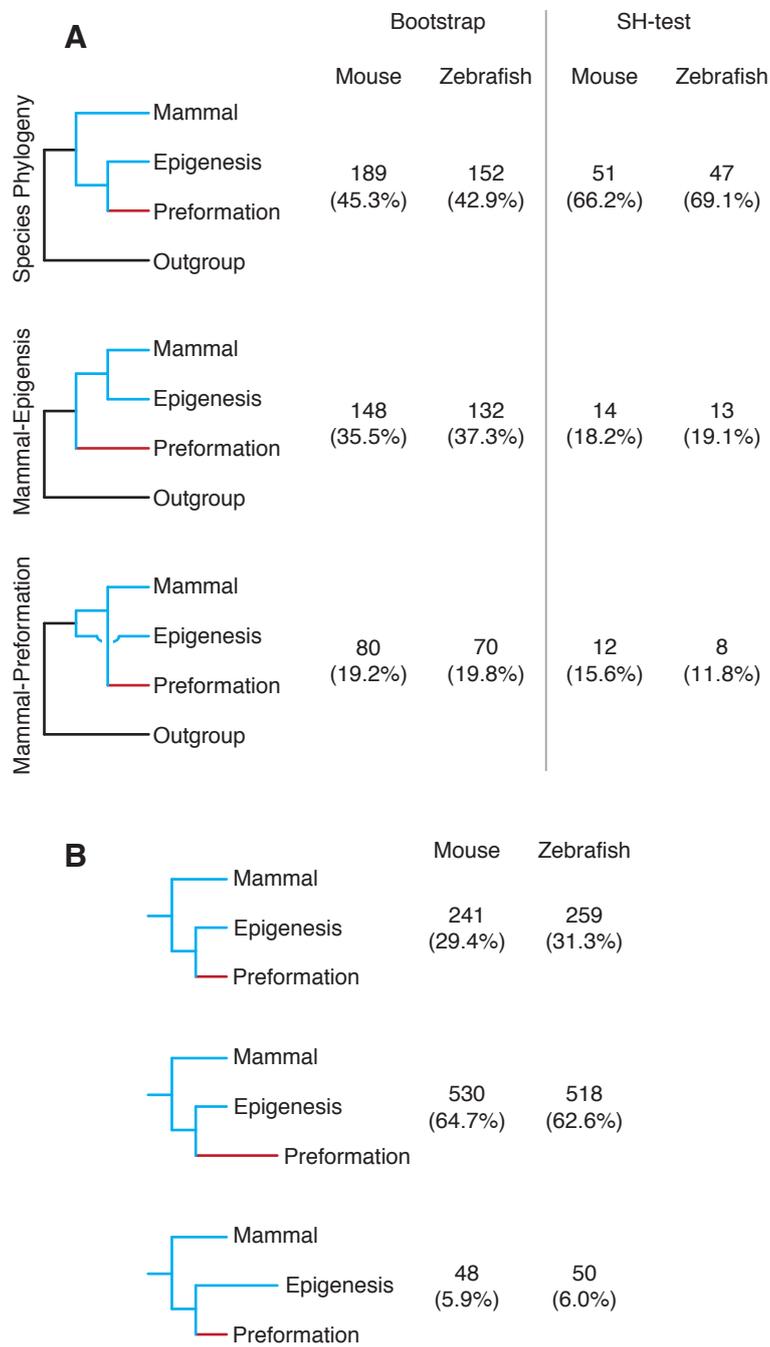


Figure 4.4: The tree and rate results for genes identified as transcription factors. The bootstrap, SH-test (A) and RRT results (B) mapped to the mouse and zebrafish transcription factors.

4.3 Gene Expression

As well as studying the function of mapped genes we looked at their expression profile. We investigated whether any time or location of expression was associated with either a Mammal-Epigenesis topology or a significantly faster rate of evolution in the ortholog undergoing preformation.

4.3.1 Time of Expression

To measure the time of expression, we first plotted the known expression data for the 18,841 mouse genes (Figure 4.5). The developmental stage with the highest number of genes known to be expressed is stage 23, late in development, approximately embryonic day (E)15. The stages with the least information are stages 6-8, this is during implantation and therefore the most difficult stages to work on experimentally.

Figure B.8 (Appendix, page 254) shows the number of genes expressed at each stage of zebrafish development. This shows that the stages with the highest number of expressed genes are late in development, just prior to hatching. There are also a large number of genes with no known expression data.

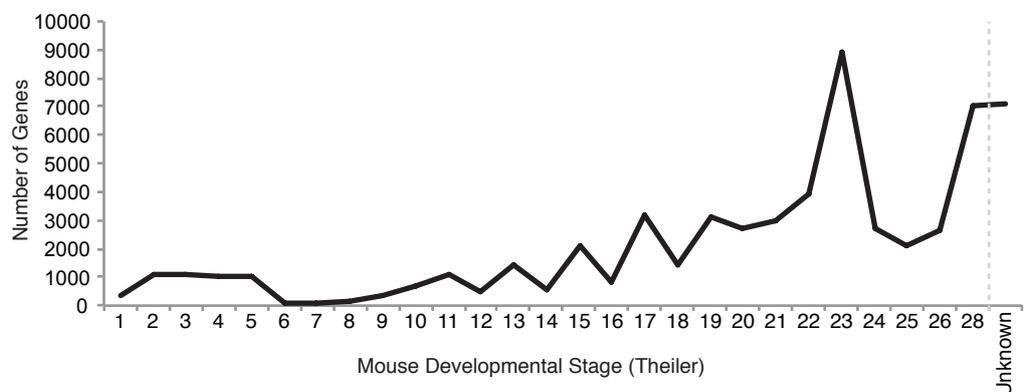


Figure 4.5: The number of mouse genes known to be expressed at each stage. For each stage of mouse development (Theiler, 1989) we have plotted the number of expressed genes. We have also plotted the number of genes with no known expression.

To assess whether there is an association between the time of expression and our results, we counted the proportion of genes expressed at each developmental stage with a particular tree topology or relative rate result. We also calculated the proportion for the genes with unknown expression data, and for all genes tested. We then compared each value to the proportion of all genes

with known expression data and assessed the difference using the chi-squared test.

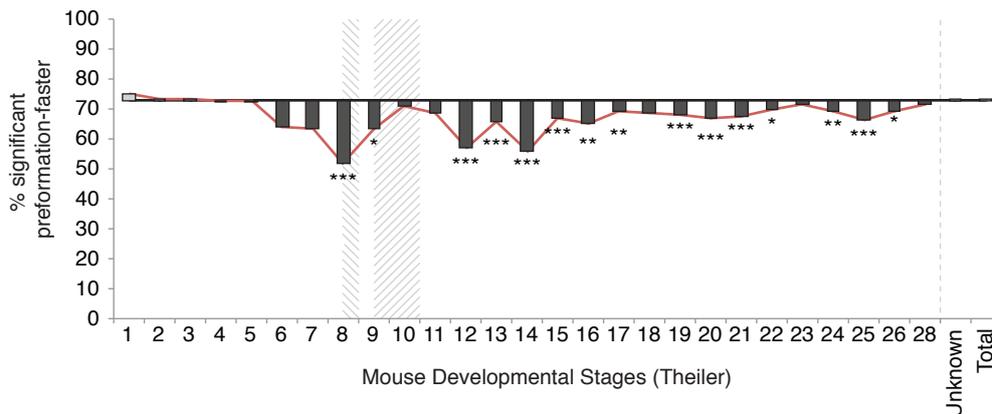


Figure 4.6: The proportion of mouse genes with a preformation-faster relative rate result expressed at each developmental stage. For each stage of development the proportion of expressed genes with a preformation-faster RRT result are shown. The horizontal line represents the average for all genes with known expression information. The value at each stage is compared to the average using the Chi-squared test (Bonferroni corrected; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$; 1 d.f.). The stages of PGC induction (late TS 8) and gastrulation (TS 9.5-10) are hashed.

Figure 4.6 shows that there is no over or under-representation of genes with a preformation-faster RRT result during the first stages of development. From the point where PGCs are induced (TS 8), the proportion of preformation-faster genes drops considerably and is significantly less than the average for almost all of the remaining stages of development. This suggests that the sequences evolving significantly faster in taxa that have acquired preformation are orthologs to the mouse genes typically expressed at the earliest stages of development.

The equivalent graph for zebrafish is shown in Figure 4.7, in this case there is significant over-representation of genes with a Preformation-faster result during the earliest stages. This value steadily drops through development until hatching, where it suddenly drops to below average. Therefore in zebrafish as well as in mice there is an over-representation of Preformation-faster genes during the earliest stages of development. However, in zebrafish this over-representation continues beyond gastrulation whereas in mouse we see significant under-representation once PGCs are specified.

Using the SH-test results mapped to mouse, the total number of genes available to analyse drops (Figure 4.8). Mouse genes with orthologs that have a Mammal-Epigenesis topology appear over-represented between TS 2-4 but this

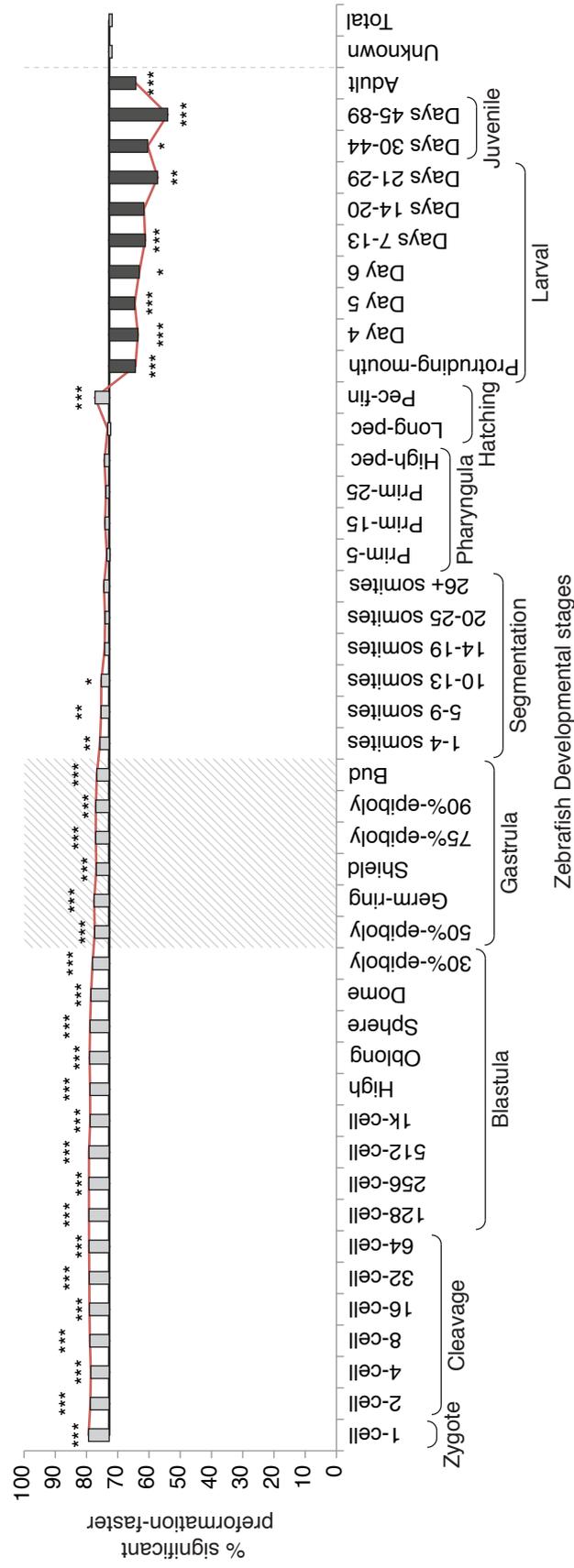


Figure 4.7: The proportion of zebrafish genes with a preformation-faster RRT result expressed at each developmental stage. This graph is the equivalent to Figure 4.6 but for zebrafish genes instead of mouse. The difference between the proportion at each stage and the average is deemed significant using the Chi-squared test (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 1 d.f.). Zebrafish gastrulation is hashed.



Figure 4.8: The proportion of mouse genes with a Mammal-Epigenesis SH-test result expressed at each developmental stage. For each stage of development the proportion of expressed genes with orthologs that produced a Mammal-Epigenesis topology significant by the SH-test are shown. The horizontal line represents the average for all genes with known expression information. The value at each stage is compared to the average using the Chi-squared test (Bonferroni corrected; $*p < 0.05$; 1 d.f.). The stages of PGC induction (late TS 8) and gastrulation (TS 9.5-10) are hashed. Stages with fewer than 20 total genes are shown as dashed outlines.

is not significant after correcting for multiple testing. The remaining later stages of embryo development show no significant over or under-representation of Mammal-Epigenesis genes. Using the bootstrap results, the over-representation persists until the PGCs are specified, after which the developmental stages are under-represented with Mammal-Epigenesis results (Figure B.9; Appendix, page 255). However, none of these stages have a value that significantly differs to the average.

The zebrafish results using the Mammal-Epigenesis topologies show the same pattern as the relative rate test results. The trees significant by bootstrap (Figure 4.9) show a significant over-representation of Mammal-Epigenesis topologies in the earliest stages of development which drops to a significant under-representation after hatching. This same pattern is shown for the trees significant by the SH-test but the differences between each stage and the average are no longer significant (Figure B.10; Appendix, page 256).

Overall we have shown that the genes of interest (those with a vertebrate ortholog with a Mammal-Epigenesis topology or evolving significantly faster in the Preformation taxa), are typically expressed in early development of both mouse and zebrafish. In mouse this over-representation ends once the PGCs

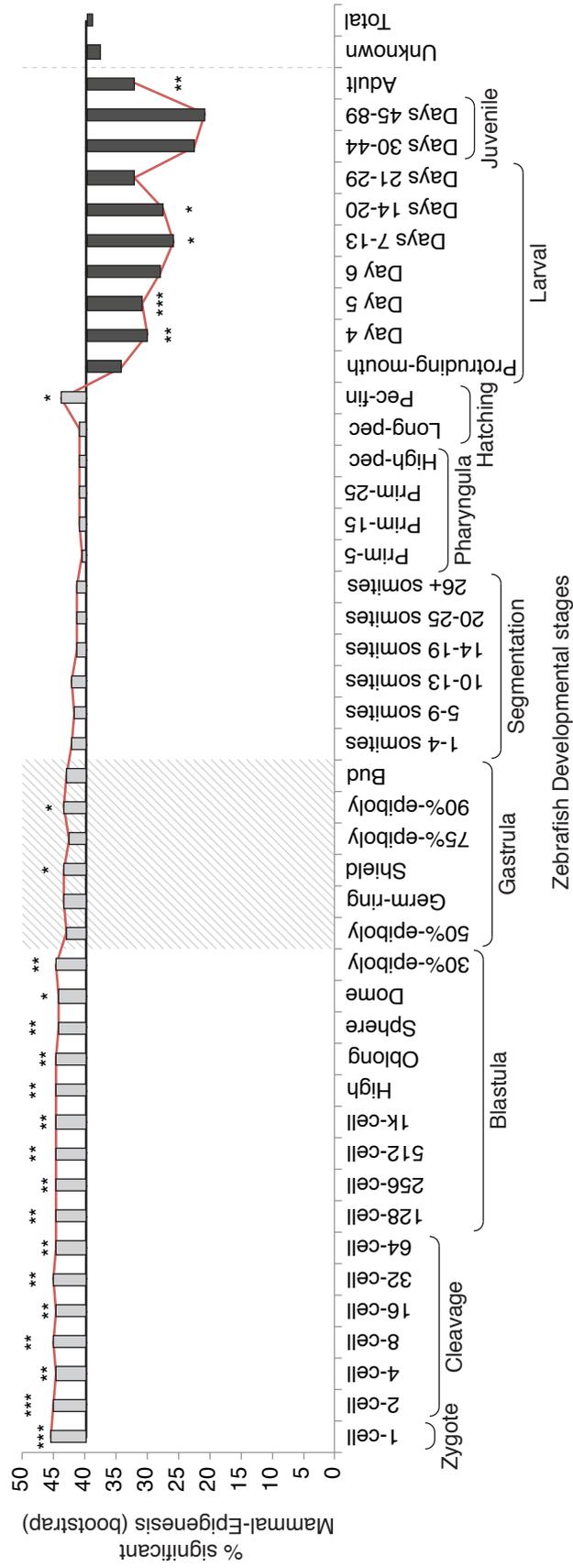


Figure 4.9: The proportion of zebrafish genes with a Mammal-Epigenesis bootstrap result expressed at each developmental stage. This graph is the equivalent to Figure B.9 (Appendix, page 255) but for zebrafish genes instead of mouse. The difference between the proportion at each stage and the average is deemed significant using the Chi-squared test (Bonferroni corrected; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$; 1 d.f.). Zebrafish gastrulation is hashed.

have been induced. In zebrafish, whose PGCs are specified by preformation, the over-representation persists beyond gastrulation.

4.3.2 Location of Expression

To examine whether there was any correlation between our genes of interest and the location of expression we have used the same method as before. For each known location of expression we have compared the proportion of Preformation-faster/Mammal-Epigenesis genes to the average across all genes with known expression.

For mouse, we took all expression data from the Mouse Genome Informatics site (downloaded 20/01/12), and for each location marker, simplified the term to a system (e.g. alimentary) and organ (e.g. pancreas). Not all terms are part of a system or organ, but we used a similar level of description based on the available information (e.g. extraembryonic, endoderm). This two tier system allowed us to summarise the location information into 28 higher-tier categories, or use both descriptors for a more precise analysis. This process was required to summarise data of mixed quality, where many locations comprised either too much or too little detail (for example, 'sinus venosus; left horn' and 'embryo')

The proportion of mouse genes with an ortholog that had a preformation-faster RRT result in each higher-tier location is shown in Figure 4.10. As with the timing graphs, each stage is shown relative to the average across all genes with known expression information. Only four locations have a value higher than average; gland, haemolymphoid, integumental and neural. None of these are significantly different to the average. Interestingly, one of germ layers, mesoderm, is significantly under-represented with Preformation-faster genes.

Using the more detailed list of locations (Table A.23, Appendix page 231) shows that 14 locations had a significantly different proportion of Preformation-faster genes than the average (Chi-squared test; Bonferroni corrected; $p < 0.05$; 1d.f.); all of which were under-represented. The most significant locations were 'embryo, head' and 'limb, hindlimb', where only 65% of genes had a Preformation-faster result compared to the average of 73%.

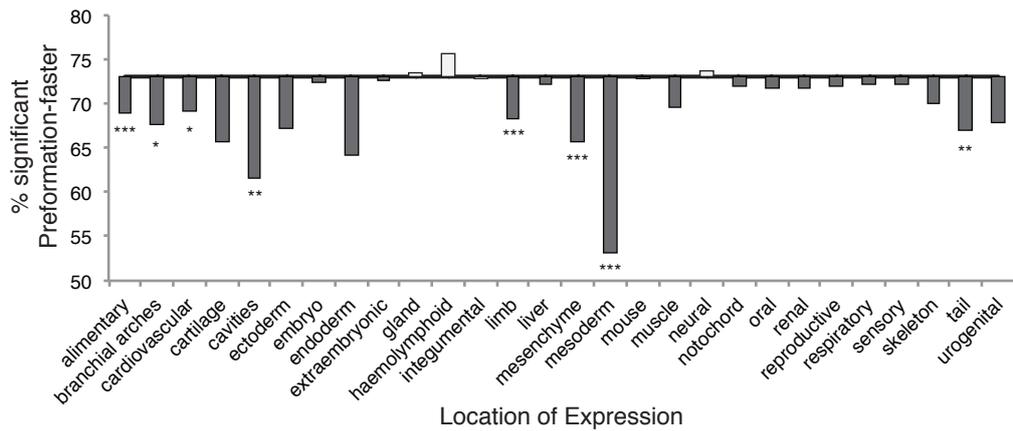


Figure 4.10: The proportion of Preformation-faster genes expressed in each mouse higher-tier location. For each higher-tier location within the mouse embryo, the proportion of genes with a Preformation-faster result is shown relative to the average (solid line). The difference between the two is judged to be significant using the Chi-squared test (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 1 d.f.).

The mouse genes allocated a Mammal-Epigenesis bootstrapped result show no significant over-representation at any location (Figure 4.11). The only higher-tier location with a significant under-representation is cartilage. Table A.24 (Appendix, page 232) shows only 2 detailed locations with a significant difference in the proportion of Mammal-Epigenesis bootstrapped genes after applying Bonferroni correction. These two locations were 'limb, hindlimb' ($p = 0.0212$) and 'limb, forelimb' ($p = 0.0214$), with Mammal-Epigenesis proportions of only 29%.

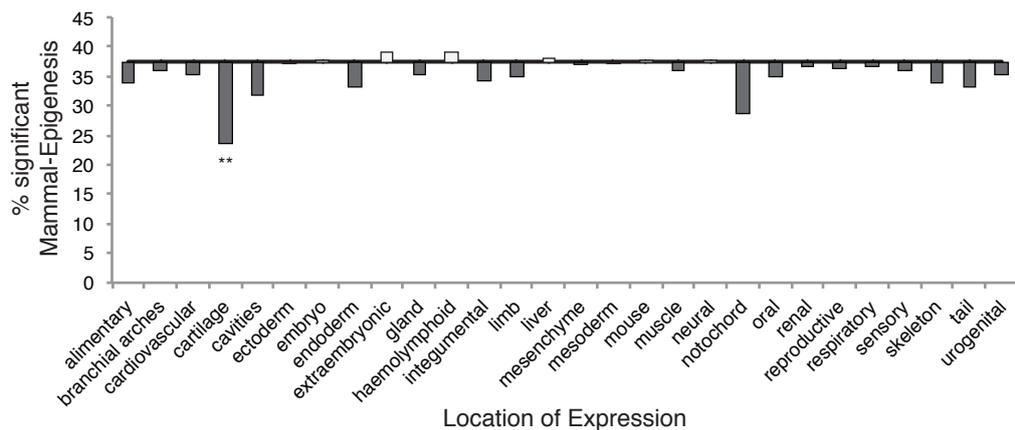


Figure 4.11: Proportion of Mammal-Epigenesis trees significant by bootstrapping expressed in each mouse higher-tier location. For each higher-tier location within the mouse embryo, the proportion of genes with a Mammal-Epigenesis result significant by bootstrapping is shown relative to the average (solid line). The difference between the two is judged to be significant using the Chi-squared test (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$; 1 d.f.).

The proportion of mouse genes with a Mammal-Epigenesis SH-test result, expressed in each location are shown in Figure B.11 (Appendix, page 257). None of the locations have a proportion of Mammal-Epigenesis genes that significantly differs to the average (Chi-squared test; Bonferroni corrected; $p \geq 0.05$; 1 d.f.). Furthermore none of the detailed locations show a significant over or under-representation of Mammal-Epigenesis genes compared to the average.

The expression data for zebrafish was kept as raw data, without simplifying it to a more accessible list of terms. The results when using the relative rate test results, specifically looking at the proportion of Preformation-faster genes, are shown in Table A.26 (Appendix, page 232). There are no locations with a significant over-representation after applying Bonferroni correction, prior to this correction there are only four terms over-represented: 'whole organism', 'immature eye', 'proliferative region' and 'forerunner cell group'. There are 23 terms with a significant under-representation, the one with the lowest p-value being 'tegmentum'.

Using the bootstrapping results, the proportion of Mammal-Epigenesis genes per location are shown in Table A.27 (Appendix, page 232). None of these locations are significantly over or under-represented after applying Bonferroni correction. However, prior to correcting for multiple testing only one location is over-represented, 'neural keel'. There were 14 locations with a significant under-representation, the location with the lowest p-value is 'gill' which shows only 22% of genes have a Mammal-Epigenesis result compared to the 38% on average.

The SH-test results per zebrafish gene expression location are shown in Table A.28 (Appendix, page 232). Once again there are no significantly over or under-represented locations after correcting for multiple testing. The one term with a significant over-representation prior to Bonferroni correction was 'pectoral fin musculature'. This location shows that 25% of the genes expressed here have a Mammal-Epigenesis tree within their orthologs significant by the SH-test, this compares to the 17% across all genes with known expression.

Looking at the proportion of genes with an interesting result across the location of expression has shown no clear pattern. We have identified no locations with any significant over-representation and few with a significant under-representation. Also, there appears to be no consistent pattern between the three tests and the two species.

Considering what we know about the time of expression in mouse (Section 4.3.1), it was surprising not see early locations such as the inner cell mass having an over-represented proportion of interesting genes. Conversely it was not surprising to see such under-representation considering that almost all of the locations described are not present until after gastrulation, at which point the interesting genes are less abundant. This includes the higher-tier categories ‘endoderm’, ‘mesoderm’ and ‘ectoderm’ which are not described until TS9; by which point there is already an under-representation of our genes of interest (Figure 4.6, page 110).

The data from zebrafish indicates that the early over-representation of interesting genes is not limited to any particular cell-type and is instead more generalized. This is particularly clear when we consider that very few locations had any over-representation and the ones that did included ‘whole organism’, although it was insignificant after correcting for multiple testing .

4.4 Gene Age

Recent studies have linked the age of genes, i.e. how far back in natural history homologs can be identified, with the rate of evolution (Wolf et al., 2009) as well as the time of expression (Domazet-Lošo and Tautz, 2010). We therefore investigated whether there was a link between our results and gene age.

For each mouse gene we searched the known orthologs in Ensembl identifying the species with the ‘oldest’ last common ancestor, similar to the methods devised by Domazet-Lošo et al., 2007. Therefore, a gene assigned the age ‘Deuterostomia’ has a known ortholog whose last common ancestor was at the base of deuterostomes, e.g. in the echinoderm *Strongylocentrotus purpuratus*. We identified orthologs as far back as the Eukaryotic last common ancestor. The total number of genes in each age category are shown in Figure 4.12. The majority of mouse genes had orthologs in the Eukaryotic common ancestor.

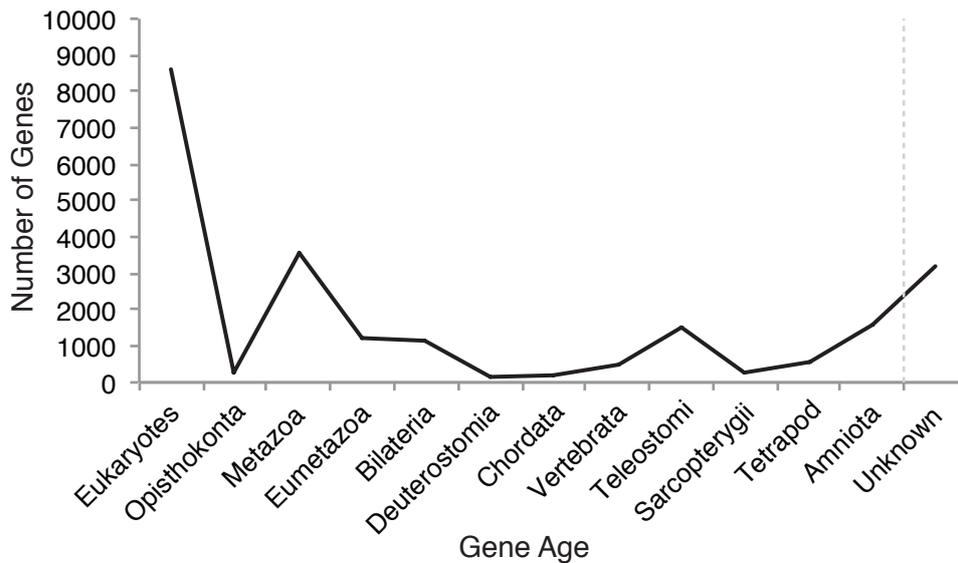


Figure 4.12: Mouse genes per age category. For each age category the total number of mouse genes are shown. There are also a number of genes with no identifiable age category.

As before, we compared the proportion of genes in each age category with a particular result (e.g. Mammal-Epigenesis topology) against the average of all genes with known information with that same result. The proportion of mouse genes with an ortholog evolving significantly faster in the taxa that has acquired preformation per age category are shown in Figure 4.13. This shows a significant over-representation of Preformation-faster genes that have a Eukaryote common ancestor. The proportion of Preformation-faster genes then decreases and younger genes are significantly under-represented. There are only 21 genes in total which can be dated to the Sarcopterygii common ancestor, and so the over-representation result is unreliable as well as insignificant. There is therefore a trend for mouse genes with a Preformation-faster result to be typically older genes.

Looking at the bootstrapped Mammal-Epigenesis genes, shows the same pattern as the relative rate results (Figure 4.14). There is a trend towards the genes with Mammal-Epigenesis topologies to be older genes. There are too few genes with significant SH-test results outside of the Eukaryote age category to analyse their association with gene age.

The zebrafish genes show a similar number of genes in each age category as the mammal, although of course the youngest category is now 'Teleostomi'

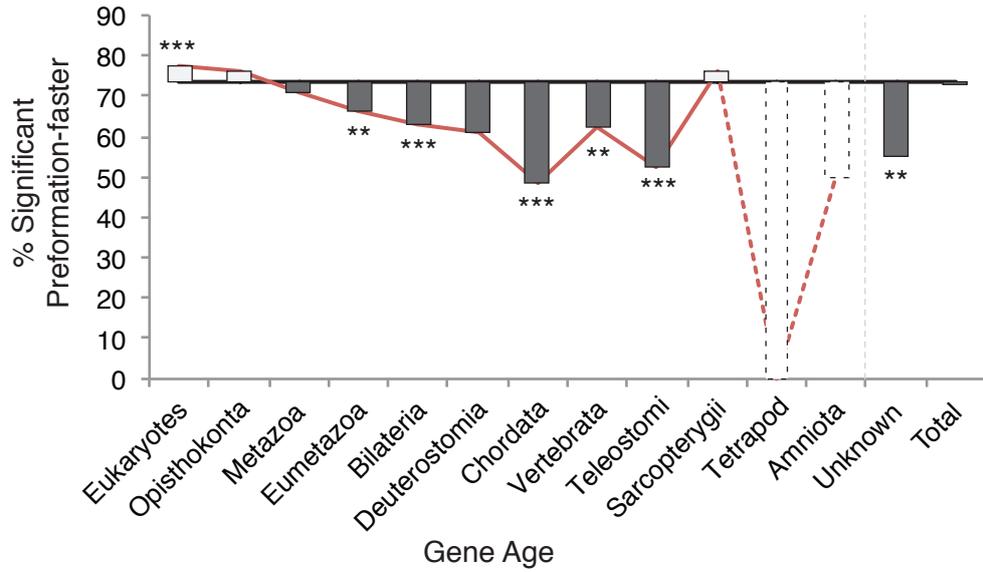


Figure 4.13: The proportion of Preformation-faster mouse genes per age category. For each age the proportion of genes with a Preformation-faster result within their orthologs are compared to the average across all age categories. The proportion of genes with no known age and the total (inc. unknown) is also shown. The Chi-squared test was used to examine if the values significantly differed to the average (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 1 d.f.). The age categories with ≤ 20 genes in total are shown as dashed outlines.

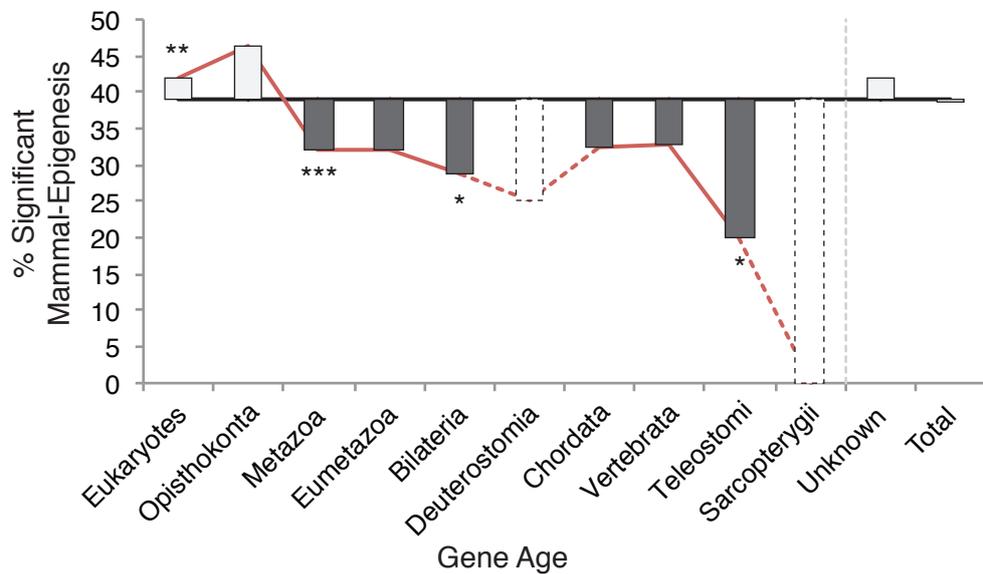


Figure 4.14: The proportion of Mammal-Epigenesis mouse genes per age category. For each age the proportion of genes with a Mammal-Epigenesis result significant by bootstrapping within their orthologs are compared to the average across all age categories. The proportion of genes with no known age and the total (inc. unknown) is also shown. The Chi-squared test was used to examine if the values significantly differed to the average (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 1 d.f.). The age categories with ≤ 20 genes in total are shown as dashed outlines.

(data not shown). Looking at the proportion of genes with a preformation-faster result (Figure 4.15), shows a very similar result to the equivalent mouse genes. There is a significant over-representation of genes with a faster rate of evolution in the preformation ortholog in the oldest gene category. The younger gene categories have a significant under-representation of genes with the same result.

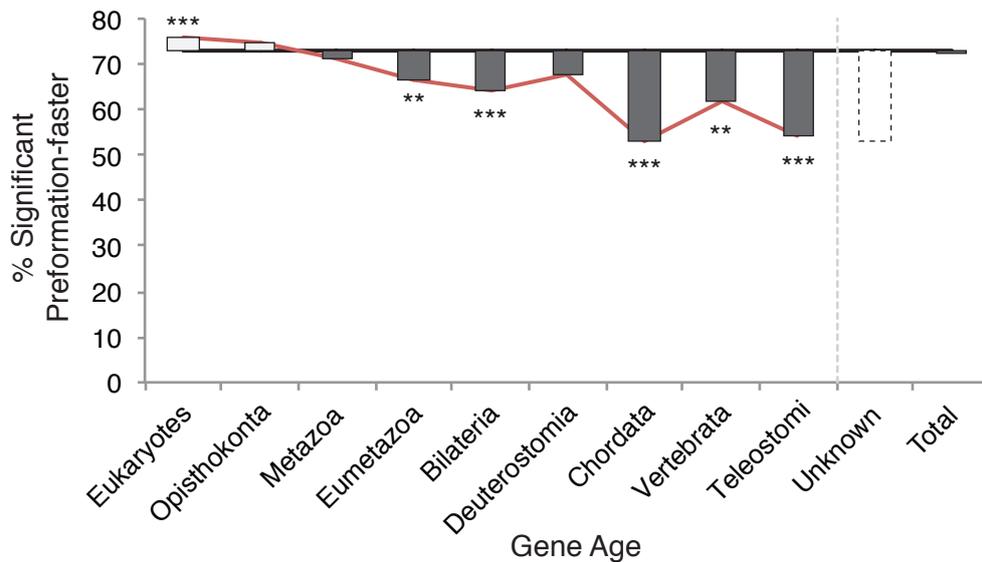


Figure 4.15: Zebrafish genes with preformation-faster result per age category. As in Figure 4.13 but for the results mapped to zebrafish genes instead of mouse. The Chi-squared test was used to examine if the values significantly differed to the average (Bonferroni corrected; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$; 1 d.f.). The age categories with ≤ 20 genes in total are shown as dashed outlines.

This pattern of interesting genes correlating with an old gene age continues for the bootstrapped Mammal-Epigenesis zebrafish genes (Figure 4.16). There is a significant over-representation of these genes at the Eukaryotic common ancestor. Meanwhile, there is a significant under-representation of these Mammal-Epigenesis genes in younger age categories. As in mouse, there are too few genes with significant SH-test results to analyse their association with gene age.

These results have shown a strong association towards our genes of interest arising in the Eukaryotic last common ancestor. On first viewing, this seems to differ to the results presented in Wolf et al., 2009 where they showed that ancient genes had a slower short term evolutionary rate than younger genes. However, their work consisted of relative rate analyses between closely related species (such as between human and macaque) that share the same mode of PGC specification. Our results have instead shown that these ancient genes

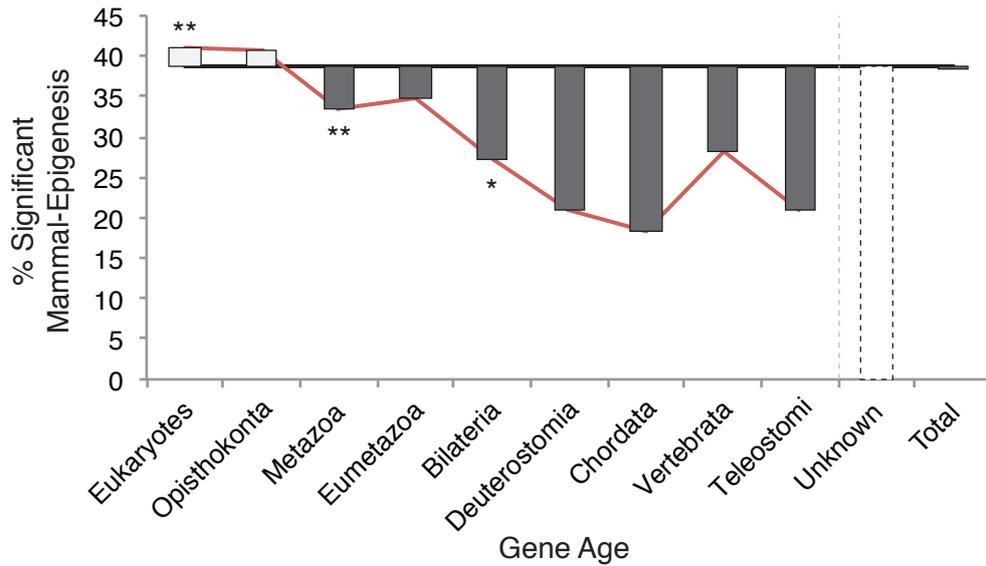


Figure 4.16: Zebrafish genes with bootstrapped Mammal-Epigenesis result per age category. As in Figure 4.14 but for the results mapped to zebrafish genes instead of mouse. The Chi-squared test was used to examine if the values significantly differed to the average (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 1 d.f.). The age categories with ≤ 20 genes in total are shown as dashed outlines.

are more likely to have an accelerated rate of evolution in species that have acquired preformation compared to their sister taxa which have retained epigenesis. They may therefore continue to show a slower rate in primates compared to younger genes, but are more likely to have an accelerated rate in the branches leading to teleosts, anurans, birds and snake.

4.5 Conclusions

By mapping the query sequences to single genomes we have been able to analyse the association between function, expression and gene age relative to the tree and RRT results. We have also been able to summarise the four-taxon tree and relative rate test results from Chapter 3 without any duplications. This demonstrated that the bias within the incongruent topologies and genes with a significant difference in rate stands.

We were unable to find a functional relationship between the genes with either a preformation-faster rate or a Mammal-Epigenesis topology. We were also unable to associate a particular location of expression with these genes. Together these data suggest that the genes with a preformation-faster or Mammal-Epigenesis result are widespread and not specific to either function or location

of expression. This is supported by the substantial number of genes with a ortholog with a preformation-faster result, over 7,500 genes.

We were however able to find a correlation between these genes of interest and the time of expression. The results suggest that genes with an ortholog that has a significantly faster rate of evolution in the taxon that has acquired preformation are typically expressed during early development. This is also true for those with a Mammal-Epigenesis tree topology within the vertebrate orthologs. These data support the theory of constraint/constraint release as genes expressed early in development are known to be under the highest levels of constraint (Roux and Robinson-Rechavi, 2008).

Furthermore we also saw a correlation between these genes of interest and their age. Those with preformation-faster or Mammal-Epigenesis results were typically ancient genes with known orthologs as far back as the last common ancestor of Eukaryotes.

These data therefore suggest that the genes with a molecular evolution that correlates with the mode of PGC specification tend to be expressed early in development and have been mostly present since the Eukaryotic common ancestor. They are expressed throughout multiple tissues and share no common function. This suggests that these genes are fundamental, and are not associated with derived tissues or specialised functions.

Expanding the Global Analysis

In Chapter 3 the majority of our sequences came from the amphibians and actinopterygians, primarily due to our sequencing of the axolotl and sturgeon transcriptomes. This meant that almost all anuran and teleost sequences were being compared against a single species. To test whether this had biased our results we expanded the dataset to include other transcriptomes. We have therefore investigated whether this expanded dataset continues to show a correlation between the mode of PGC specification and patterns of molecular evolution. To do this we have used the same methods as before, namely 4-taxon trees and the relative rate test.

Within the amphibians we have incorporated the transcriptome data from two species of anuran, *Rana kukunoris* and *Rana chensinensis*, and one species of urodele, *Notophthalmus viridescens*. We are therefore able to reanalyse problematic species such as *Notophthalmus viridescens* and *Rana chensinensis* using a larger number of sequences, as well as studying a new anuran species. We also combined another *Ambystoma mexicanum* transcriptome (Stewart et al., 2013) with our own data. The anurans will therefore be compared against two different urodele transcriptomes.

In Actinopterygii we have incorporated the whole genome from *Lepisosteus oculatus*, the spotted gar. This species is more closely related to teleosts than sturgeon (Figure 1.1, page 2). Although there has been no experimental verification on the mode of PGC specification we would predict gar to have retained epigenesis, mostly due to its primitive morphology. By studying the phylogenetic incongruence and rates of evolution we will be able to examine whether gar has undergone the same molecular changes as teleosts or whether it resembles sturgeon.

We have acquired new genomes and transcriptomes from sauropsids, allowing us to further explore these previously under sequenced taxa. This includes the two whole genomes released in turtles, *Pelodiscus sinensis* and *Chrysemys picta bellii*. We have also added transcriptomes from 3 species of Scincidae *Carlia rubrigularis*, *Lampropholis coggeri* and *Saproscincus basiliscus* and one gecko, *Eublepharis macularius*. The skinks are thought to be undergoing preformation while geckos have retained epigenesis (Figure 1.9, page 20). This allowed us to investigate whether the lack of phylogenetic incongruence observed previously was a true observation of sauropsid evolution or an artefact due to a lack of data.

We have also sought to investigate sequence evolution in previously unstudied lineages, specifically coelacanth, lungfish and sharks. Each of these are thought to be undergoing epigenesis yet none of them have a sister taxon in which preformation has evolved. We have therefore developed new methods for studying their sequence evolution, although we have continued to use four-taxon trees and the relative rate test.

5.1 Four-taxon Trees

As in Section 3.2, we built 4-taxon trees comprising two sister taxa, a mammal species and an outgroup. We investigated how many of these trees recapitulate the species phylogeny, and how many are incongruent. Within those that are incongruent, we observed whether there is a bias that correlates with the mode of PGC specification.

Within amphibians, by incorporating new sequences for *Rana kukunoris*, *Rana chensinensis*, *Notophthalmus viridescens* and *Ambystoma mexicanum*, we built 32,291 four-taxon trees. As before we used a teleost outgroup. The bootstrapped trees (Figure 5.1) show the same result as before. The bias in incongruent trees is clear across all of the amphibian species and is far more striking than when only the *Ambystoma mexicanum* transcriptome was included. The bias has particularly strengthened for those species with an enlarged sequence dataset such as *Rana chensinensis* and *Notophthalmus viridescens*.

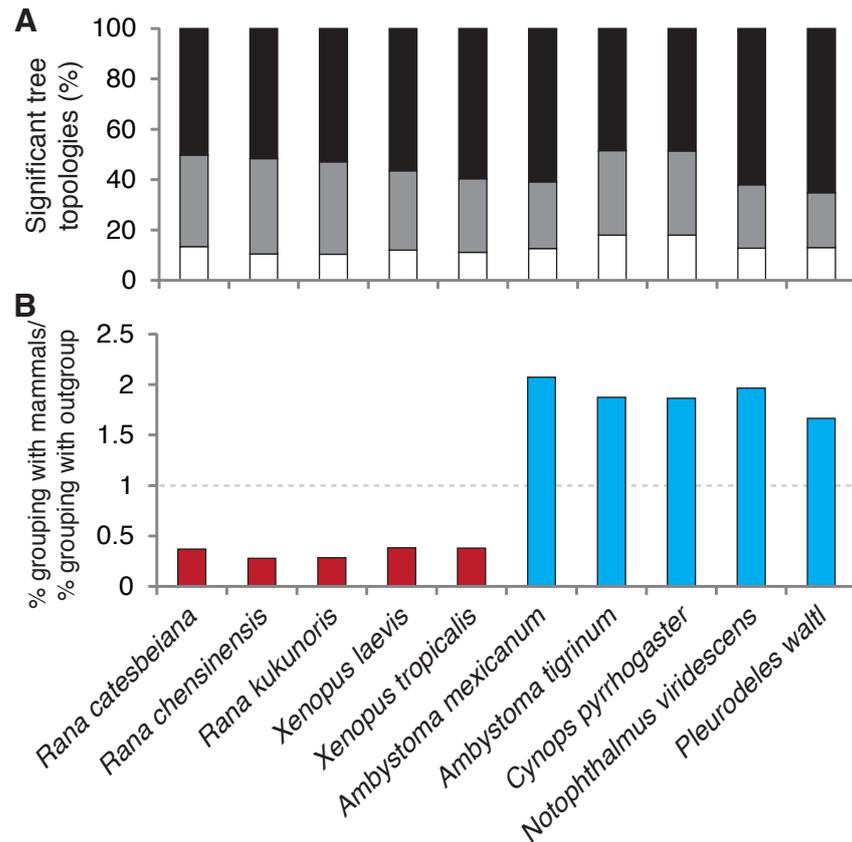


Figure 5.1: Amphibian bootstrap results. As in Figure 3.14 the top panel (A) shows the proportion of significant tree topologies; the species phylogeny in black, Mammal-Urodele in grey and Mammal-Anuran in white. (B) shows the likelihood of each species grouping with mammals. The anurans are shown in red, the urodeles in blue.

These results show that there is still a strong bias within the incongruent trees which groups the mammals and urodeles together, the two species undergoing epigenesis. This suggests that the previous results were not an artefact of the *Ambystoma mexicanum* transcriptome.

Within Actinopterygii we added the gar genome, which, along with sturgeon, was compared against teleosts with an amphioxus outgroup. Although there is no experimental evidence that gars are undergoing epigenesis, their primitive morphology and basal position suggests that they might.

The bootstrap significant results are shown in Figure 5.2. Once again these results strongly resemble the previous data. Comparing gar against sturgeon, shows a larger proportion of species phylogeny trees, probably reflecting the closer relationship between gars and teleosts. Within the incongruent trees, gar

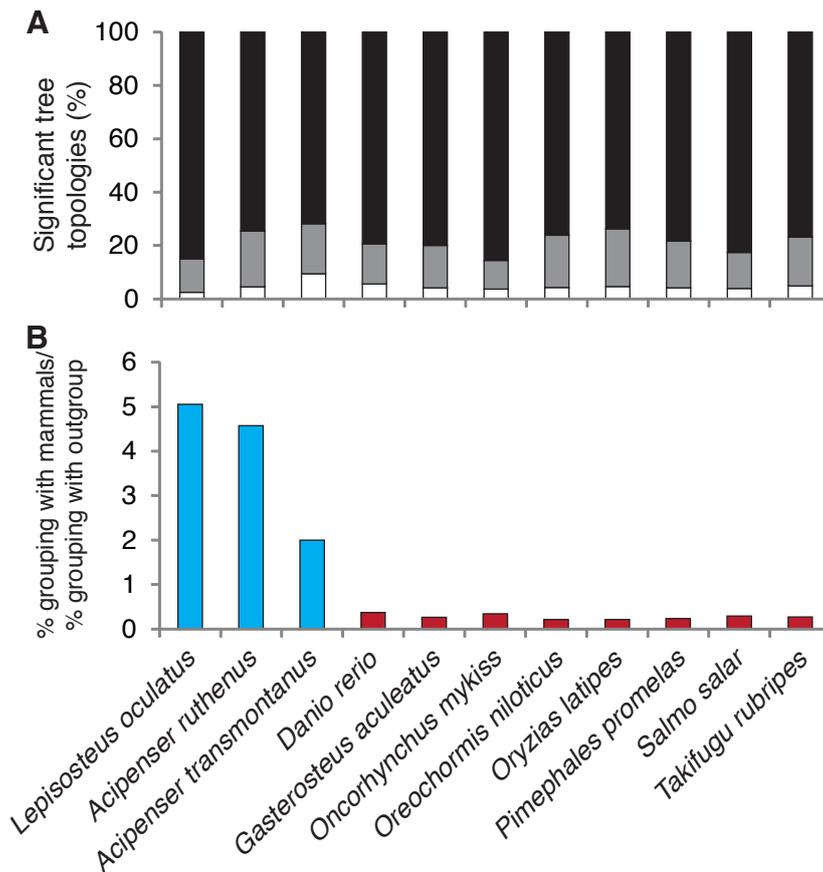


Figure 5.2: Actinopterygian bootstrap results. The results are shown in the same format as Figure 5.1. The black in (A) shows the proportion of Species Phylogeny trees, the grey shows the Mammal-Sturgeon/Gar trees, and the white is the proportion of Mammal-Teleost trees. In (B) the gar (*Lepisosteus*) and sturgeons (*Acipenser*) are shown in blue, the teleosts in red.

shows the same bias as sturgeon with a high likelihood of grouping with mammals. Therefore, in conjunction with other observations, these results suggest that gars undergo epigenesis.

With the addition of the complete genomes' worth of data from two species of turtle we were able to build 35,856 archosaur and testudine 4-taxon trees rooted on the anuran outgroup. This compares to the previous 1,082 trees we had been able to build using only the EST based data.

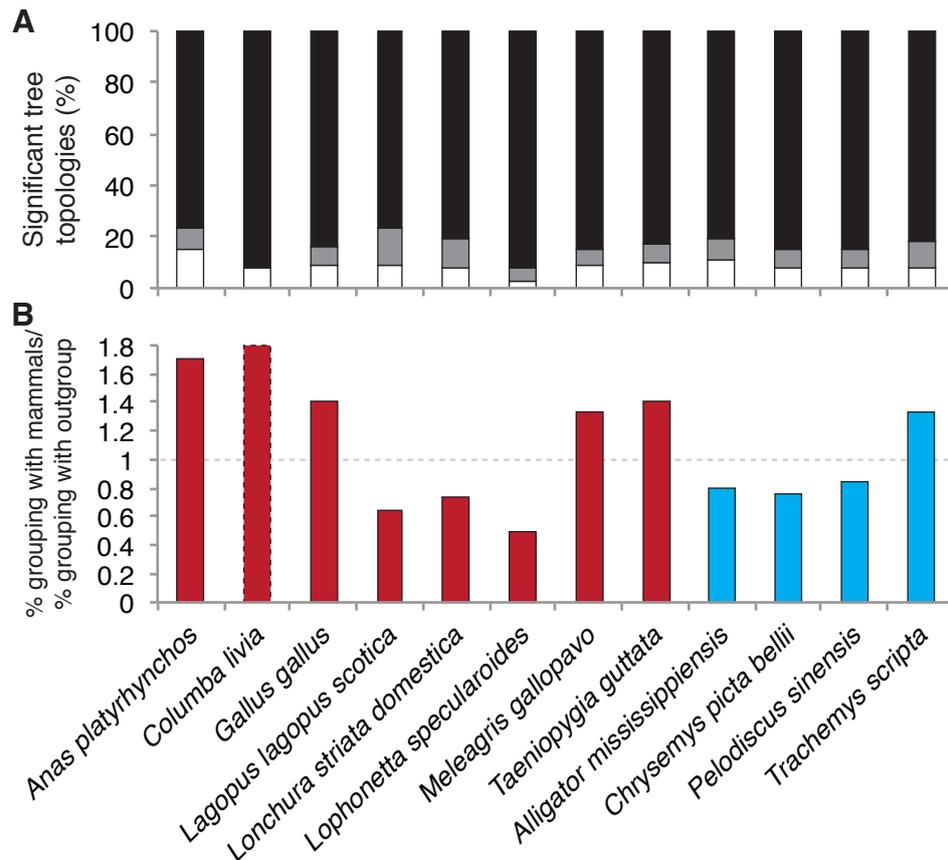


Figure 5.3: Archosaur bootstrap results. The results are shown in the same format as Figure 5.1. The black in (A) shows the proportion of Species Phylogeny trees, the grey shows the Mammal-Crocodile/Turtle trees, and the white is the proportion of Mammal-Bird trees. In (B) the birds are shown in red, the crocodile and turtles in blue. *Columba livia* only grouped with mammals when the tree was incongruent, as shown by the dashed bar.

The results from the bootstrap-significant trees show that approximately 80% of the trees in each species reflect the species phylogeny (Figure 5.3). For those that are incongruent there is no consistent bias as to whether the lineage is likely to group with mammals or not. This is the same result that we saw previously (Figure 3.24) when we used only the EST based data.

In lepidosaurs we have added transcriptomes from a species of Gekkota (epigenesis), as well as three species of Scincoidea (preformation). The relationships between these new species and the previous species (from the Serpentes and Pleurodonta families) are shown in Figure 1.9 (page 20). Since the term lizard encompasses taxa undergoing both epigenesis and preformation, the groups are instead referred to according to their mode of PGC specification. We built a total of 28,982 four-taxon trees using this larger lepidosaur dataset.

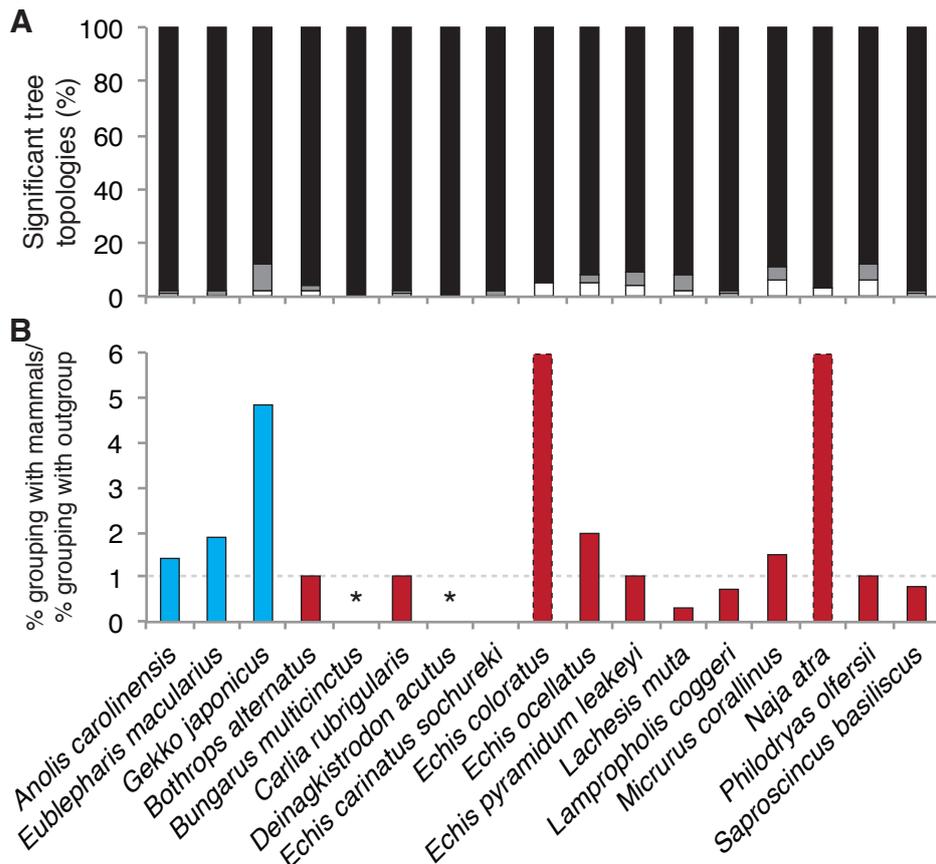


Figure 5.4: Lepidosaur bootstrap results. The results are shown in the same format as Figure 5.1. The black in (A) shows the proportion of Species Phylogeny trees, the grey shows the Mammal-Epigenesis trees, and the white is the proportion of Mammal-Preformation trees. In (B) the species undergoing epigenesis are shown in blue, those that have acquired preformation are shown in red. *Echis coloratus* and *Naja atra* only grouped with mammals when the tree was incongruent. **Bungarus multicinctus* and *Deinagkistrodon acutus* had no incongruent trees.

The results in Figure 5.4 resemble what we saw before the addition of the transcriptomes (Figure 3.26). In each species almost all of the tree topologies reflect the species phylogeny. Within the few trees that are incongruent, there is

no consistent bias between the Mammal-Epigenesis and Mammal-Preformation topologies.

Summary

The total number of trees built in each class are shown in Table 5.1. By adding in the latest transcriptome data we have seen that the amphibian results with a previously unclear bias, were due to a lack of sequence data. Within the actinopterygii, the 4-taxon trees suggest that gar is also undergoing epigenesis as the result resembles sturgeon. In both groups of sauropsids adding in the extra data has not affected the result we saw previously, there is still no consistent bias within the incongruent trees. Therefore, increasing the number of sequences for each species reinforces the result obtained previously but does not change the direction of bias.

Table 5.1: Summary of the expanded dataset four-taxon tree results.

(A) Bootstrap results.

Class	Total	Species Phylogeny	Mammal-Epi.	Mammal-Pre.
Amphibians	32 291	9077	4893	1928
Actinopterygii	43 005	18 893	4317	1336
Archosaurs	35 856	18 950	1533	1990
Lepidosaur	28 982	24 429	333	238

(B) SH-test results.

Class	Species Phylogeny	Mammal-Epi.	Mammal-Pre.
Amphibians	1251	335	87
Actinopterygii	3921	273	30
Archosaurs	5313	31	164
Lepidosaur	14 669	28	6

Across all vertebrates we have built 140,134 four-taxon trees, of which 87,917 were significant using bootstrapping. Of these trees, 71,349 (81.2%) reflected the species phylogeny, 11,076 (12.6%) showed a Mammal-Epigenesis topology and 5,492 (6.2%) grouped mammals with the species that has acquired preformation. Therefore, even with the inclusion of the sauropsids, there are still twice as many Mammal-Epigenesis trees as there are Mammal-Preformation trees. The bias towards grouping the species undergoing epigenesis together is still present after adding in the new data.

5.2 Relative Rate Test

As well as building 4-taxon trees, we wished to test the new data using the relative rate test (RRT). As before (Section 3.4, page 87), we used the three taxon alignments, and with mammals as the reference compared the rate of evolution between the two sister taxa.

With the additional urodele and anuran sequences we were now able to test 49,425 Amphibian 3-taxon alignments. The results per species (Figure 5.5) show that approximately 35% of sequences in each species are evolving at significantly different rates. In each case there are more sequences evolving significantly faster in anurans compared to urodeles.

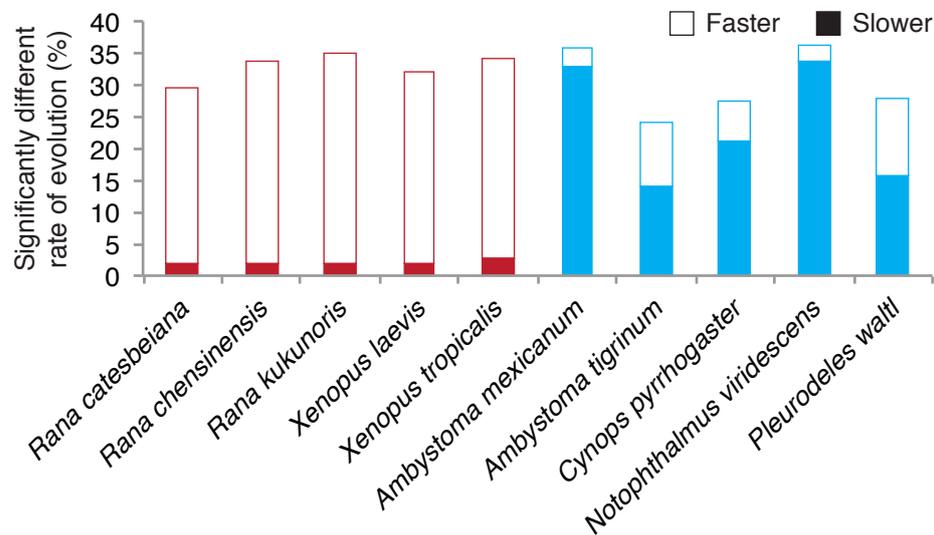


Figure 5.5: Amphibian relative rate test results. The figure shows the proportion of significantly faster evolving sequences as clear bars, and the proportion of significantly slower evolving sequences as filled bars. The anurans are shown in red, urodeles in blue.

The addition of the new transcriptomes has particularly affected *Notophthalmus viridescens*, as can be seen in Figure 5.6. The previous results (Figure 3.31) showed that only 20% of sequences were evolving at significantly different rates, and there was no clear bias in terms of the direction. By including the new transcriptomes, *Notophthalmus viridescens* now shows 35% of sequences are evolving at significantly different rates, the vast majority of which have a slower rate in the urodele than in anurans. This suggests that if we were able to increase the quantity of data for the other urodeles then these would also show a greater bias.

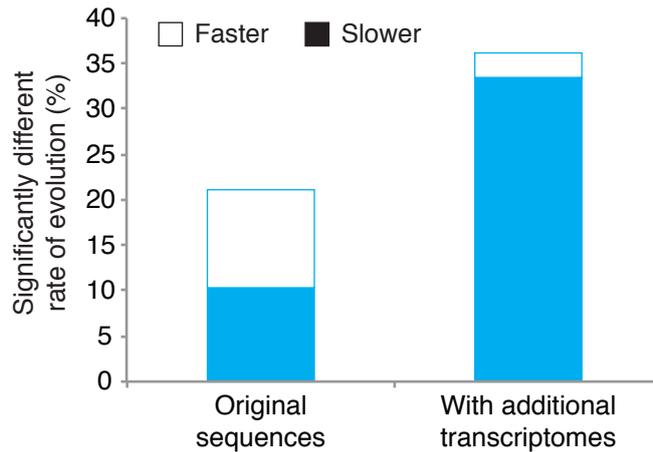


Figure 5.6: RRT results for *Notophthalmus viridescens*. The results are shown in the same format as Figure 5.5 for *Notophthalmus viridescens* using the original (Figure 3.31) and new data.

Using this larger amphibian dataset we see that 16,696 sequences show a significantly different rate of evolution between the sister taxa (Table A.37, Appendix page 236). Within these, 91.6% (15,289) have a faster rate in the anuran than urodele; only 1,407 sequences showed the opposite result.

With the inclusion of the gar genome we were able to test 118,691 actinopterygian 3-taxon alignments. The results shown in Figure 5.7, show that approximately 50-60% of sequences are evolving at significantly different rates. They also show that for each species almost all of these sequences have a significantly faster rate in the teleost than in the urodele. This includes the gar, *Lepisosteus oculatus*, which has an almost identical result to *Acipenser ruthenus* (Table A.38, Appendix page 236).

In total, across all species of Actinopterygii tested, there were 57,283 alignments with a significantly different rate in evolution between the sister taxa. Of these, 98.5% (56,424) showed a significantly faster rate of evolution in teleost than in sturgeon or gar. Only 859 sequences showed a significantly slower rate of evolution in teleosts.

Within the archosaurs, the addition of two turtle genomes meant we tested 51,248 alignments. Figure 5.8 shows that in each species there are approximately 35% of sequences evolving at significantly different rates. Within these alignments, there are more sequences with a significantly faster rate in the bird than either crocodile or turtle.

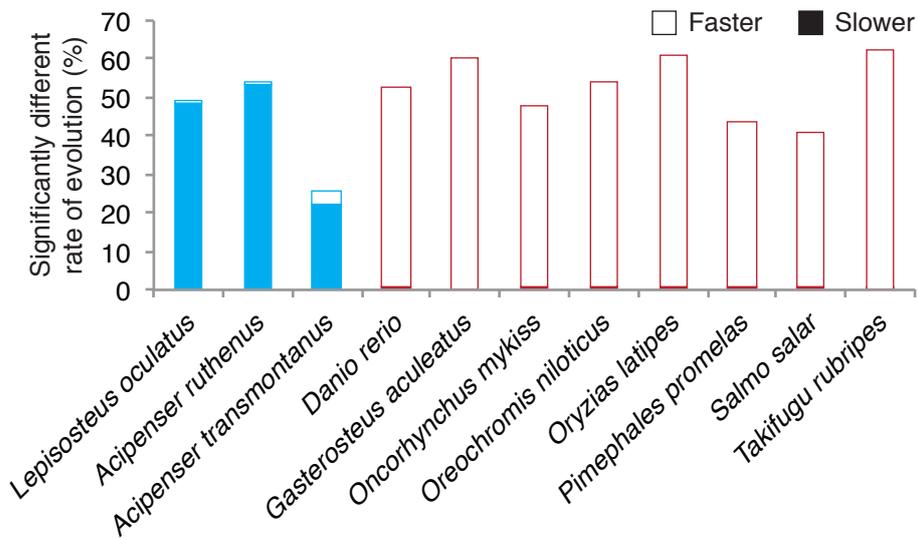


Figure 5.7: Actinopterygian relative rate test results. The results are shown in the same format as Figure 5.5. The gar and sturgeons are shown in blue, the teleosts in red. Only a few teleost species are shown, the rest are in Figure B.12 (Appendix, page 258).

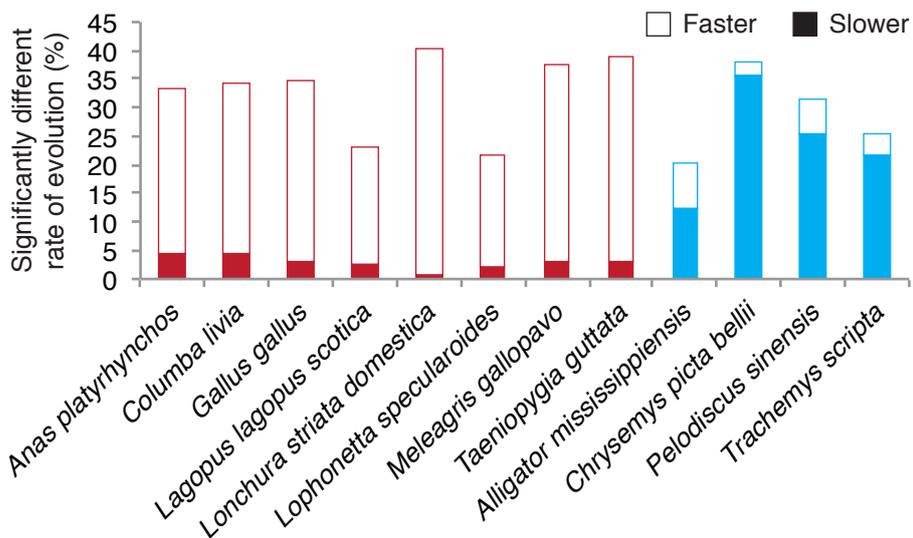


Figure 5.8: Archosaur and Testudine RRT results. The results are shown in the same format as Figure 5.5. The Birds are shown in red, the Crocodile and Turtles in blue.

This result is consistent across all species of bird and turtle, and is more striking than observed previously (Figure 3.34). Considering what we have seen already with the addition of more sequences it suggests that the weaker result in *Alligator mississippiensis* is due to a lack of data (Table A.39, Appendix page 237).

Overall in archosaurs there were 18,443 alignments with a significantly different rate of evolution between the sister taxa. Within these 90.5% (16,695) showed a significantly faster rate in the birds, compared to the 1,748 alignments which showed a significantly faster rate in either crocodiles or turtles.

The larger lepidosaur dataset allowed us to test 42,147 3-taxon alignments. As explained earlier the lizards are polyphyletic and are therefore referred to by their mode of PGC specification. Figure 5.9 shows that between 15-25% of sequences are evolving at significantly different rates between the sister taxa. Lepidosaurians also show that on the whole there are more sequences evolving at a significantly faster rate in the species that have acquired preformation.

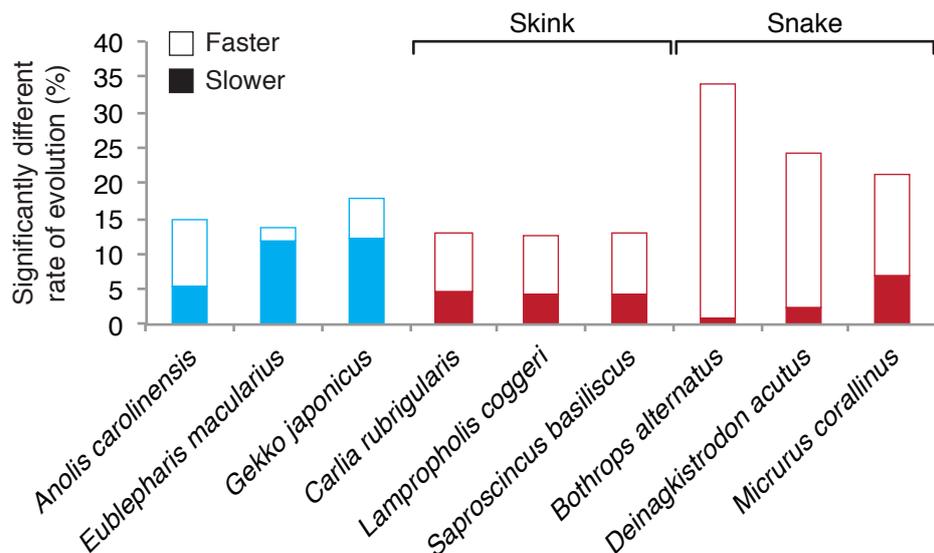


Figure 5.9: Lepidosaur relative rate test results. The results are shown in the same format as Figure 5.5. The species undergoing epigenesis are shown in blue, those that have acquired preformation are shown in red. The sequence numbers are shown in Table A.40 (Appendix, page 237).

Interestingly, the result in *Anolis carolinensis* shows more sequences evolving significantly faster than slower, the opposite to what we obtained previously (Figure 3.35, page 91). There is also a difference between the result in skinks (*Carlia rubrigularis*, *Lampropholis coggeri* and *Saproscincus basiliscus*) and snakes (*Bothrops alternatus*, *Deinagkistrodon acutus* and *Micrurus corallinus*). The

snakes appear to have a larger proportion of sequences evolving significantly faster than skinks. To explore this result in more detail, we have re-run the RRT, comparing each species against each other (Figure 5.10; Table A.41, Appendix page 238).

		Epigenesis			Preformation						
		Gekkos			Skinks			Snakes			
		<i>Anolis carolinensis</i>	<i>Eublepharis macularius</i>	<i>Gekko japonicus</i>	<i>Carlia rubrigularis</i>	<i>Lampropholis coggeri</i>	<i>Saproscincus basiliscus</i>	<i>Bothrops alternatus</i>	<i>Deinagkistrodon acutus</i>	<i>Philodryas olfersii</i>	
Epigenesis	Gekkos	<i>Anolis carolinensis</i>	88.90	66.67	66.28	63.62	63.76	11.11	11.76	0	
	<i>Eublepharis macularius</i>	11.10		48.72	18.71	17.62	17.10	0	0	25	
	<i>Gekko japonicus</i>	33.33	51.28		34.88	28.13	37.84	50	50	100	
Preformation	Skinks	<i>Carlia rubrigularis</i>	33.72	81.29	65.12		44.27	36.23	0	9.09	25
		<i>Lampropholis coggeri</i>	36.38	82.38	71.88	55.73		47.03	9.09	15.38	33.33
		<i>Saproscincus basiliscus</i>	36.24	82.90	62.16	63.77	52.97		0	27.27	20
	Snakes	<i>Bothrops alternatus</i>	88.89	100	50	100	90.91	100		100	
		<i>Deinagkistrodon acutus</i>	88.24	100	50	90.91	84.62	72.73	0		
		<i>Philodryas olfersii</i>	100	75	0	75	66.67	80			

Figure 5.10: The RRT results in Lepidosauers, comparing each species against each other. The bias within the sequences with a significantly different rate of evolution is shown. Each square shows the proportion evolving significantly faster in the species to the left, and therefore a slower rate in the species above. The squares are coloured in a scale from red (0%) through to yellow (50%) and green (100%). Not all of the comparisons shown have more than 20 significant results; the blank squares had no alignments with a significant difference. Each relative rate test used a human sequence as the reference.

These results show that when there is a significant difference in rate, snakes are generally evolving faster than skinks. The two species of gecko (*Eublepharis macularius* and *Gekko japonicus*) appear to be evolving slower than all the other sauropsids. *Anolis carolinensis* is evolving faster than skinks, but slower than snakes. The difference in result for *Anolis carolinensis* is summarised in Figure 5.11 where we have totalled the results described above for snakes and skinks.

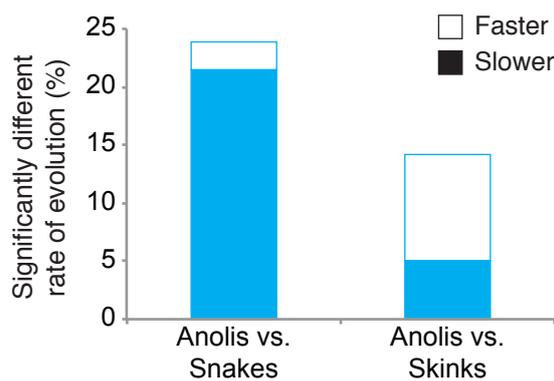


Figure 5.11: The RRT results for *Anolis carolinensis* when compared to snakes or skinks. The comparisons against snakes and skinks from Figure 5.10 are combined for *Anolis carolinensis*.

Figure 5.11 shows that when *Anolis carolinensis* is compared against snakes there are a lot of alignments where anolis is evolving significantly slower, this resembles the result we obtained in Figure 3.35. However, when *Anolis carolinensis* is compared against skinks, there are fewer sequences with a significant difference in rate, and the bias is for skinks to be evolving slower.

These data suggest one of two possibilities, that either the predicted mode of germ cell specification is wrong, or that there has been a significant change in rate at the base of the Serpentes-Pleurodonta branch. However, considering the species phylogeny is debatable and that there is no clear evidence for the mode of PGC specification, I do not believe these questions can currently be answered.

Summary

In total, across all vertebrates we have been able to test 261,511 three-taxon alignments, 37.5% (98,206) of which are evolving at significantly different rates between the sister taxa. Of these, 93.8% (92,133) show a significantly faster rate of evolution in the species that has acquired preformation. There are only 6,073 where the species undergoing epigenesis is evolving at a significantly faster rate. This includes the lepidosaurs, where we have just seen a more complicated result than the other groups

Figure 5.12 shows the species ordered according to the proportion of sequences evolving significantly slower. Since the situation in lepidosaurs is unclear these results are excluded. The remaining groups of amphibians, actinopterygii and archosaurs are clearly divided according to the mode of PGC

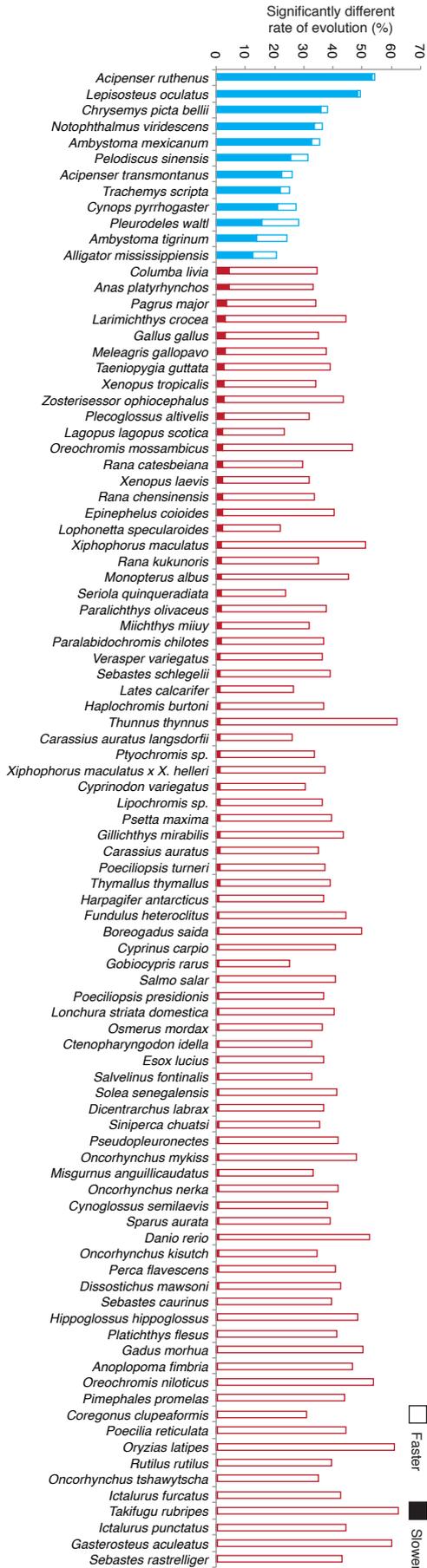


Figure 5.12: Relative Rate Test results for amphibians, actinopterygii and archosaurs, including recent transcriptomes. This figure shows that for each species which had over 20 significant results, the proportion of those with a significantly faster rate of evolution in filled bars and the proportion with a significantly slower rate of evolution in clear bars. The species undergoing epigenesis are coloured blue, those that have acquired preformation are shown in red. The species are sorted according to the proportion evolving at a significantly slower rate.

specification. There is in fact a clear ‘step’ between the proportion evolving significantly slower in the two modes.

5.2.1 Analysing the rate of evolution within lineages

To investigate whether this bias in rate of evolution is due to the difference in mode of PGC specification and not another factor that differs between sister taxa we have sought a control comparison. Ideally, this would be another group of taxa which diverged at the same time but shares the same mode of PGC specification. For example, urodeles could be compared against caecilians, and sturgeons against gar. Unfortunately, for the majority of species these taxa are not available and so we have therefore compared within their own lineage, for example comparing axolotl against other urodeles.

As before, we began this process in amphibians, building alignments with either two urodele or two anuran sequences. A large decrease in the number of alignments was immediately apparent, for example there were 1,502 *Ambystoma tigrinum* sequences when compared to anurans, but only 325 when compared to urodeles. This was due to the quality parameters applied onto the alignments (Section 3.1.1, page 58), particularly the requirement for the minimum uncorrected distance to be greater than 0.02 (Table 5.2).

Table 5.2: Number of *Ambystoma tigrinum* alignments. The total number of alignments by minimum uncorrected distance, when compared against either anurans or urodeles.

Minimum distance	Compared against	
	Anurans	Urodeles
≥ 0.0	1503	1157
≥ 0.01	1503	632
≥ 0.02	1502	327
≥ 0.04	1502	194
≥ 0.06	1502	156
≥ 0.08	1501	129
≥ 0.1	1496	112
≥ 0.2	1089	14
≥ 0.3	200	1

When compared against urodeles, the majority of alignments had a minimum uncorrected distance less than 0.02 and so had been removed. This difference in minimum distance was presumably due to the much closer relationship

between species. We therefore removed the minimum uncorrected distance parameter from all analyses which involved a within order comparison.

The final results from this analysis, along with the same sequences compared between orders, are shown in Figure 5.13. The key result we are interested in is when anurans are evolving significantly faster than urodeles, as shown by the dark blue bars for anurans in Figure 5.13A and urodeles in Figure 5.13B. The high proportion of significant results seen in these species are never recapitulated in the comparisons within orders.

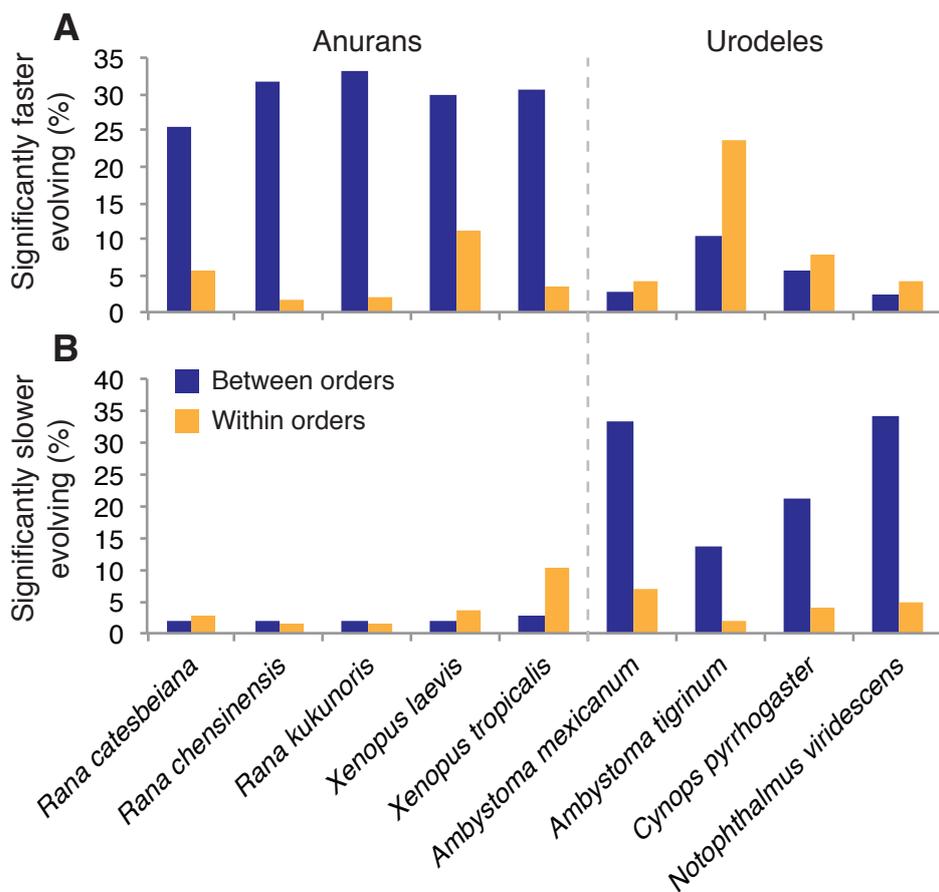


Figure 5.13: Amphibian RRT results between and within orders. The proportion of sequences evolving significantly faster (A) and slower (B) are shown for comparisons between orders (dark blue) and within orders (orange). The number of sequences are shown in Table A.42 (Appendix, page 239).

Interestingly the few ‘background’ results, where urodeles are evolving significantly faster than anurans, show a similar proportion of significant results to the within order comparisons. This is true in all cases except for *Ambystoma tigrinum* which appears to have a large number of sequences evolving significantly faster than other urodeles.

These results suggest that there are always some sequences evolving at significantly different rates, as we would expect since not all genes will follow a fixed molecular clock across all taxa. However, the comparisons between orders which differ in their mode of PGC specification clearly exceeds this ‘normal’ variation in rate. This suggests that the breadth of change between anurans and urodeles is far greater than what you would normally see between any two taxa.

The actinopterygian species comparing the results between orders against those conducted within orders are shown in Figure 5.14. In this case the sturgeons and gars provide a comparison between species which diverged at a similar time as the sturgeon-teleost split but have not altered the mode of PGC specification.

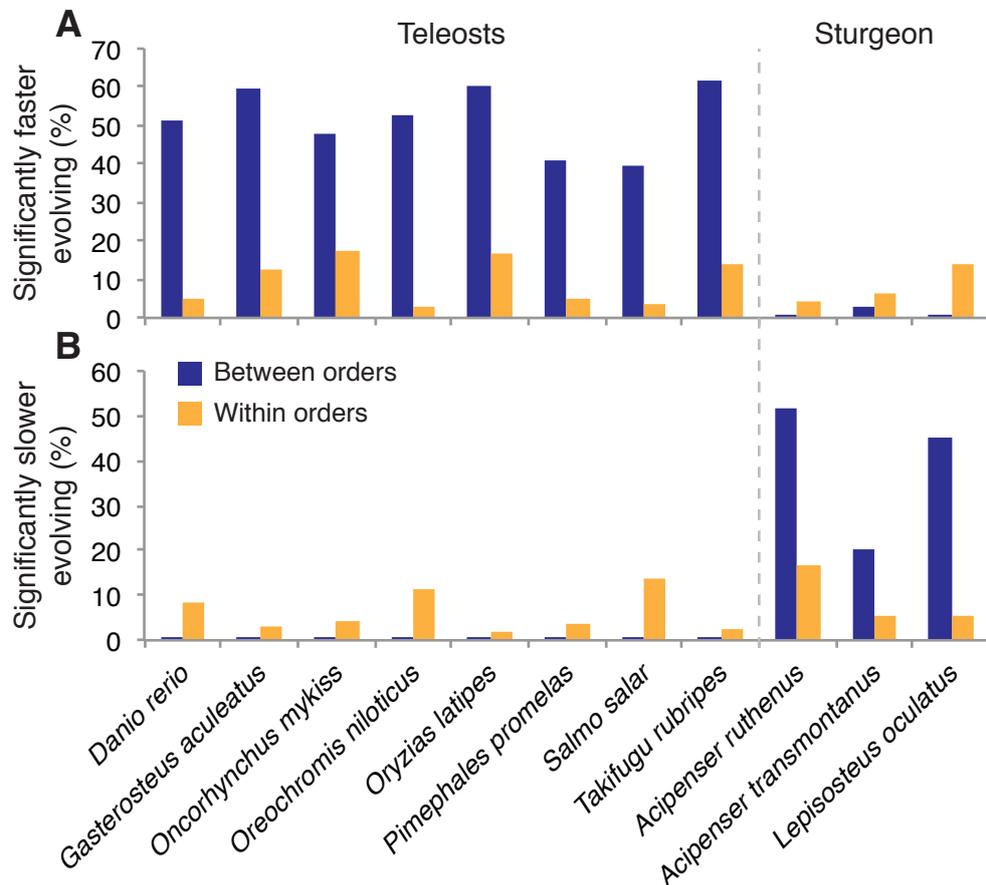


Figure 5.14: Actinopterygian RRT results between and within orders. The proportion of sequences evolving significantly faster (A) and slower (B) are shown for comparisons between orders (dark blue) and within orders (orange).

The results in Figure 5.14 resemble the equivalent comparison in amphibians, there are more significant sequences with a faster rate in the teleost compared to sturgeon/gar than there are in any within order comparisons (Table A.43, Appendix page 240). A number of sequences are evolving slower in *Acipenser ruthenus* than in gar, but this proportion is overshadowed by the number evolving slower than in teleosts.

This analysis was repeated in archosaurs, as shown in Figure 5.15 and Table A.44 (Appendix, page 240). Interestingly, the difference between the two studies (between and within orders) is not as substantial as what was observed in amphibians and actinopterygians. This suggests that archosaur sequences have a greater variation in rates of evolution, irrespective of the mode of PGC specification, than either amphibians and actinopterygians.

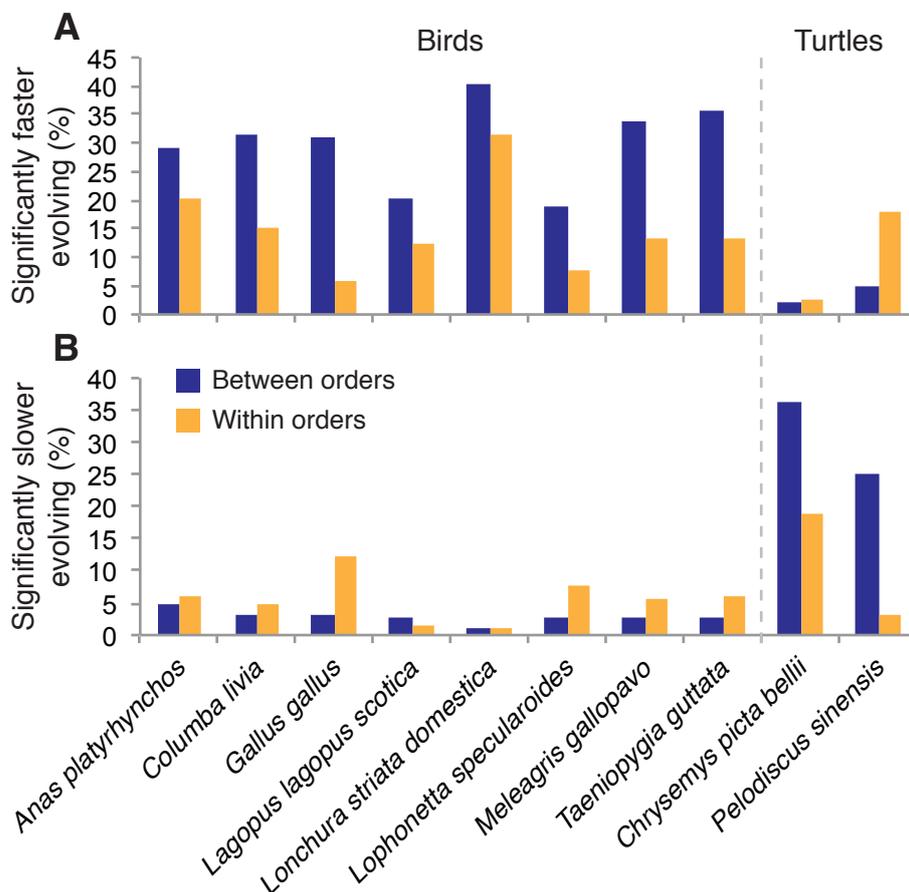


Figure 5.15: Archosaur and testudine RRT results between and within orders. The proportion of sequences evolving significantly faster (A) and slower (B) are shown for comparisons between orders (dark blue) and within orders (orange).

However, the proportion of sequences evolving significantly faster in birds than crocodile/turtles exceeds the proportion evolving at a different rate within

the order. Therefore, the overall result is the same as in amphibians and actinopterygians.

Summary

For the three classes of taxa with reliable relative rate test results, amphibians, actinopterygians and archosaurs, we have shown that the number of sequences with a significant difference in rate is higher in comparisons between orders than when comparing within orders. This pattern was consistent even when comparing between species which diverged at a similar point in time (sturgeons, gar and teleosts). This therefore suggests that the changes in rate of evolution associated with the mode of PGC specification exceed the normal variation in rate seen between species. As such it supports our results that the observed acceleration in anurans, teleosts and birds is unique to these taxa, all of which have acquired preformation.

5.3 Gene Characterisation

As in Chapter 4 we wished to annotate the genes which show an incongruent phylogeny or significant difference in the rate of evolution. Using the same methods as before, the sequences were mapped to a single genome. To simplify the analysis, we have only mapped the results to mouse. The new sequences were able to map to 18,953 mouse genes, of which 13,747 had been tested using the relative rate test, and 11,592 using four-taxon trees. The result was allocated based on whether any orthologs had shown either an incongruent topology or significantly different rate of evolution.

Considering the situation in lepidosaurs, we have calculated the results with and without these species. The results from the four-taxon trees are shown in Figure 5.16, comparing these to the original results (Figure 4.1, page 103) shows that there are now higher proportions of species phylogeny trees, particularly when using the SH-test. This is unsurprising considering the increased number of sauropsid sequences in this enlarged dataset.

Mapping the relative rate test results to mouse (Figure 5.17), shows that the majority of mouse genes have an ortholog with a significantly faster rate in the taxa that has acquired preformation. In fact, comparing this to the data from Figure 4.3 shows that the proportion of genes with this result has increased from

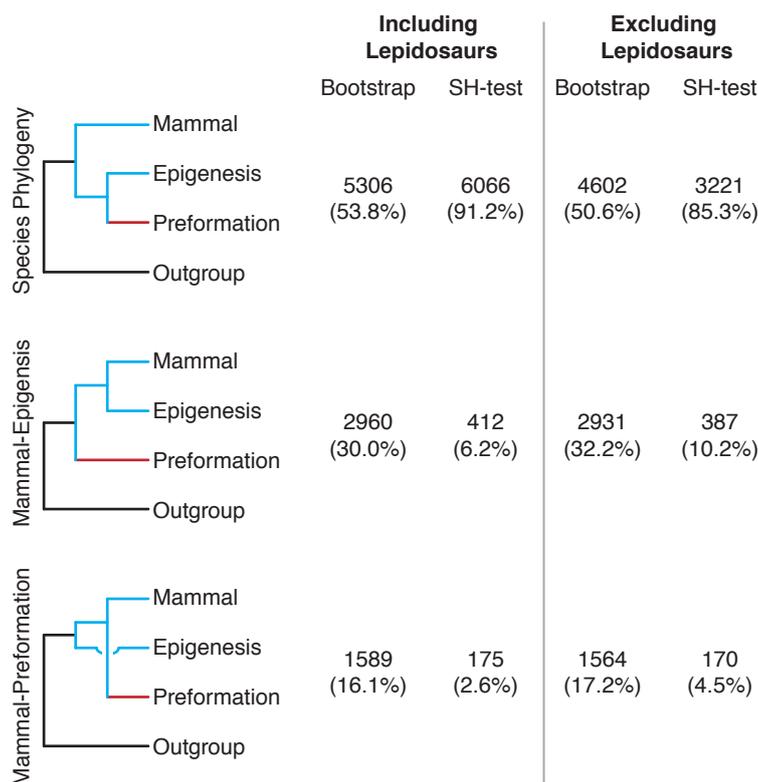


Figure 5.16: The four-taxon tree results mapped to mouse. The bootstrap and SH-test results from Section 5.1 (page 124) are mapped to the mouse genes. The proportion of mouse genes with an ortholog showing a significant tree topology are shown. These results are shown including and excluding lepidosaurs.

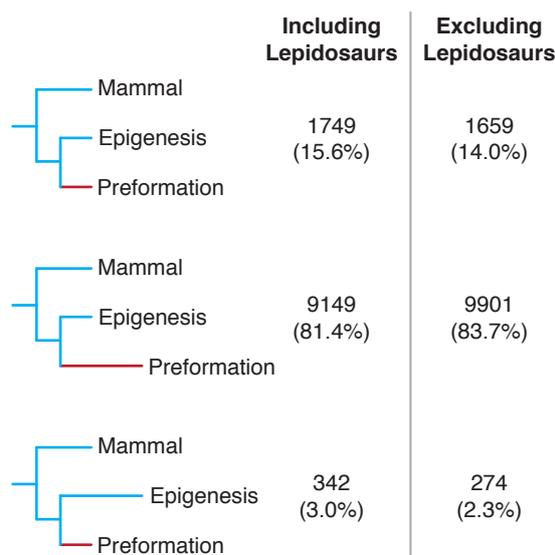


Figure 5.17: The relative rate test results mapped to mouse. As in Figure 5.16, but showing the relative rate test results, including and excluding the lepidosaurs.

70% to over 80%. This proportion is highest when excluding the lepidosaurs, although not by a great extent. Interestingly there are more mouse genes with results we can analyse when the lepidosaurs are excluded, this is because the number of ambiguous results decreases from 2,507 to 1,613.

The genes with an assigned Mammal-Epigenesis topology significant by the SH-test are shown according to the time of expression in Figure 5.18. This shows that there is an over-representation of Mammal-Epigenesis genes in early developmental stages. However, this difference is not significantly different to the average (Chi-squared test, Bonferroni corrected; $p > 0.05$; 1 d.f.). As we saw with the total proportion of tree results, excluding the lepidosaurs does not drastically alter the result.

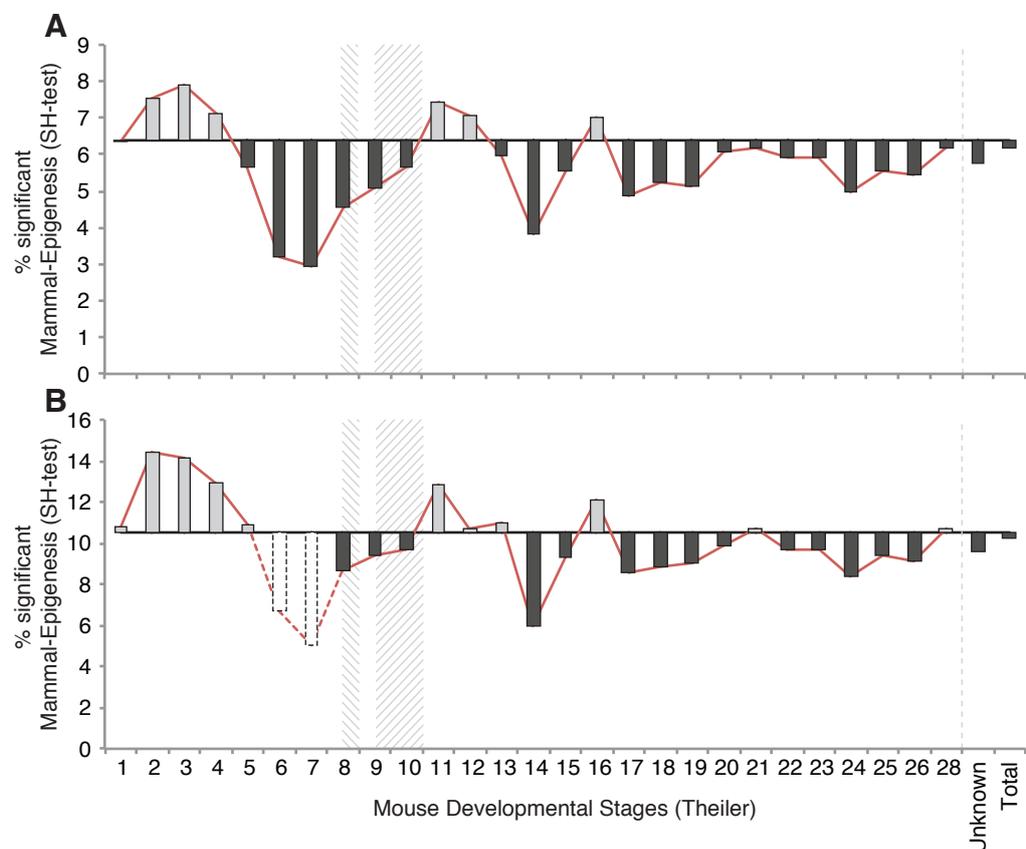


Figure 5.18: The mouse genes with Mammal-Epigenesis trees by the SH-test are plotted by stage of expression. The proportion of genes expressed at each stage with a Mammal-Epigenesis tree significant by the SH-test are compared to the total proportion of Mammal-Epigenesis genes with known expression (horizontal line). (A) includes lepidosaurs, (B) excludes the lepidosaurs. The difference in the results is tested by the chi-squared test (Bonferroni corrected; $*p < 0.05$; 1 d.f.). The stages of PGC induction (late TS 8) and gastrulation (TS 9.5-10) are hashed. Stages with fewer than 20 total genes are showed as dashed outlines.

The relative rate test results are mapped to the mouse genes and plotted according to when they are expressed in Figure 5.19. Compared to the previous results (Figure 4.6), this now shows an over-representation of preformation-faster genes expressed in TS2-5, however it is insignificant. All stages after this point have an under-representation of these genes, the same pattern observed previously. Interestingly, there is no difference in the pattern of gene expression when the Lepidosauria are excluded (Figure 5.19B), even though the overall proportion of preformation-faster genes increases.

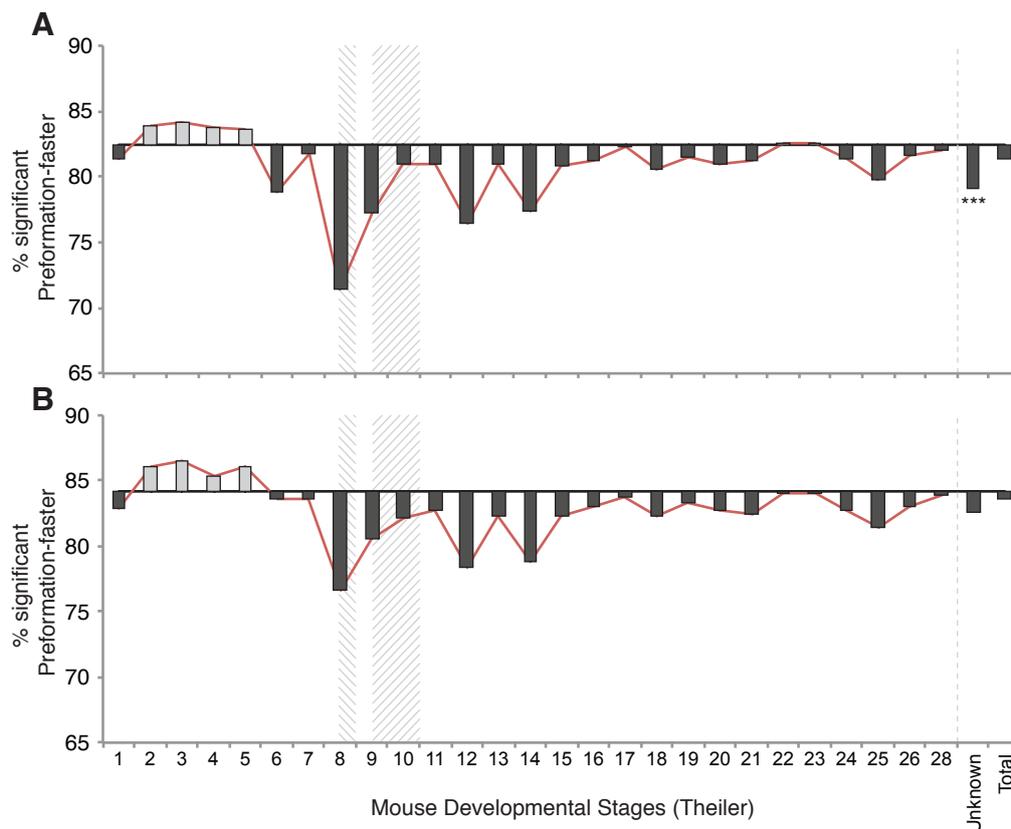


Figure 5.19: The mouse genes with an assigned preformation-faster RRT result are plotted by stage of expression. The figure is presented as in Figure 5.18. (A) includes lepidosaurs, (B) excludes the lepidosaurs. The difference between the proportion of preformation-faster genes in each stage compared to the overall proportion is tested by the chi-squared test (Bonferroni corrected; $*p < 0.05$, $**p < 0.01$; 1 d.f.). The stages of PGC induction (late TS 8) and gastrulation (TS 9.5-10) are hashed.

The only other test of gene identification that had proved interesting in Chapter 4 was the age of the genes. We therefore used the same age classification from Section 4.4 (page 117), and looked at the new subset of mapped genes. We asked whether the proportion of genes with a preformation-faster result in each age category differed to the proportion for all genes with a known age

(Figure 5.20). This shows that, as we saw before, the preformation-faster genes are significantly over-represented in the oldest gene ages. Genes which appear to have arisen from Eumetazoa onwards are significantly under-represented with preformation-faster results. This result is independent of the lepidosaurs (Figure 5.20B).

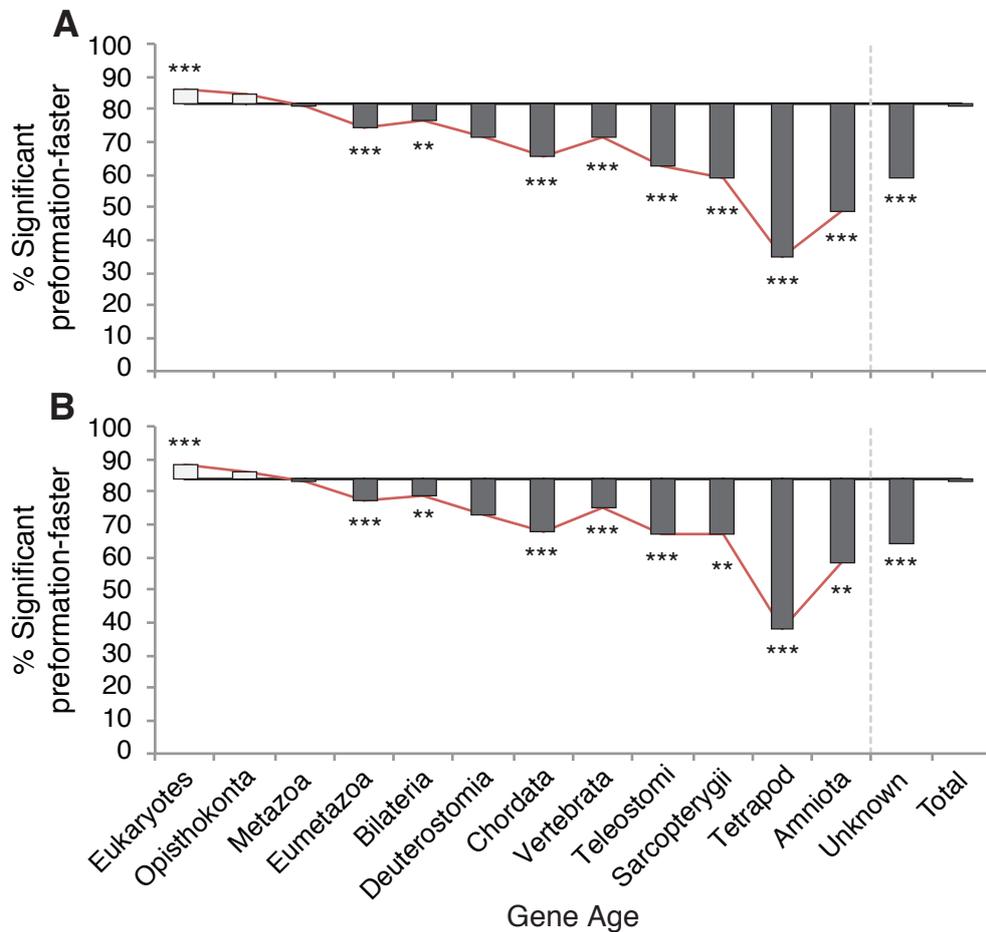


Figure 5.20: The mouse genes with an assigned preformation-faster RRT result are plotted by their age category. The proportion of genes in each age category (last common ancestor of known orthologs) with a preformation-faster results in the relative rate test is shown. This is compared to the overall proportion (horizontal line) of all genes with a known age category. (A) includes lepidosaurs, (B) excludes the lepidosaurs. The difference between each age category and the overall value is tested by the chi-squared test (Bonferroni corrected; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; 1 d.f.).

Using the larger dataset that incorporates more amphibian, sauropsid and actinopterygian sequences has allowed us to map our results to more mouse genes than the original query dataset. However, the genes with a Mammal-Epigenesis tree topology or a preformation-faster RRT result appear to follow the same patterns. In this regard we have observed that, as in our original

dataset, these genes are typically expressed during early development and are ancient genes with a known ortholog in the last common ancestor of Eukaryotes.

5.4 Coelacanth, Lungfish and Sharks

To expand our analysis beyond amphibians, actinopterygians and sauropsids we have investigated coelacanth, lungfish and sharks; all of which are predicted to be undergoing epigenesis (Section 1.3). However, none of these have a sister taxa that has acquired preformation, and therefore cannot be analysed using the same methods as in previous sections. Instead, we compared them against their closest relatives which contain both modes of PGC specification. Sharks are compared against teleosts and sturgeon/gar, while coelacanth and lungfish are compared against anurans and urodeles.

Along with the sequence collection described previously we have added our own novel transcriptome from the Australian lungfish *Neoceratodus forsteri*. We have also added in the publicly available transcriptomes from the lungfish *Protopterus annectans*, the skate *Leucoraja erinacea* and the dogfish *Scyliorhinus canicula*. We also used the Ensembl known cDNA collection for the coelacanth *Latimeria chalumnae*.

5.4.1 Four-taxon trees

The first type of 4-taxon tree we built for these species included their closest relative undergoing preformation, a mammal and an appropriate outgroup. The three possible topologies can be seen in Figure 5.21. For sharks we have used amphioxus as the outgroup, for coelacanth and lungfish the outgroup is a teleost.

The tree topologies that are significant by bootstrapping are shown in Figure 5.22. This shows that for coelacanth and lungfish, 70-80% of the trees reflect the species phylogeny. Within the remaining incongruent trees, there are more Mammal-Lungfish topologies than Anuran-Lungfish in all species excepting *Protopterus annectans*. In sharks, each species shows only 25% of the significant tree topologies reflect the species phylogeny (Table A.45, Appendix page 241). The vast majority group mammals with sharks, the two species that have retained epigenesis.

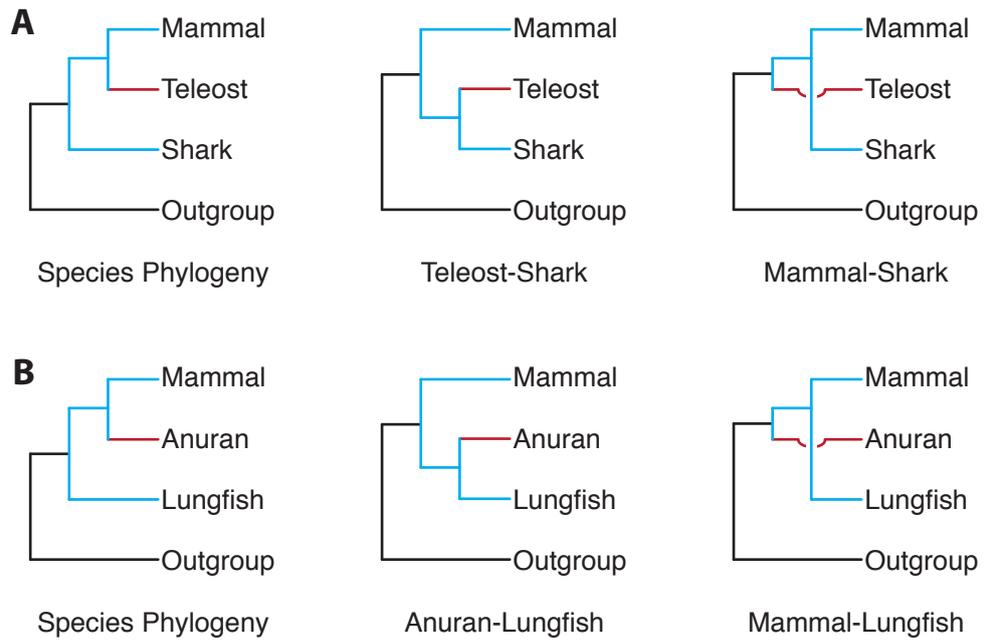


Figure 5.21: The three possible tree topologies. Using either sharks (A) or lungfish (B) the figure shows the three possible tree topologies. In both cases they will either show the species phylogeny, the query species grouping with the species undergoing preformation, or shark/lungfish grouping with mammals. Coelacanth will have the same possible tree topologies as lungfish.

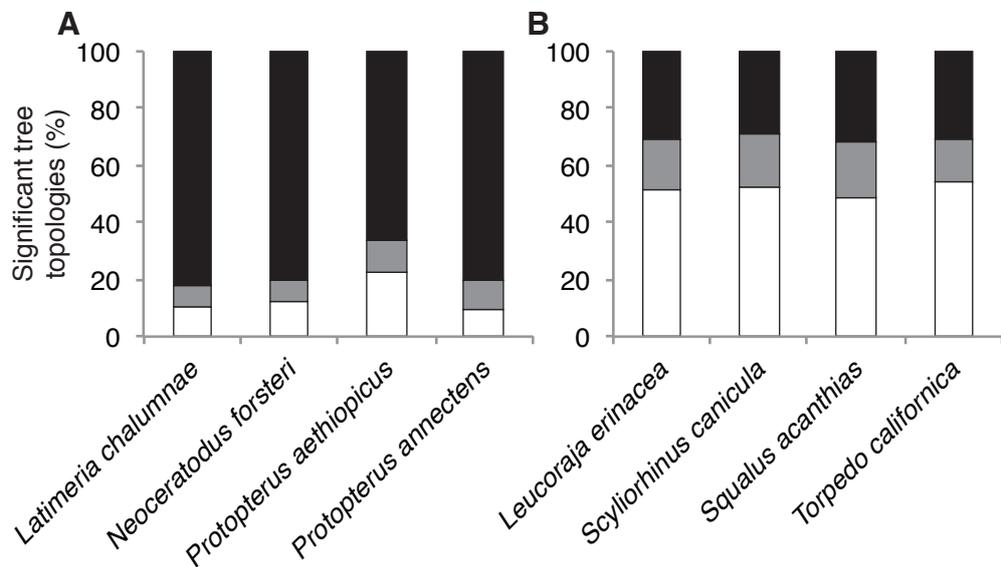


Figure 5.22: The tree topologies significant by bootstrapping. (A) The coelacanth and lungfish trees show either the species phylogeny (black), an Anuran-Lungfish topology (grey) or a Mammal-Lungfish topology (white). (B) For the sharks, the species phylogeny trees are shown in black, the Teleost-Shark topologies are shown in grey, and the Mammal-Shark trees in white.

To explore whether this pattern continued to occur under stringent conditions, we used the SH test. Figure 5.23 shows that for coelacanth and lungfish almost all of the significant tree topologies reflect the species phylogeny. There are almost no incongruent topologies, with no clear bias within them (Table A.46, Appendix page 241). The sharks however continue to show that the majority of significant tree topologies show a Mammal-Shark grouping. In fact, the proportion of these topologies has increased using the more stringent method, something that only the species phylogeny has done until now.

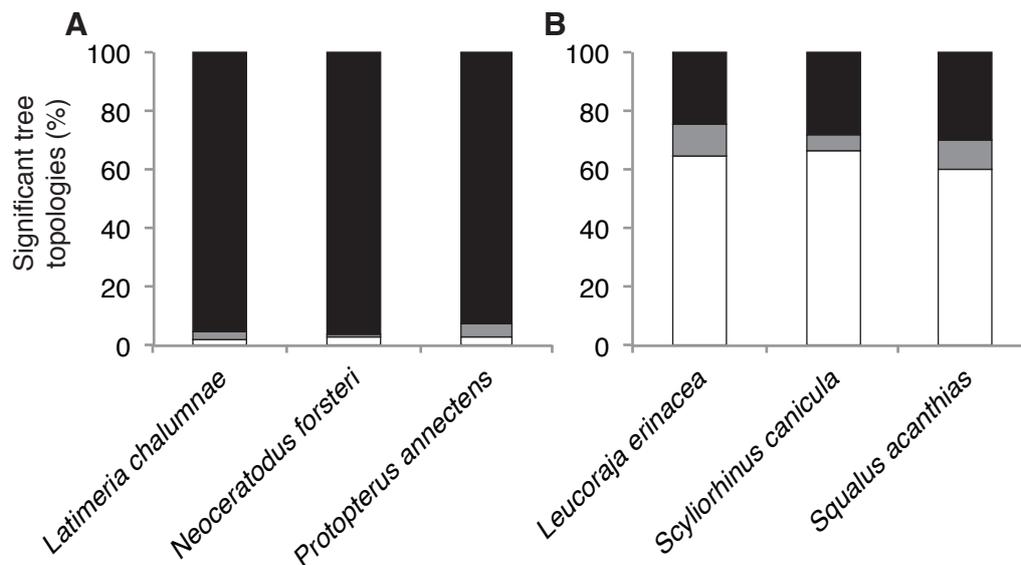


Figure 5.23: The tree topologies significant by the SH test. The tree topologies are shown in the same colours as Figure 5.22 for the coelacanth and lungfish (A) and the sharks (B).

In total, for coelacanth and lungfish we see that of the 7,896 four-taxon trees analysed, 4,740 were significant by bootstrapping. Of these, 3,832 (80.8%) reflected the species phylogeny, 394 (8.3%) grouped the Lungfish/Coelacanth with Anurans and 514 (10.8%) showed the opposite incongruent topology. Within sharks we were able to build a total of 3,178 four-taxon trees, of which 1,534 were significant by bootstrapping. Within these, 466 (30.4%) recapitulated the species phylogeny, 275 (17.9%) showed a Teleost-Shark topology and 793 (51.7%) showed sharks to be grouped with mammals. Therefore there is still a bias within the incongruent topologies which groups species that have retained epigenesis together. The bias towards a Mammal-Epigenesis topology is much stronger in sharks than in any other species analysed.

Four-taxon phylogenies excluding mammals

We also investigated the proportion of incongruent topologies for trees built using both closely related species and an outgroup (Figure 5.24). This provides a positive and negative result within the incongruent trees, i.e. whether the species groups with the taxon that has also retained epigenesis, or the taxon that has acquired preformation. Lungfish and coelacanth trees are therefore being built with both anurans and urodeles while sharks are analysed with teleosts and sturgeon. The outgroups remained the same.

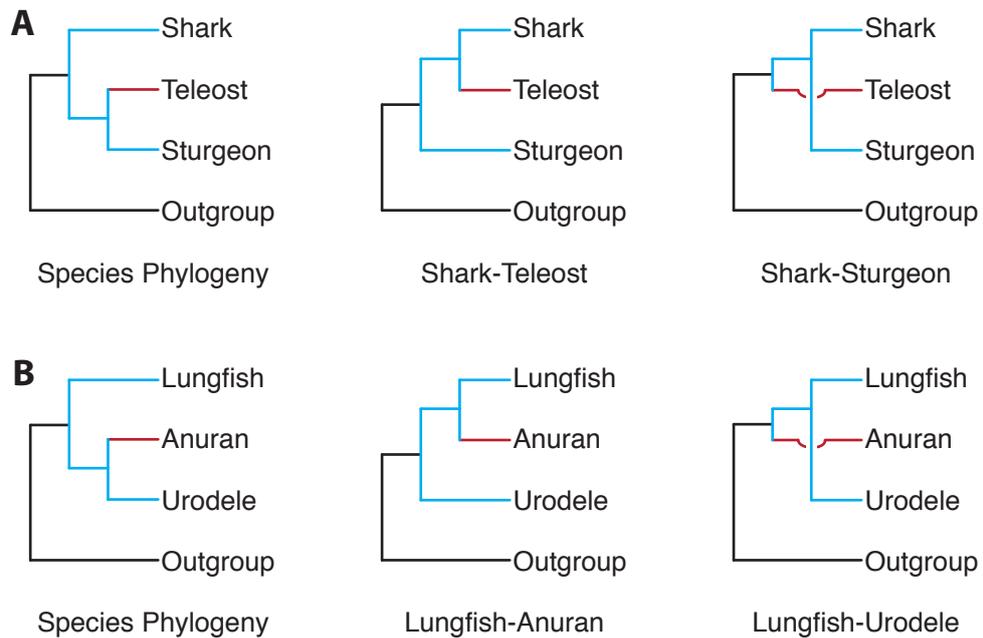


Figure 5.24: The three possible tree topologies when including both of the closest related species. For sharks (A) and lungfish (B) the three possible topologies are shown, the species phylogeny, where the species tested groups with the taxon undergoing preformation, or when it groups with the taxon undergoing epigenesis.

Figure 5.25 shows that contrary to what we saw previously (Figure 5.22), each species shows the species phylogeny in the majority. For coelacanth and lungfish approximately 90% of the trees in each species show the species phylogeny. In sharks there are approximately 70% of significant trees reflecting the species phylogeny. In both groups, within the incongruent trees there is a bias towards grouping with the other taxon that has retained epigenesis (Table A.47, Appendix page 242).

To confirm these results we used the stringent SH test, this shows an increase in the proportion of species phylogeny trees across all taxa (Figure 5.26;

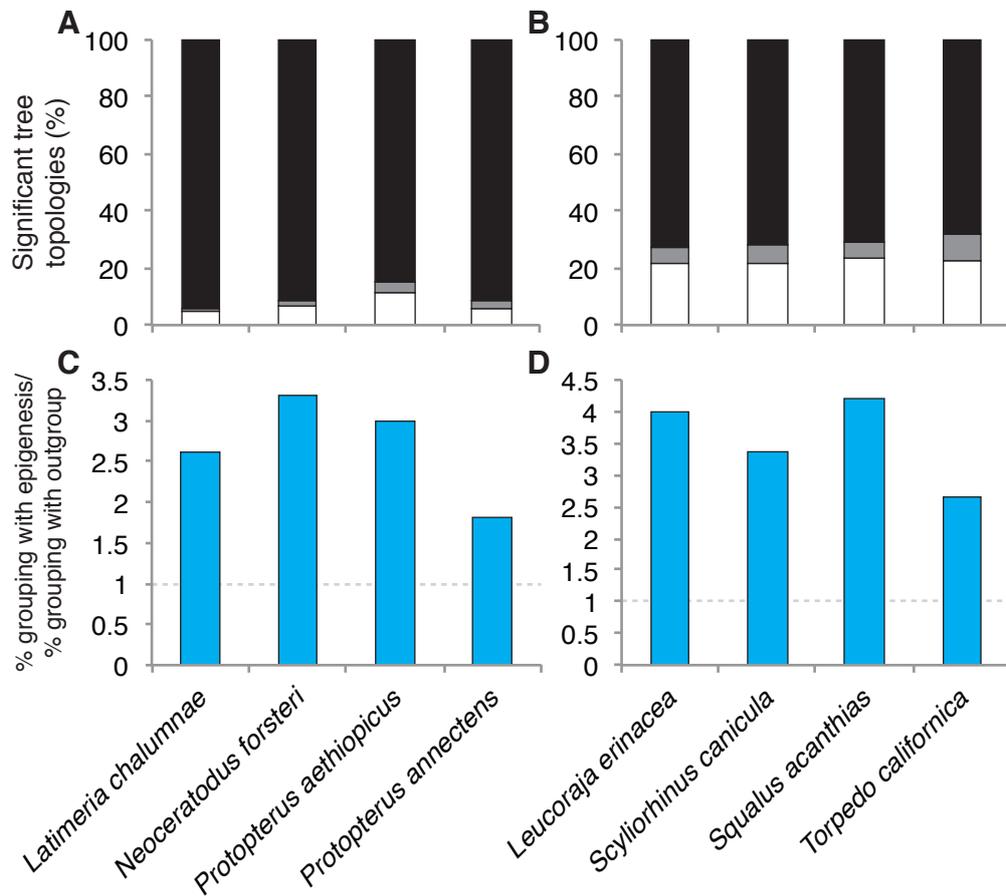


Figure 5.25: The significant bootstrapped topologies. The top panels (A and B) show the proportions of the three possible tree topologies, for Coelacanth and Lungfish (A) and Sharks (B). In (A) the species phylogeny is shown in black, the Lungfish-Anuran trees in grey and Lungfish-Urodele in white. In (B) the Shark-Teleost topologies are in grey and the Shark-Sturgeon trees in white. The bottom panels (C and D) show the likelihood of each species grouping with the taxon undergoing epigenesis when the tree is incongruent.

Table A.48, Appendix page 242). Within the few incongruent trees, there is a very strong bias towards the two species undergoing epigenesis grouping together. In four of the species (*Latimeria chalumnae*, *Protopterus aethiopicus*, *Leucoraja erinacea* and *Scyliorhinus canicula*) this is the only incongruent topology.

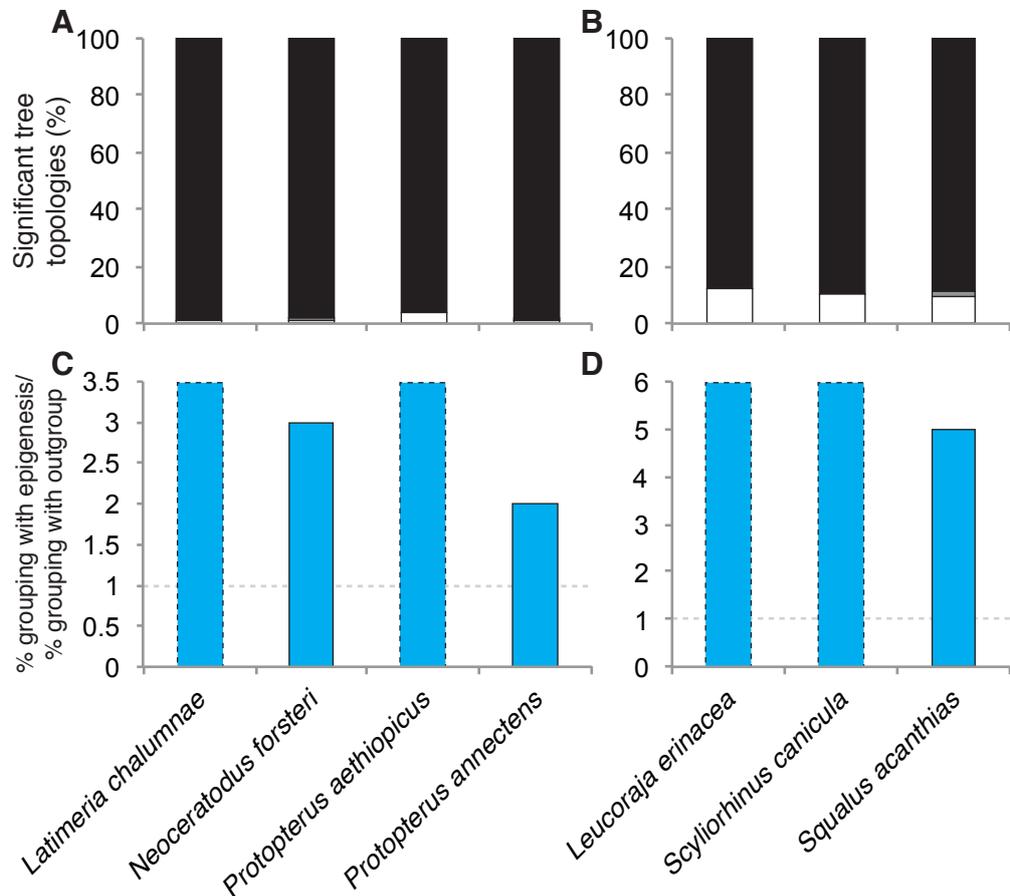


Figure 5.26: The tree topologies significant by the SH test. The results are shown in the same format as Figure 5.25. *Latimeria chalumnae*, *Protopterus aethiopicus*, *Leucoraja erinacea* and *Scyliorhinus canicula* only group with the taxon undergoing epigenesis when the tree is incongruent and are therefore shown with full bars and dashed outlines.

Overall we were able to build 6,791 four-taxon trees utilising coelacanth and lungfish, 4,978 of these were significant according to the bootstrap support. Within these, 4,597 (92.3%) reflected the species phylogeny, 111 (2.2%) grouped the species with anurans and 270 (5.4%) grouped lungfish and urodeles together. Using sharks we were able to build 2,593 trees, 1,449 of which were significant by bootstrapping. Of these, 1,035 (71.4%) recapitulated the species phylogeny, 91 (6.3%) had a Shark-Teleost topology and 323 (22.3%) grouped the two species undergoing epigenesis together.

The 4-taxon trees in these species have shown that even when we cannot compare between sister taxa there is still a bias towards incongruent trees that group the modes of PGC specification together. We have also seen that this bias is very strong in sharks, so much so that sharks are more likely to group with mammals than recapitulate their true species phylogeny.

5.4.2 Relative Rate Test

Since we must always compare sister taxa in the relative rate test, we cannot use mammals as the reference. Therefore amphioxus is used for the reference and each taxon is compared against its closest relatives, either undergoing epigenesis or preformation. So for example the relative rates between shark and teleost are tested, as well as between shark and sturgeon.

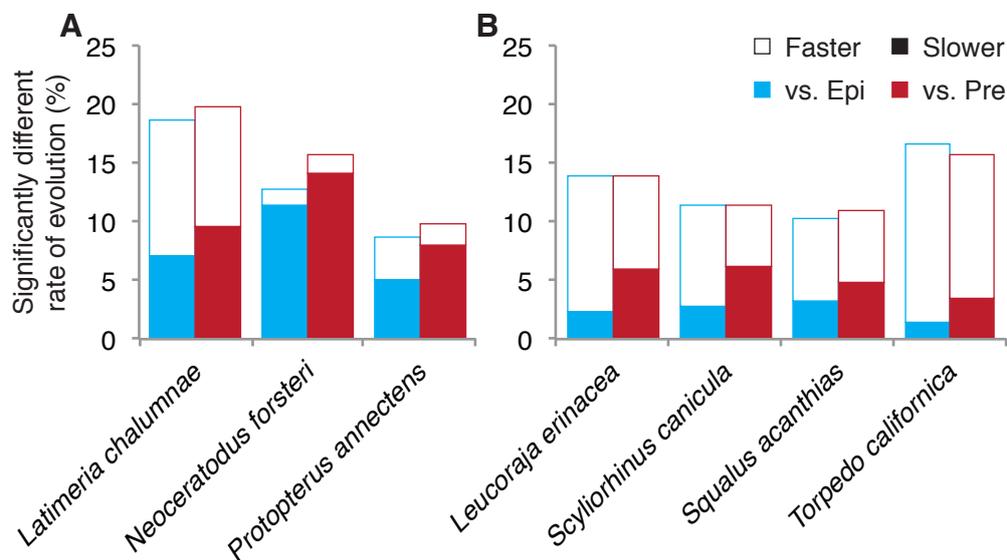


Figure 5.27: The Relative Rate Test results. For Coelacanth and Lungfish (A) and Sharks (B) the proportion of sequences evolving significantly slower (filled bars) and faster (clear bars) is shown when compared to either the closest species undergoing epigenesis (blue) or preformation (red). Number of sequences are shown in Table A.49 (Appendix, page 243).

The results in Figure 5.27 show that in each species there are more slower evolving sequences when compared to the taxon undergoing preformation than there are compared to the taxon using epigenesis. This reflects the difference in rate between sister taxa (anurans and urodeles, and teleosts and sturgeon) that we saw in Section 3.4 (page 87).

However, within sharks there are consistently more sequences evolving faster when compared to both teleosts and sturgeon. This high proportion of

sequences evolving significantly faster is also shown in coelacanth (*Latimeria chalumnae*) but not lungfish. The two species of lungfish appear to have more sequences evolving significantly slower than both anurans and urodeles.

These results are surprising considering the phylogenetic trees showed a bias towards coelacanth, lungfish and sharks all grouping with the species using epigenesis. Based on the results in Section 3.5, we expected to see a large number of sequences with a significantly faster rate in the taxon using preformation, leading to LBA. Indeed this is the result we observe in lungfish, which show a slower rate than anurans. However, the majority of coelacanth and shark sequences with a significant difference in rate are evolving faster than in anurans and teleosts.

5.5 Conclusion

By adding new datasets that have recently become available we have been able to study a larger number of genes and species. We have enlarged the analyses conducted in Chapters 3 and 4, as well as analysing new species such as coelacanth, lungfish and sharks. Enlarging the datasets for amphibians, sauropsids and actinopterygii has shown that, as when we added the axolotl transcriptome, the bias becomes more pronounced but does not change direction. This was true for both the four-taxon tree phylogenies, which showed a bias towards Mammal-Epigenesis incongruent trees, and the relative rate tests which demonstrated a significantly faster rate of evolution in the taxa that have acquired preformation. We also showed that these genes continue to be predominantly expressed during early mouse development and are ancient genes with orthologs in the eukaryote common ancestor.

Within the lepidosaurs, however, where we added a new lineage (skinks) we did observe a discrepancy in the results. The skinks, undergoing preformation, are evolving significantly slower than the Iguanidae lizards which have supposedly retained epigenesis. We do not know whether this truly contradicts the results we have observed in all other vertebrates, or whether it is due to our limited knowledge on the mode of PGC specification in these organisms. It would be interesting to further explore lepidosaurs in the future, especially since two snake genomes have been recently published (Castoe et al., 2013; Vonk et al., 2013).

The analyses on coelacanth, lungfish and sharks have shown that once again there is a consistent bias within gene trees that incongruently group the species which have retained epigenesis together. This bias was particularly strong for the shark genes in a tree with mammals and teleosts, the vast majority grouped mammals and shark together contrary to the species phylogeny. However we were unable to observe a corresponding pattern in the relative rate test results, in fact the majority of shark genes with a significant difference in rate were evolving faster than teleosts. This was surprising considering whole genome studies have placed sharks and coelacanth as the slowest evolving vertebrates (Amemiya et al., 2013; Venkatesh et al., 2014). However, these studies were based on four-fold degenerate sites and so differed from the analyses we have performed. Since we cannot compare sister taxa in these species, the rate results may reflect evolutionary changes that are irrespective of the mode of PGC specification, for example an acceleration at the base of Teleostomi.

Analysing Pluripotency Genes

In previous analyses we took an unbiased approach, analysing all available genes using simple phylogeny and rate tests. To complement this, we studied specific genes in detail. We chose to investigate pluripotency factors because they are expressed at the earliest stages of development, which we showed is associated with a faster rate of evolution in taxa undergoing preformation (Section 4.3, page 109). These genes are also expressed in primordial germ cells and so may be affected by PGC specification.

The pluripotency genes we have analysed are Oct4 (otherwise known as Oct3 or POU5F1), Sox2, Klf4 and Nanog. The first three are part of the Yamanaka factors, the first genes known to reprogram a somatic cell back into a pluripotent state (Maherali et al., 2007; Okita et al., 2007; Takahashi et al., 2007; Takahashi and Yamanaka, 2006; Wernig et al., 2007). Nanog is required for the final stages of reprogramming and the acquisition of pluripotency (Silva et al., 2009; Yu et al., 2007). Other pluripotency factors include Myc and Lin28 but due to time constraints have not been analysed here.

Oct4 was discovered by its ability to bind to the octamer binding motif located in promoter and enhancer regions (Okamoto et al., 1990; Schöler et al., 1989). The gene was then cloned in mice and revealed to contain both a homeodomain and a POU-specific domain (Rosner et al., 1990; Schöler et al., 1990b). It was shown that Oct4 was expressed in oocytes, blastocysts and the epiblast before being localised to the PGCs (Schöler et al., 1990a). When mouse embryos are deficient of Oct4, the embryo develops to the blastocyst stage but the inner cell mass is not pluripotent and instead differentiates into extraembryonic trophoblast (Nichols et al., 1998). Conversely, over-expression of Oct4 leads to differentiation of the epiblast into primitive endoderm and mesoderm (Niwa et al., 2000). Oct4 can also partner Sox17 to specify endoderm (Aksoy et al., 2013).

Oct4 expression is limited to the PGCs in mouse after gastrulation; it remains expressed in these cells through migration, proliferation and differentiation into the germ cells (Schöler et al., 1990a; Yeom et al., 1996). If Oct4 is specifically downregulated in the PGCs then the cells apoptose and the resulting organism is sterile (Kehler et al., 2004). Oct4 is therefore required for the maintenance of the PGCs, but not their early specification.

Sox2 interacts with Oct4 to activate FGF4, and is thereby required for the maintenance of the pluripotent inner cell mass and the establishment of the epiblast (Avilion et al., 2003; Yuan et al., 1995). This combination of Oct4 and Sox2 also regulates expression of Nanog (Rodda et al., 2005). Sox2 is known to interact with Sall4 and Esrrb, both of which are involved in maintenance of pluripotency (Hutchins et al., 2013; Tanimura et al., 2013). Sox2 also functions to repress differentiation inducers such as Eomes, Esx1 and Pax6 (Boyer et al., 2005; Masui et al., 2007). In addition to the role of Sox2 in pluripotency, it is also involved in lens (Kamachi et al., 1998) and neural development (Graham et al., 2003).

Krüppel-like factor 4 (Klf4, previously known as GSKF) is a zinc-finger transcription factor enriched in epithelial cells, particularly those of the gastrointestinal tract (Garrett-Sinha et al., 1996; Shields et al., 1996; Ton-That et al., 1997). Klf4 is associated with tumour formation and depending on the context functions either as a tumour suppressor or as an oncogene (Rowland et al., 2005; Ton-That et al., 1997; Yu et al., 2011). Klf4 is also involved in epithelial cell differentiation, for example it is required for the skin's barrier function as well as goblet cell differentiation in the colon (Katz et al., 2002; Segre et al., 1999). Within pluripotency, Klf4 functions along with Sox2 and Oct4 to regulate gene expression, for example all three are required for activation of the Lefty1 promoter (Nakatake et al., 2006). Klf4 is also able to revert stem cells derived from the mouse epiblast back to the naïve ground state of pluripotency (Guo et al., 2009).

Nanog was discovered by its ability to rescue mouse ESC self-renewal in the absence of LIF (leukemia inhibitory factor; Chambers et al., 2003). This function of Nanog is dependent on the formation of a homodimer, which is in turn dependent on the tryptophan repeat (WR) domain (Mullin et al., 2008; Wang et al.,

2008). This domain is not present in the axolotl *Nanog* gene and thus the protein is unable to dimerise, however *Nanog*'s function as a pluripotency factor is retained (Dixon et al., 2010). *Nanog* monomers are able to enhance reprogramming and to limit differentiation in embryoid bodies. Mouse *Nanog* is required for the ICM to develop ground state pluripotency and to progress into a correctly specified and viable epiblast independent of LIF (Mitsui et al., 2003; Silva et al., 2009).

For each of these pluripotency factors we have explored their phylogeny, rate of evolution and pattern of gene loss/gain using the methods described in Sections 2.6, 2.7 and 2.9. We have investigated whether any of these specific genes show a change in their molecular evolution that correlates with the mode of PGC specification.

6.1 *Oct4*

Oct4 is part of the POU gene family, a large group of genes that have been divided into 6 classes of which *Oct4* is a member of the class V group (for review, see Tantin, 2013). The first class consists of just one mammalian gene, *POU1F1* (*Pit1*), and is expressed in the pituitary (Bodner et al., 1988). The second class consists of 3 genes (*POU2F1*, *POU2F2* and *POU2F3*), the first of these is expressed ubiquitously, the second is expressed in the brain and blood and the third is expressed in the epidermis and taste receptors (Tantin, 2013). The third class of POU genes consists of 4 mammalian genes, all of which are expressed in the brain. The fourth class comprises 3 genes that are expressed in the brain and retina. The final class (*POU6*), consists of two genes both expressed in the brain and central nervous system (Zhang et al., 2013).

The *POU5* class consists of *Oct4* (*POU5F1*), as well as two other genes, *POU5F2* and *POU5F1B*. The latter of these is a processed pseudogene and may play a role in cancer development (Kastler et al., 2010). *POU5F2* is expressed in developing spermatids and is required for optimal function of the male germ cell (Pearse et al., 1997). Outside of mammals, a *POU5F1* homolog was identified in zebrafish and named *pou2* (Takeda et al., 1994). As with the mammalian homolog, *pou2* is expressed in all cells of the developing blastula. After gastrulation *pou2* expression is localized to the epiblast, it is then localized to the dorsal midline and a transverse band in the lateral region. This latter expression

pattern overlaps the future hindbrain and indeed *pou2* has been shown to be crucial for formation of the midbrain and hindbrain (Burgess et al., 2002). Interestingly, zebrafish embryos that develop with no maternal or zygotic *pou2* are still able to develop germ cells and gastrulate. They are however unable to develop endoderm, and so it appears that *pou2* and Oct4 have different functions (Lunde et al., 2004).

Within *Xenopus*, three Oct4 homologs were isolated, Oct25, Oct91 and Oct60 (Frank and Harland, 1992; Hinkley et al., 1992; Whitfield et al., 1993). These genes are expressed sequentially, with Oct60 expressed in the oocyte and early cleavage stages until gastrulation, Oct25 is then expressed from mid-blastula to gastrulation (Hinkley et al., 1992; Whitfield et al., 1995). Oct91 is also expressed from mid-blastula but the expression peaks at late gastrula before the levels drop off. All three genes are involved in pluripotency during *Xenopus* development and a combined knockdown leads to severe posterior truncations and anterior neural defects (Morrison and Brickman, 2006). Overexpression of these genes blocks formation of mesendoderm by inhibiting FGF and Nodal signalling (Cao et al., 2006). The three *Xenopus* genes are therefore functional homologs and indeed they are found clustered within the *Xenopus tropicalis* genome (Cao et al., 2006; Morrison and Brickman, 2006).

The vertebrate Oct4 genes described above were initially thought to be orthologous (Burgess et al., 2002; Morrison and Brickman, 2006). However, with the advancement of whole genome sequencing it became clear that the *Xenopus* and mammalian genes existed within different syntenic positions (Niwa et al., 2008). Both POU5F1 and *pou2* were discovered in marsupials and monotremes, confirming the suggestion that they are separate genes (Niwa et al., 2008). At this time, it was suggested that the POU5 duplication occurred at the base of mammals, and that the *pou2* gene was lost in eutherians.

Not long after this, the two genes were discovered to co-exist in axolotl (Bachvarova et al., 2004; Tapia et al., 2012). This suggested that the duplication within the gene family had occurred at least at the base of tetrapods (Frankenberg et al., 2010; Tapia et al., 2012). These data proposed that the *pou2* gene in teleosts was ancestral to both genes in tetrapods, and accordingly the *pou2* gene in fish was renamed to POU5F1 by the zebrafish nomenclature committee

(www.zfin.org; Onichtchouk, 2012). This renaming also eliminated the potential confusion between pou2 and the POU2 class of genes.

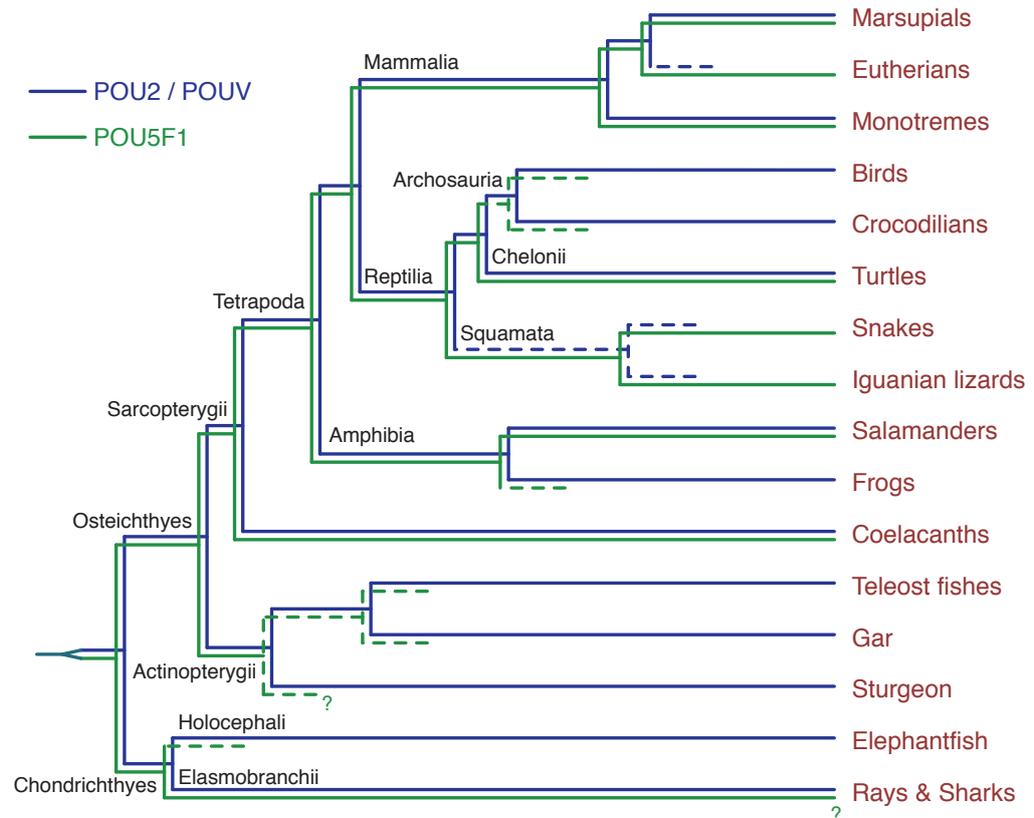


Figure 6.1: The proposed evolution of the POU5 genes. It is currently thought that the gene duplication between POU5F1 and pou2 occurred at the base of gnathostomes and that POU5F1 was lost from Holocephali (Elephant shark), Actinopterygii, frogs and archosaurs (Frankenberg and Renfree, 2013). Pou2 has been predicted to have been lost in squamates and eutherian mammals. Figure adapted from Frankenberg and Renfree, 2013.

Most recently it was suggested that the duplication occurred at the base of vertebrates, based on the presence of EST sequences in the skate, *Leucoraja erinacea* (Figure 6.1; Frankenberg and Renfree, 2013). This was substantiated by the presence of two copies of NPDC1, the neighbouring gene to pou2. In this light, the teleost homolog was once more renamed to POU5F3 (www.zfin.org), a name which reflects the paralogous relationship to POU5F1 in mammals. Throughout the remainder of this section I shall use the term POU5F3 to describe this gene.

Utilising the data in our own novel transcriptomes and the recent genome data that have been published (such as gar and elephant shark; Flicek et al., 2013; Venkatesh et al., 2014) we have sought to further elucidate the relationships within the POU5 class. We have also attempted to locate POU5 class

genes beyond vertebrates, as well as identify their relationship to the other POU classes. To do this we have built comprehensive phylogenetic trees as well as studying the available synteny.

6.1.1 POU gene family

We first sought to place the POU5 genes in terms of their relationship to the other POU classes. We therefore built a phylogenetic tree using a few selected deuterostome species as well as non-deuterostome species (Figure 6.2 and Figure B.13, Appendix page 259). As expected, the topologies within each class are not well defined but the internal branches between POU classes are generally well supported.

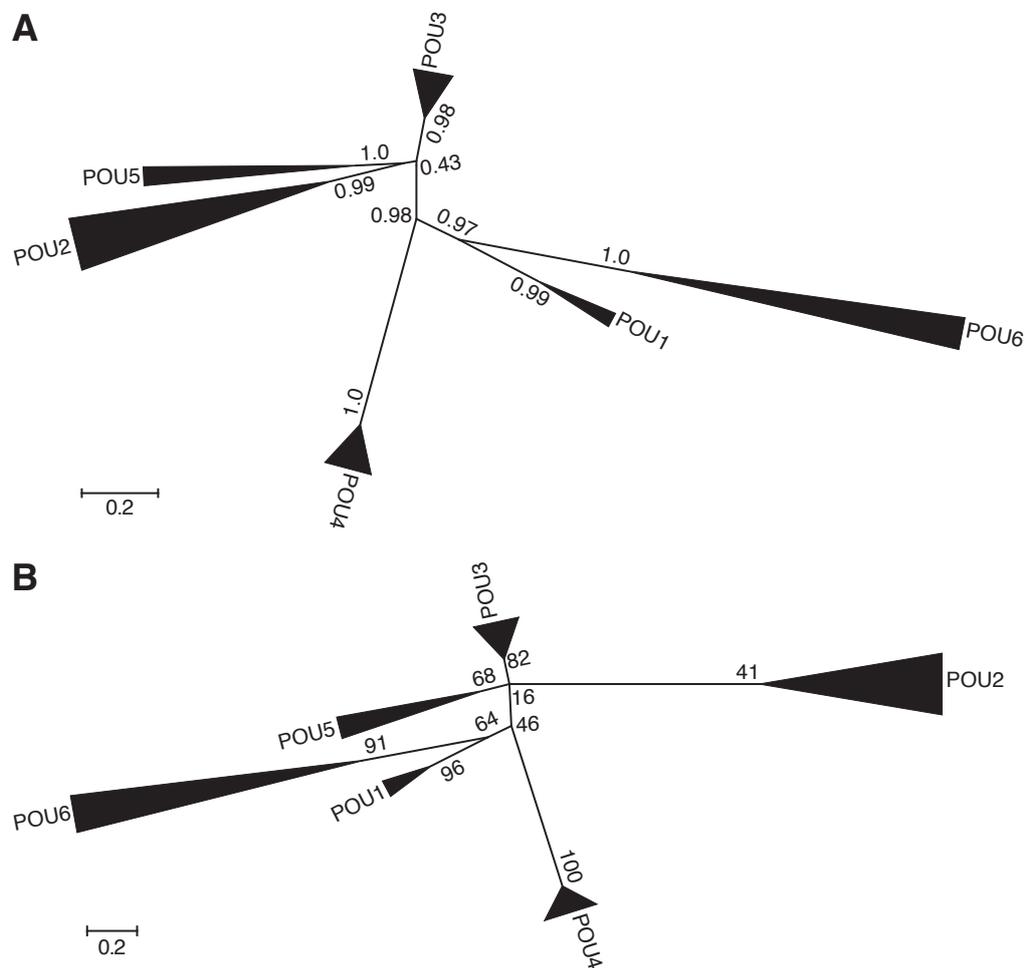


Figure 6.2: POU family phylogeny. The Bayesian (A) and maximum-likelihood (B) trees were created using the protein sequences from a few selected species, each POU class branch was then compressed. The non-compressed tree is shown in Figure B.13 (Appendix, page 259).

POU5 is distinct from the other POU classes and is not grouped within any other clade. In the Bayesian tree POU5 appears most closely related to the POU2 class, poorly supported by a posterior probability of 0.43. The maximum-likelihood tree shows a trifurcating branch between POU5, POU2 and POU3 supported by only 16% of the bootstrap replicates.

The POU6 and POU3 classes were both present in the last common metazoan ancestor as orthologs were identified in *Amphimedon queenslandica* and *Trichoplax adhaerens* respectively. POU1 and POU4 both had orthologs in species which suggest an ancestry within the eumetazoan common ancestor. The POU2 class is found in species which share a common ancestor at the base of bilateria. The only species in the POU5 class are vertebrates, which suggests it is the most recent class within the POU gene family to have evolved. However, the identity of the POU5 sister class cannot be determined.

6.1.2 The POU5 class

Looking in detail at the POU5 class of genes, we first built Bayesian and ML phylogenetic trees using the full-length sequences (Figure B.14, Appendix page 259). Poorly supported branches on the Bayesian tree were then collapsed and the ML bootstrap values added where possible (Figure 6.3). Both the Bayesian and ML methodologies show that the POU5F2 gene is unique to mammals and groups with the POU5F1 mammalian genes. This suggests that a gene duplication at the base of mammals led to the two genes. It is possible that the duplication occurred at the base of eutherians since POU5F2 was not identified in any monotreme or marsupial.

In both trees, an axolotl sequence is grouped with the mammals, suggesting this is a POU5F1 gene. This positioning is significant in the Bayesian phylogeny (Figure 6.3). The Bayesian tree also shows genes from lizard (*Anolis carolinensis*), painted turtle (*Chrysemys picta bellii*) and coelacanth (*Latimeria chalumnae*) grouped in this position but this is not conserved in the ML tree (Figure B.14, Appendix page 259). The POU5F3 marsupial genes are grouped with birds, turtle, alligator and other urodele genes in both phylogenies.

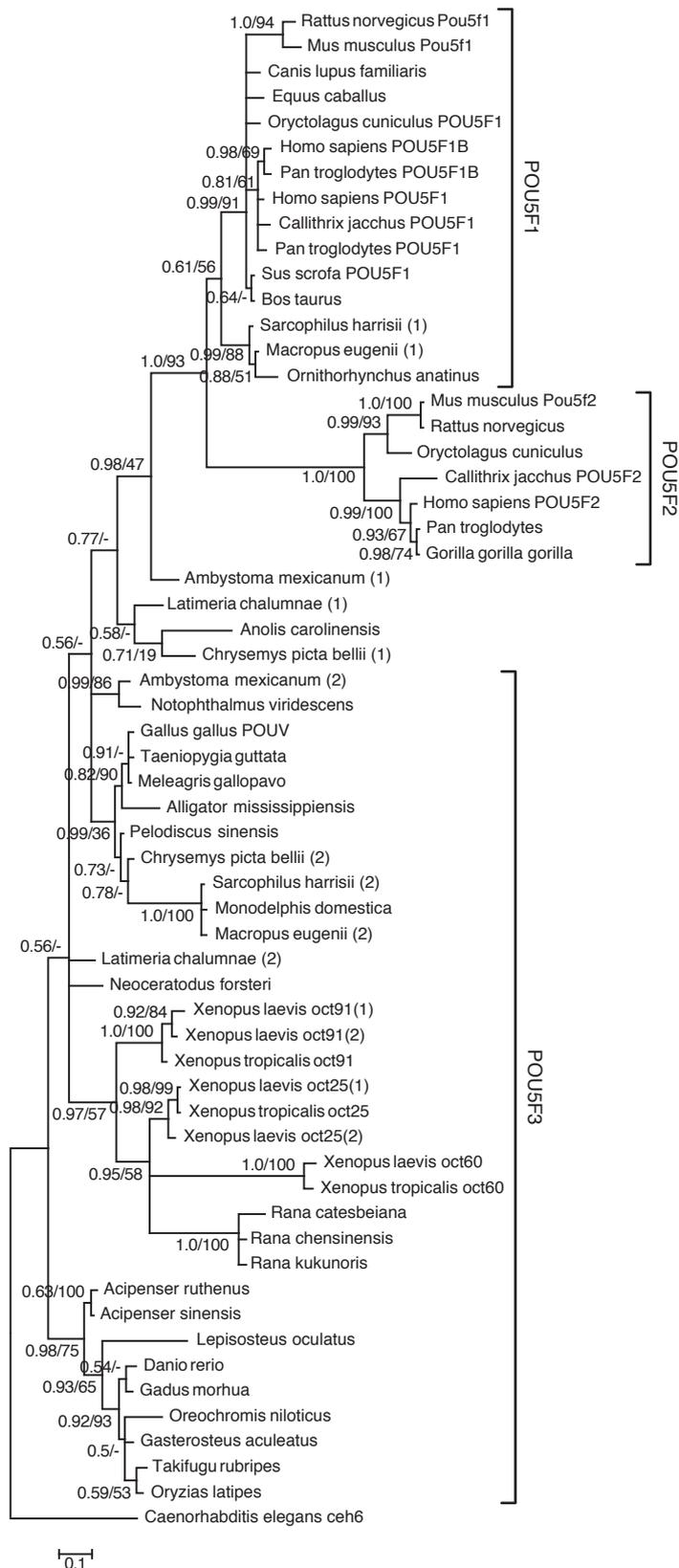


Figure 6.3: The POU5 phylogeny. The Bayesian and ML trees (Figure B.14, Appendix page 259) were built using the full length protein sequences and rooted on the *Caenorhabditis elegans* ceH6 gene (POU3). The Bayesian tree is shown with branches supported by a posterior probability less than 0.5 collapsed. Support values show both the posterior probabilities and ML bootstrap results (PP/BS).

Figure 6.3 shows the three *Xenopus* genes, oct60, oct25 and oct91, grouped together. However, only one gene was identified in each *Rana* species, grouping with oct60 although poorly supported. This suggests that this gene expansion of the POU5 class is unique to *Xenopus* species and not all anurans. The actinopterygians all group together, with high support in both phylogenies. In the Bayesian phylogeny they are positioned at the base of the tree, in the ML phylogeny they are positioned within the POU5F3 group (Figure B.14, Appendix page 259).

Both phylogenies show weak support in the internal branches that form the backbone of the tree. It is therefore difficult to clearly separate the POU5F1 and POU5F3 genes from one another, particularly for species such as lungfish and coelacanth which do not form a well supported or consistent clade with any other species. We therefore required a different method to deduce exactly which gene was present in each species.

We used the available genomic information to compare the neighbouring groups of genes in different species. Figure 6.4 shows that POU5F1 neighbours the TCF19 gene while POU5F3 resides between FUT7 and NPDC1. It is immediately obvious that the synteny surrounding POU5F3 is better conserved than the region surrounding POU5F1. It is also apparent that neither region is conserved within the lamprey, *Petromyzon marinus*.

Looking at the synteny for POU5F3 shows that the region between NPDC1 and CLIC3/A has been reversed in the archosaurs and testudines. The multiple genes in *Xenopus*, Oct60, Oct90 and Oct25 are located at the same locus, as shown previously (Cao et al., 2006; Morrison and Brickman, 2006). *Anolis carolinensis* appears to be missing the POU5F3 gene, however there is no single scaffold across the syntenic region so it could be missing due to incomplete sequencing or assembly. There does appear to be a FUT7 gene in this species, however it is unannotated and due to the highly conserved nature of the FUT genes it is not possible to tell whether this gene is truly FUT7 or another paralog. The POU5F3 gene has definitely been lost from eutherian mammals as the syntenic region is well sequenced and annotated but there is no remaining similarity to the POU5F3 gene.

The entire POU5F1 syntenic region appears to have been lost in birds, I was able to identify some of the far markers (FLOT1, VARS and LSM2) but only

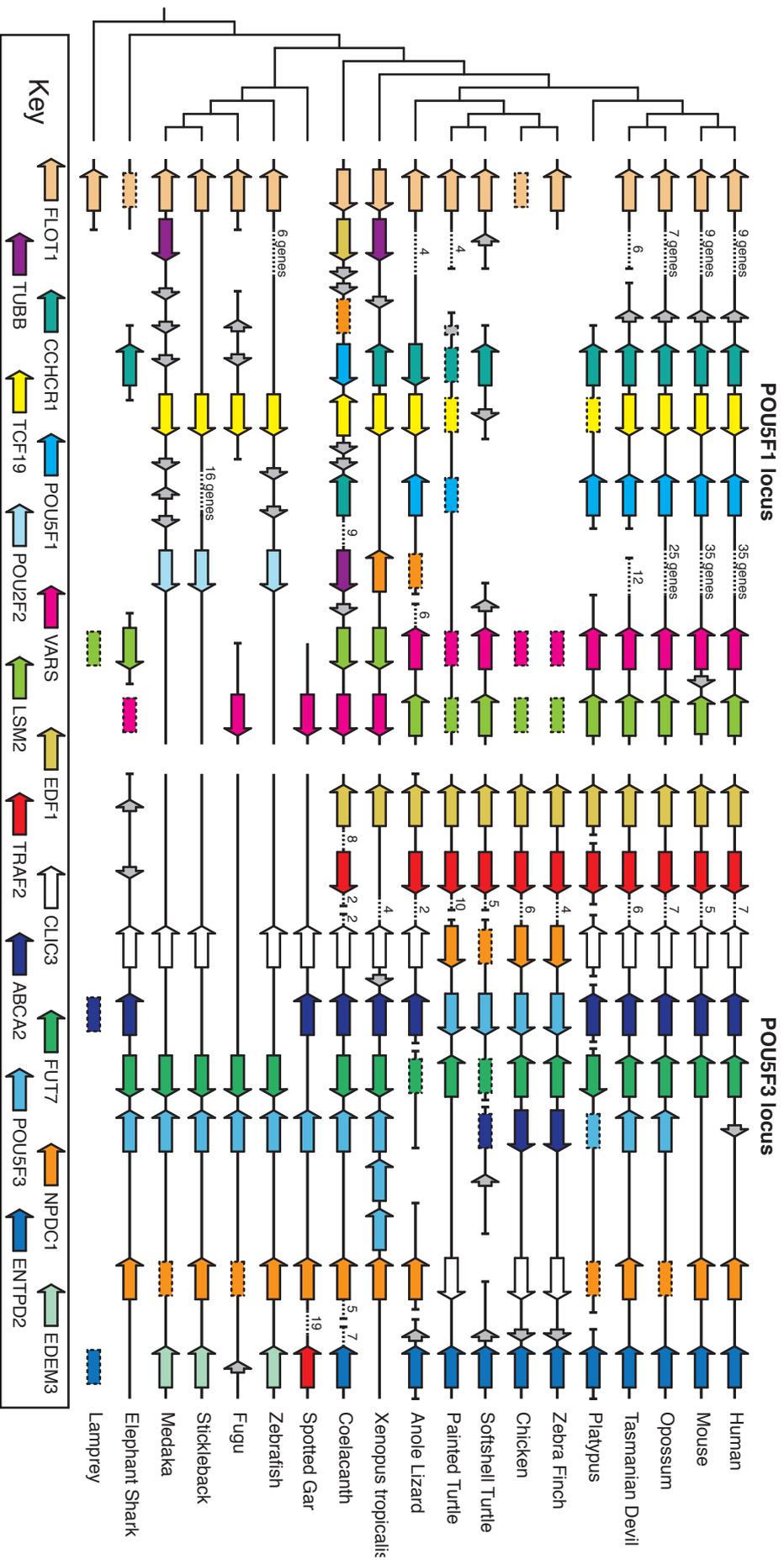


Figure 6.4: The syntenic relationships for POU5F1 and POU5F3. The horizontal lines represent the genomic scaffolds, the ends of which are represented by vertical lines. BLAST results that correspond to a location with no known gene are shown with dashed outlines. Unlocalised BLAST results (i.e. ESTs) are shown without the horizontal line. The distances between genes are arbitrary.

within EST collections. The FLOT1 gene could be located in the zebra finch (*Taeniopygia guttata*) genome, but there were no recognisable genes surrounding it. This region was also poorly conserved, or sequenced, in the soft-shell turtle (*Pelodiscus sinensis*) and elephant shark (*Callorhynchus milii*). Although POU5F1 could be located, this region was also incompletely assembled in the platypus and tasmanian devil.

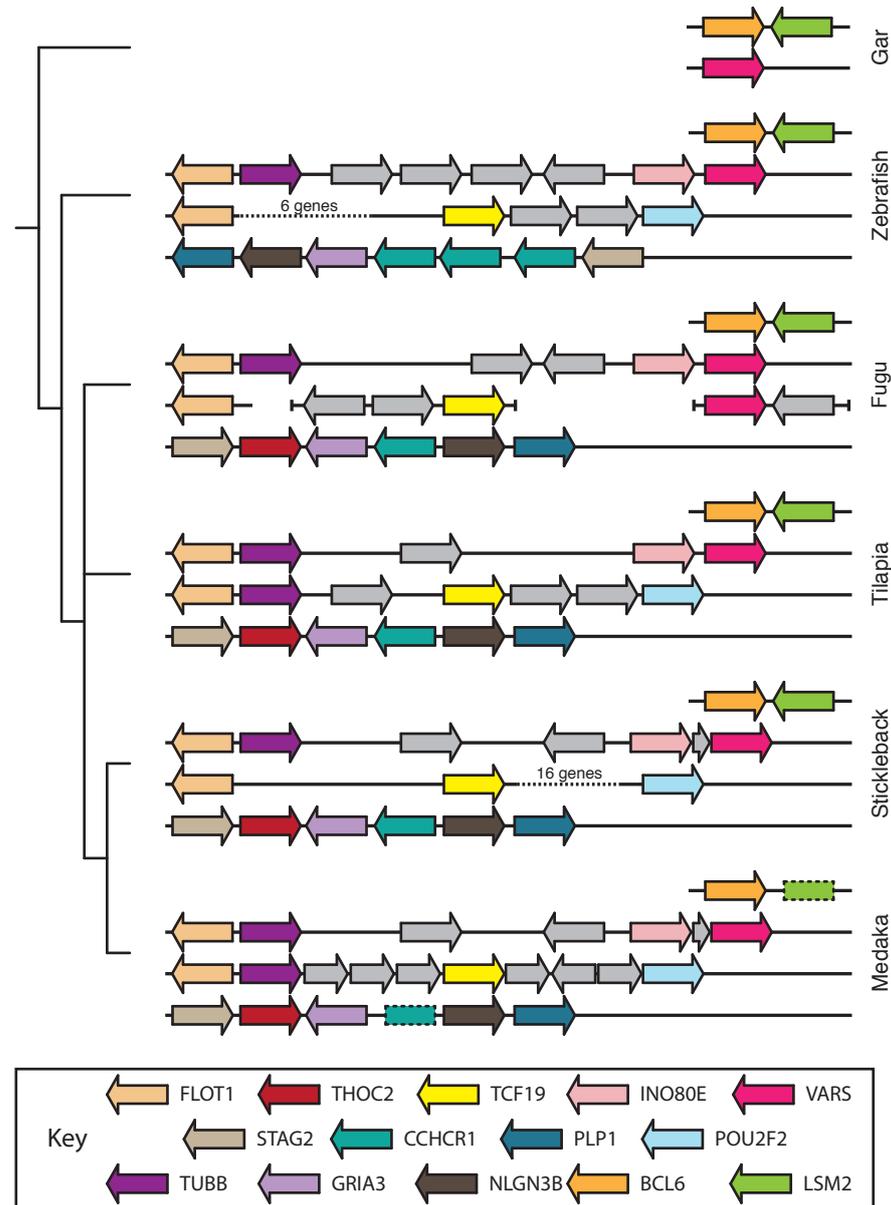


Figure 6.5: POU5F1 syntenic region in Actinopterygii. The syntenic regions are shown as in Figure 6.4 but in more detail over the actinopterygians.

The actinopterygian synteny surrounding POU5F1 is shown in more detail in Figure 6.5. LSM2 and VARS are no longer neighbouring and CCHCR1 is located on a separate chromosome. The whole genome duplication that occurred

at the base of teleost fish appears to have resulted in two copies of the FLOT1 gene, one of which is associated with TCF19 and the other with VARS. Neither of these genomic regions show any similarity to the POU5F1 gene. Interestingly, there is no sign of FLOT1, TCF19 or POU5F1 in the gar, even though VARS is located within the middle of a chromosome. This suggests that either the VARS gene has been translocated in gar (and that FLOT1 and other genes are located on an un-sequenced region), or that these genes have been lost in gar. Overall, this work on synteny suggests that POU5F1 is missing from birds, anurans, actinopterygians and sharks.

Table 6.1: POU5 and neighbouring gene copy numbers in vertebrates.

Species	FLOT1	CCHCR1	TCF19	POU5F1	NPDC1L	LSM2	VARS	EDF1	TRAF2	CLIC3	ABAC2	FUT7	POU5F3	NPDC1	ENTPD2	EDEM3
Human	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
Pig	1	1	1	1	0	1	3	1	1	1	1	1	0	1	1	1
Mouse	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1
Opossum	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Taz. Devil	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Platypus	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Zebra Finch	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
Chicken	1	0	0	0	0	1	1	1	1	1	1	1	1	2	1	1
Turkey	1	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1
Softshell Turtle	1	1	0	0	0	1	1	1	1	0	1	1	1	1	2	1
Painted Turtle	1	1	1	1	0	1	1	1	1	1	0	1	1	1	2	1
Anole lizard	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
Indian Python	1	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1
Axolotl	1	0	0	1	0	3	2	1	2	1	4	2	1	0	2	2
X.tropicalis	1	1	1	0	1	1	1	1	1	1	1	3	1	1	1	1
X.laevis	1	1	1	0	1	1	1	1	1	1	1	5	1	1	1	1
Lungfish	1	0	0	0	1	1	2	1	2	1	3	1	1	1	3	2
Coelacanth	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
Sturgeon	1	0	0	0	0	1	2	1	2	1	4	0	1	1	1	2
Gar	0	0	0	0	0	1	1	1	2	1	1	1	1	1	1	2
Zebrafish	2	2	1	0	0	1	1	1	3	1	1	2	1	2	2	1
Fugu	2	1	1	0	0	1	2	1	3	1	2	2	1	2	1	2
Tilapia	2	1	1	0	0	1	1	1	3	1	1	2	1	2	2	2
Stickleback	2	1	1	0	0	1	1	1	3	1	1	1	1	1	1	2
Medaka	2	1	1	0	0	1	1	1	3	1	2	2	1	2	1	2
Elephant Shark	1	1	0	0	0	1	1	1	2	2	1	1	1	1	2	2
Little Skate	1	0	1	?	0	1	1	1	1	1	1	1	1	1	0	1
Lamprey	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	2

To determine which genes are present or missing within each species, irrespective of location, we searched multiple databases including our own transcriptomes (Table 6.1). It is likely that some of the positive gene numbers in this table are inaccurate, particularly for those species with no sequenced genome or genes with high identity between paralogs. However, the genes marked

as missing have been checked manually and are as accurate as possible. Almost all of the vertebrates analysed have the POU5F3 gene and neighbours; the only exceptions are squamates and eutherian mammals, as previously reported (Frankenberg and Renfree, 2013).

POU5F1 can be identified in mammals, turtles, squamates, axolotl and coelacanth, suggesting it arose in the sarcopterygian ancestor. Indeed the coelacanth gene is located close to a copy of EDF1 and NPDC1, genes which are associated with POU5F3 (Table 6.1; Figure 6.4). This suggests that when the ancestral POU5 gene was duplicated, NPDC1 and EDF1 were also copied. This second NPDC1 gene (NPDC1L) is also found in lungfish, both *Xenopus* species and *Anolis carolinensis*. The duplication of EDF1 is only observable in coelacanth.

A copy of POU5F1 was identified in the sturgeon *Acipenser oxyrinchus* (Johnson et al., 2003a) although we were unable to locate an ortholog in our transcriptome from *Acipenser ruthenus* ovary. There are also a few ESTs in the skate *Leucoraja erinacea* that appear to belong to the POU5F1 gene (Frankenberg and Renfree, 2013). However, these skate ESTs are very short and do not conclusively group with the POU5F1 genes. Using the *Leucoraja erinacea* transcriptome we have isolated only one POU5 sequence, shown along with the sturgeon sequence in Figure 6.6 (bold font) and Figure B.15 (Appendix, page 260).

Figure 6.6 shows the *Acipenser oxyrinchus* sequence grouping with the POU5F1 genes. Although the support for this internal node is not significant (bootstrap < 20; posterior probability < 0.6), it is consistent between the two different tree building methods. Although we cannot truly distinguish between POU5F1 and POU5F3 on this tree considering the low branch supports, the *Acipenser oxyrinchus* sequence is definitely not grouping with the other sturgeon sequences. This would suggest that the *Acipenser oxyrinchus* sequence is most likely POU5F1.

The Bayesian and ML trees both show the skate (*Leucoraja erinacea*) sequence positioned at the base of the tree, in this case with significant support. Considering this position, it suggests that the skate sequence is ancestral to both POU5F1 and POU5F3, contrary to the findings of Frankenberg and Renfree, 2013. However, the presence of multiple skate ESTs, which overlap and yet have different sequences cannot be doubted. Although these skate ESTs contain insufficient

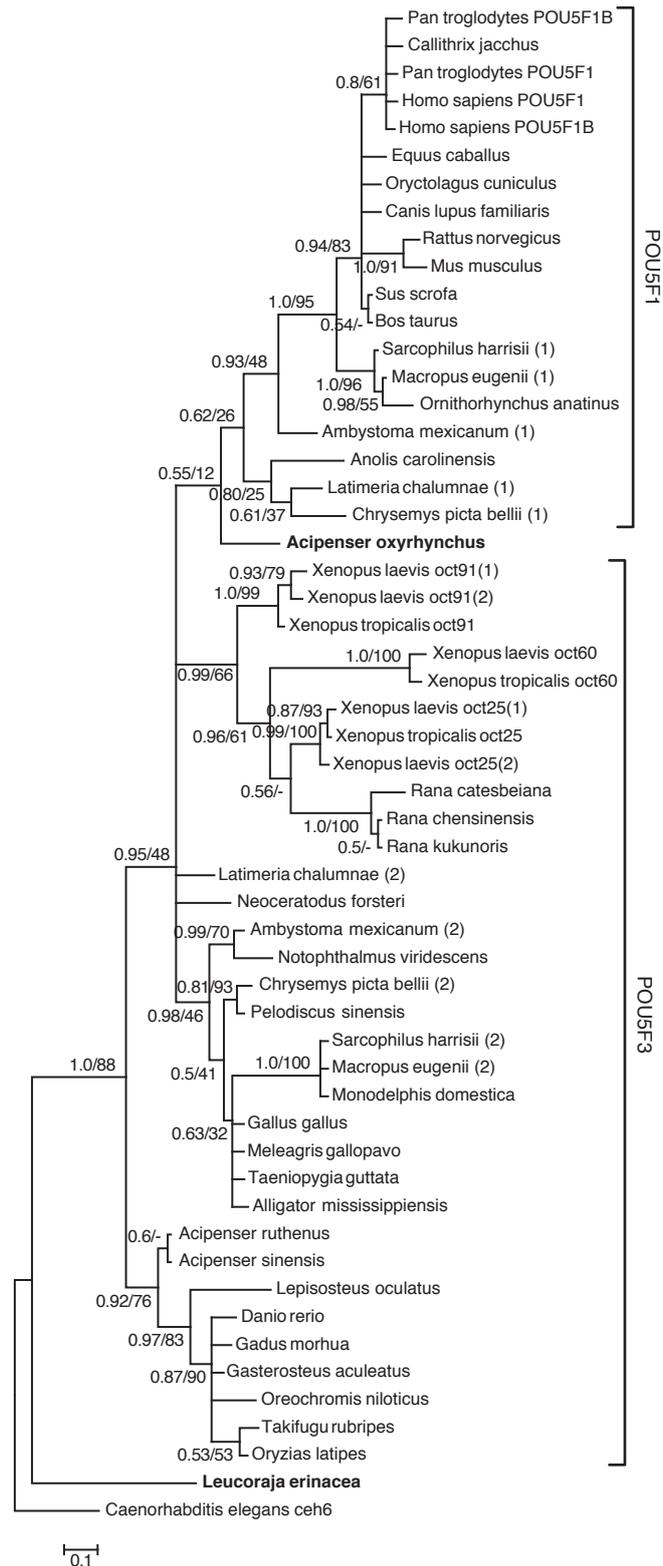


Figure 6.6: Phylogenetic POU5 tree including sturgeon and skate. The Bayesian tree with collapsed branches is shown as in Figure 6.3. The Bayesian and maximum-likelihood trees (Figure B.15, Appendix page 260) were built using the sequences in Figure 6.3 but excluding the POU5F2 gene. To this we added the sturgeon sequence from Johnson et al., 2003a and a contig from the *Leucoraja erinacea* transcriptome, both of which are highlighted in bold font.

information to build a reliable phylogenetic tree, their presence alone suggests the POU5F1/F3 divergence occurred at the base of vertebrates.

Although the syntenic evidence suggests that the POU5F1/POU5F3 duplication occurred at the base of sarcopterygii, the presence of a POU5F1 sequence in sturgeon suggests the duplication could have occurred earlier. However, since this gene cannot be located in our *Acipenser ruthenus* transcriptome nor the gar genome it still requires further investigation. The conclusions that Frankenberg and Renfree (2013) made as to the POU5F1/POU5F3 duplication occurring prior to Chondrichthyes divergence, were made on insufficient and low-quality data for both the POU5 and NPDC1 genes.

Knowing which orthologs belong to either the POU5F1 or POU5F3 genes, we can compare the rates of evolution within each gene. As before we are comparing the relative rate of evolution between taxa undergoing epigenesis and their sister taxa that have acquired preformation. The POU5F1 gene is only present in one species thought to be undergoing preformation, the indian python. The rest of the POU5F1 genes are found in species undergoing epigenesis. It is therefore only possible to perform this analysis on the POU5F3 gene, which occurs in multiple species representing both modes of PGC specification.

The results shown in Figure 6.7 show the difference in branch length for each 3-taxon tree built as part of the relative rate test. The majority of comparisons showed a longer branch length for the taxa using preformation (shown in red), particularly for anurans which had a longer branch length than any species using epigenesis. A large proportion of these comparisons were deemed significant by the relative rate test, in fact the *Rana catesbeiana* POU5F3 gene is evolving at a significantly faster rate than in turtles, crocodile, urodele, lungfish, coelacanth and sturgeon.

The birds showed a very different result, in the majority of comparisons their branch length is shorter than the taxa that has retained epigenesis. However, this difference in rate is only significant when compared against *Ambystoma mexicanum*. Interestingly, the coelacanth and sturgeon POU5F3 genes are evolving slower than all taxa which have acquired preformation, although this difference is not always significant.

The relative rate results for POU5F3 corroborate what we saw in Chapters 3 and 5, taxa that have acquired preformation are tending to evolve at a significantly faster rate than species which have retained epigenesis. This is particularly true for amphibians and actinopterygii, although interestingly the spotted gar *Lepisosteus oculatus* shows no significantly different rate of evolution compared to teleosts.

6.2 Sox2

The SOX gene family was first identified as related to the SRY gene and is characterised by the presence of a HMG (high mobility group) DNA-binding domain (Gubbay et al., 1990). 20 members of the SOX family have been identified, each of which can be classified into one of eight major groups (Bowles et al., 2000; Schepers et al., 2002). Sox2 belongs to the B1 group, along with Sox1, Sox3 and Sox19.

The phylogeny of Sox1, Sox2 and Sox3 shows that these three genes are indeed highly similar to one another, and also that they appear to have derived from gene duplications at the base of vertebrates (Figure B.16, Appendix page 260). Their close identity is evident in the short branch lengths between the three genes, and also by the ML tree showing the Sox3 coelacanth and lungfish genes grouped within the Sox2 clade. The invertebrate deuterostome species (*Saccoglossus kowalevskii*, *Paracentrotus lividus* and *Strongylocentrotus purpuratus*) are all placed at the base of the tree, suggesting they are ancestral to all three vertebrate genes.

To look at the phylogeny of Sox2 in more detail we built a tree consisting of only this gene and a few outgroup species (Figure 6.8 and Figure B.17, Appendix page 261). The DNA alignment was highly conserved, hence the very short branch lengths. The trees do not reflect the species phylogeny, particularly within the teleosts and amphibians. In the Bayesian tree the urodeles group with the majority of teleost fish while the remaining teleost species (*Danio rerio*) is at the base of the tree together with gar, sturgeon and coelacanth. In the ML tree, this latter group is now clustering with the sauropsids while the urodele-teleost clade is positioned at the base of the tree. However, when the poorly supported branches are collapsed (Figure 6.8) the majority of these incongruent branches are removed.

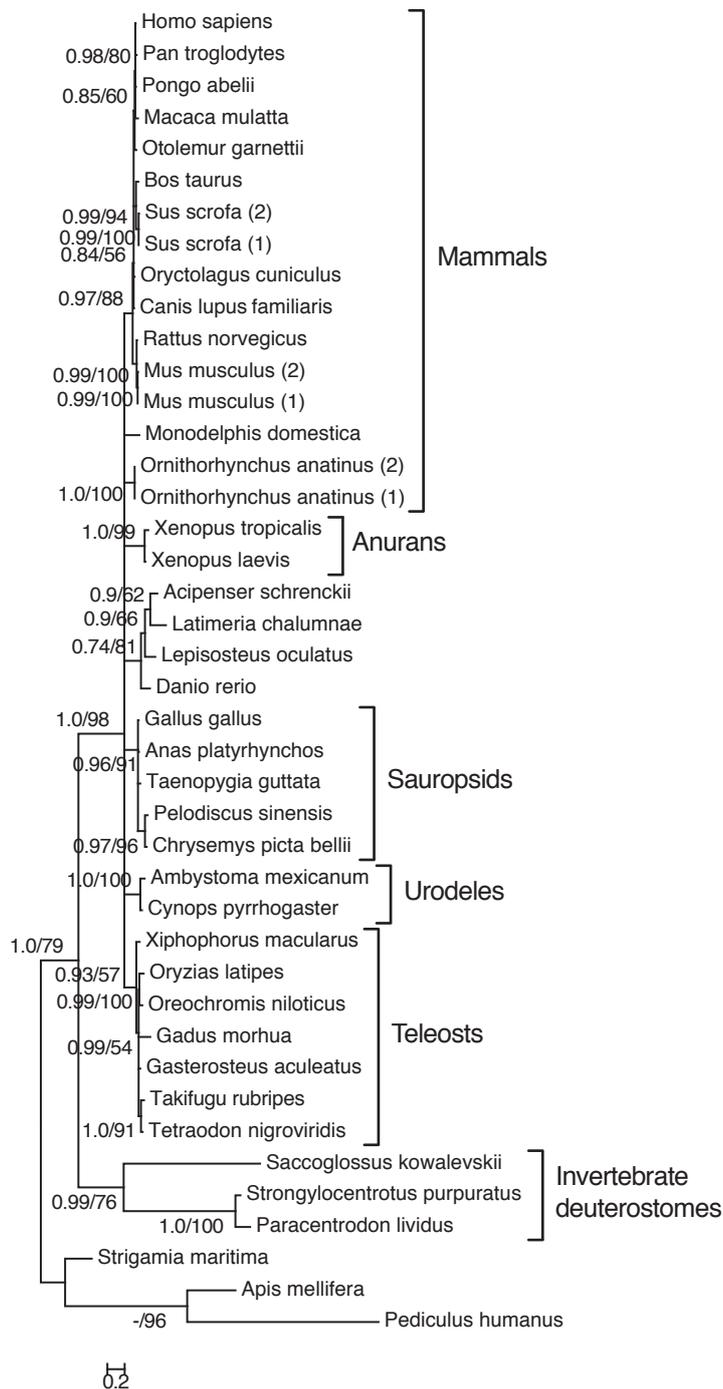


Figure 6.8: Phylogeny of Sox2 DNA sequences. The Bayesian and maximum-likelihood (Figure B.17, Appendix page 261) trees were created using full length sequences from the Sox2 gene, rooted on the protostome species. The ML tree was then collapsed based on the bootstrap values less than 50%, this tree shows both the Bayesian and ML branch supports (PP/BS).

Although the tree does not recapitulate the species phylogeny, there is no clear duplication into multiple genes. We were able to find multiple sequences for mouse, pig and platypus which contained multiple divergent sites. However, in the phylogeny (Figure 6.8) each species groups together suggesting these are independent gene duplication events. To further explore the possibility of gene duplications we analysed the syntenic relationships across vertebrates.

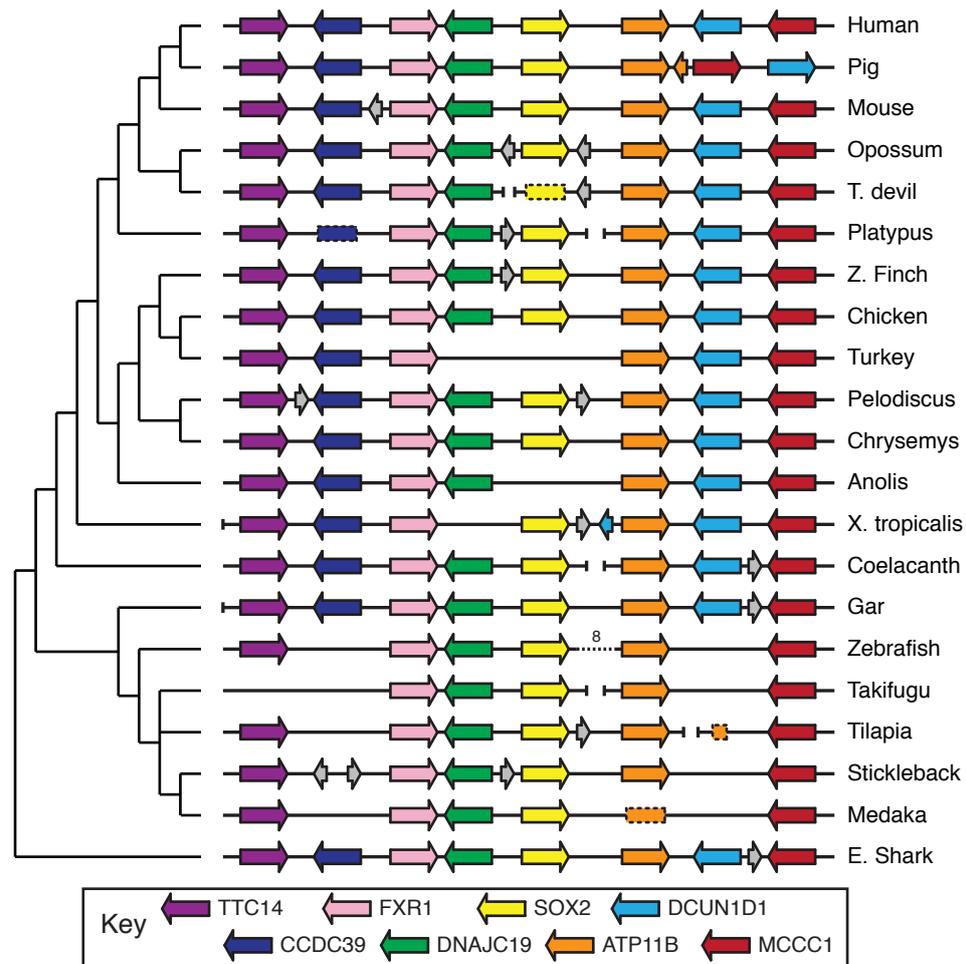


Figure 6.9: Sox2 synteny in vertebrates. The gene neighbourhood surrounding Sox2 is shown for each major vertebrate species.

For the majority of species we only found a single Sox2 gene, all of which were positioned in the same syntenic location (Figure 6.9). This suggests that there is only the one copy of the Sox2 gene in vertebrates. The additional Sox2 gene in Fugu (Table A.50, Appendix page 244) appears on an isolated scaffold, suggesting it is an assembly artefact. The multiple sequences identified in mouse, pig and platypus previously (Figure 6.8), were likely alternative alleles. We were unable to identify a Sox2 sequence in the genomes of the turkey and

anole lizard, or in the *Acipenser ruthenus* transcriptome (Table A.50, Appendix page 244).

The synteny surrounding Sox2 is well conserved in vertebrates, however while many of the genes can be identified in the lamprey genome the synteny is not conserved (not shown). Indeed the Sox2 ortholog is on an unplaced scaffold, and is moreover likely to be the ancestral sequence to Sox1, Sox2 and Sox3 (Figure B.16, Appendix page 260).

Overall our work has shown that there is only one copy of Sox2 in vertebrates, although the species phylogeny cannot be resolved. Sox2 originated at the base of vertebrates, along with Sox1 and Sox3. This differs to what we saw for Oct4; there is a clear relationship with other Sox genes and Sox2 has not been duplicated and alternatively lost in multiple species.

Using a DNA alignment we have tested the relative rate of evolution between all species differing in their mode of PGC specification (Figure 6.10). This shows that there are few comparisons with a significantly different rate of evolution. However, those that do have a significant difference in rate are evolving faster in the species that has acquired preformation. This most commonly occurred when comparing mammals against birds. The only significant differences in rate within teleosts occurred in *Takifugu rubripes* which appears to have a longer branch length than other teleosts.

6.3 Klf4

Klf4 is a member of the Krüppel-like factor family of genes, closely related to the Klf2 and Klf1 genes (Bieker, 2001; Dang et al., 2000). The 15 members of this family all function as transcriptional regulators throughout a diverse range of developmental and cellular processes.

The Klf4 Bayesian and ML phylogenies (Figure B.18, Appendix page 262) show the amphibians grouped together and generally resemble the species phylogeny. However, when the poorly supported ML branches are collapsed (Figure 6.11) it becomes apparent that most branches are not reliable. The few clades which are well supported group the eutherian mammals, archosaurs and testudines, urodeles, anurans and teleosts independently. The relationships between these orders is unclear.

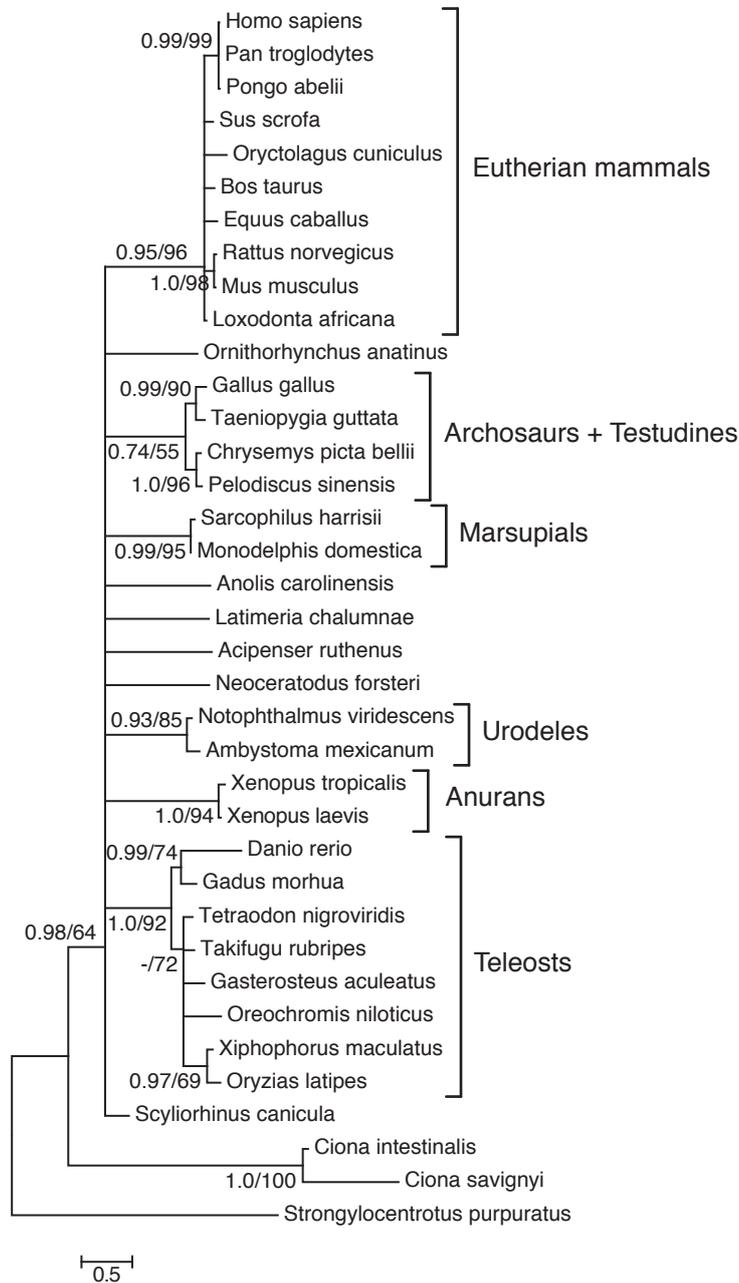


Figure 6.11: Klf4 Phylogeny. The Bayesian and ML (Figure B.18, Appendix page 262) alignments were created using full length DNA sequences. The ML tree is shown with branches supported by less than 50% collapsed, and both the posterior probabilities and bootstrap support values (PP/BS).

To further investigate the relationships among the Klf4 genes, we have analysed the synteny surrounding this gene in vertebrates. Figure 6.12 shows that there is no conserved synteny on one side of the Klf4 gene; the gene arrangement differs in most species. However, on the other side of the Klf4 gene the synteny is well conserved throughout vertebrates. Interestingly, Klf4 could not be identified in the turkey genome but this is likely due to incomplete sequencing considering the strong conservation in other vertebrates.

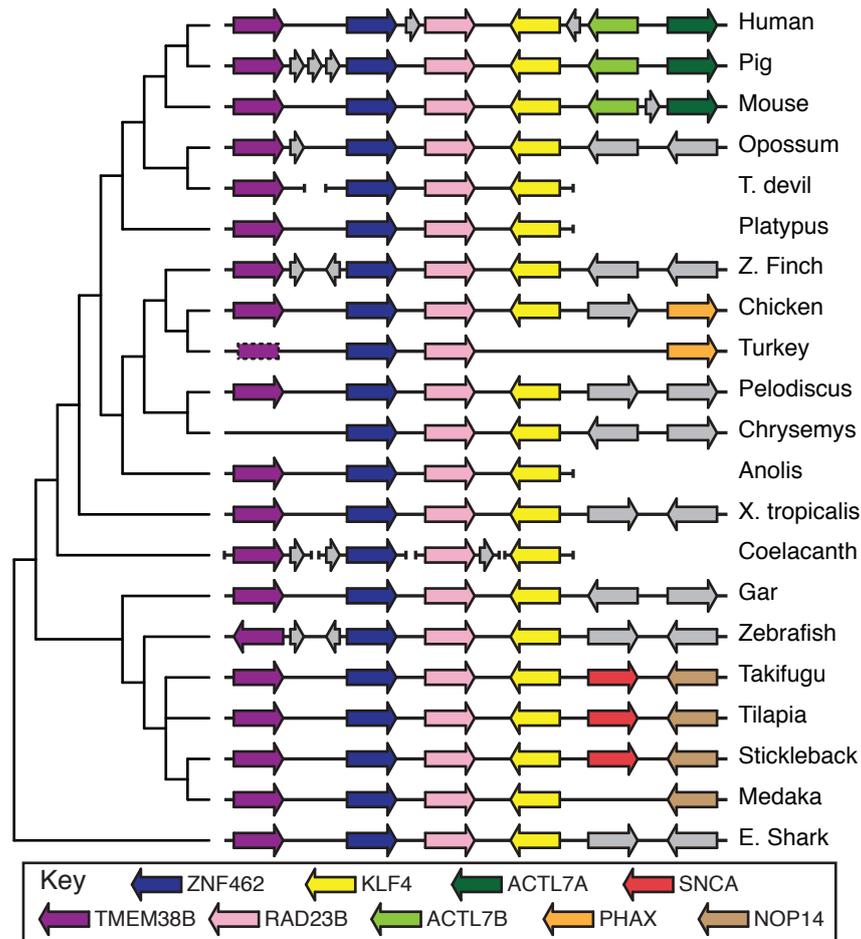


Figure 6.12: Klf4 synteny in vertebrates. The gene neighbourhood surrounding Klf4 is shown for each major vertebrate species.

The Klf4 neighbouring genes could not be located in the lamprey genome (Table A.51, Appendix page 245), suggesting the syntenic locus is missing. However, Klf4 is known to have originated at the base of vertebrates and has a well documented relationship to other Klf genes (Bieker, 2001; Dang et al., 2000).

The relative rate results for the vertebrate DNA sequences from Figure 6.11 are shown in Figure 6.13. A large proportion of the teleost sequences are evolving faster than their counterparts in species which have retained epigenesis, particularly in *Gasterosteus aculeatus*, *Oreochromis niloticus* and *Xiphophorus macularius*. Interestingly, *Gadus morhua* stands out amongst the teleosts as having the shortest branch lengths. There are few significant differences in the rate of evolution outside of teleosts, although *Xenopus laevis* Klf4 appears to be evolving significantly slower than in the turtle species.

6.4 Nanog

The Nanog protein consists of a conserved homeobox domain, surrounded by conserved stretches within the C- and N-terminus (Chambers et al., 2003; Schuff et al., 2012). A recent analysis of the evolutionary origins of Nanog has shown duplications in many species, as well as a loss of synteny between actinopterygians and sarcopterygians (Scerbo et al., 2014). Prior to this paper being released we had also studied the phylogeny and synteny of the Nanog gene.

The Nanog phylogeny we built (Figure 6.14) shows that there are multiple genes within the sauropsids, mammals and urodeles. There are 3 *Ambystoma mexicanum* genes but no nanog sequences could be identified in either *Xenopus* or *Rana*. We were also unable to find any Nanog genes from non-teleostomi species and therefore the trees are rooted on the branch between Actinopterygii and Sarcopterygii. A skate EST has previously been described as having orthology to Nanog (Schuff et al., 2012), however this sequence had no similarity to Nanog when translated (not shown) and so was not included.

The duplications within sauropsids and mammals follow the patterns previously described (Scerbo et al., 2014; Schuff et al., 2012). The additional human sequence is from one of the eleven Nanog pseudogenes present in this species (Booth and Holland, 2004; Fairbanks and Maughan, 2006). Although the relationships between the multiple genes in sauropsids is unclear (Figure 6.14), the remainder of the tree resembles the species phylogeny and many of the backbone branches are well supported.

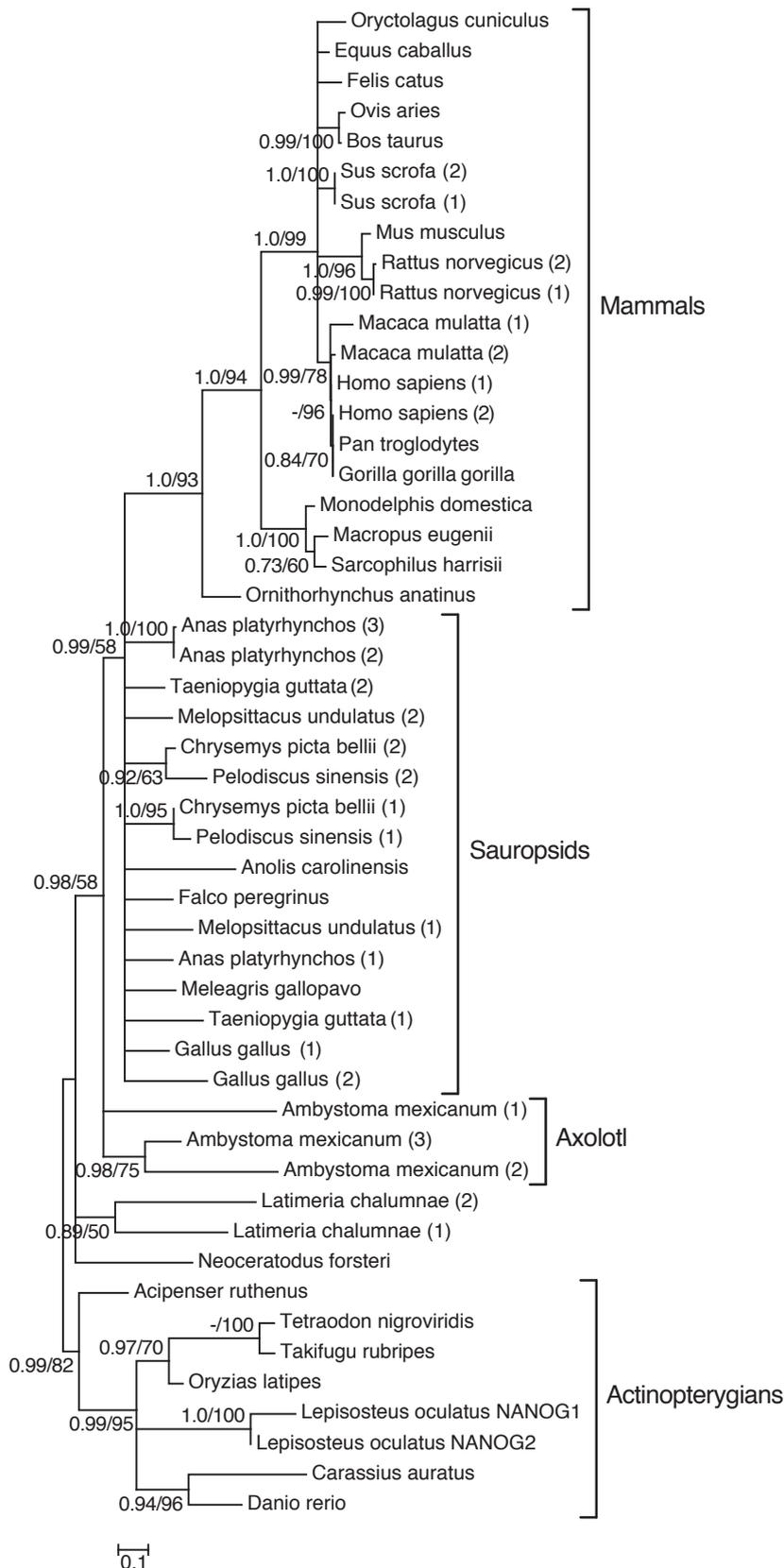


Figure 6.14: Phylogeny of Nanog DNA sequences. The Bayesian and ML (Figure B.19, Appendix page 263) phylogenies were rooted on the Actinopterygian species. The ML tree is shown with branches supported by less than 50% collapsed, the posterior probabilities and bootstrap support values are displayed at each bifurcating node (PP/BS).

To clarify the phylogeny further we have investigated the syntenic relationships between these species. Figure 6.15 shows that there has been a duplication event within sauropsids, presumably at the base of archosaurs and testudines. This supports the phylogeny shown in the maximum-likelihood tree (Figure B.19, Appendix page 263). As has previously been shown Nanog is missing from the *Xenopus* genome, although the surrounding genes are conserved (Hellsten et al., 2010; Schuff et al., 2012).

Figure 6.15 also shows that the actinopterygian Nanog gene is in a completely different location than the sarcopterygian Nanog. Although both gene neighbourhoods are reasonably well conserved across vertebrates, Nanog appears in either one location or the other. This suggests that it is a translocation event that occurred at the base of tetrapods and not a gene duplication event as for POU5F1. We might have expected more than one gene to have been translocated but this does not appear to be the case. Neither syntenic location is conserved in the shark genome and so it is not possible to tell whether there was only one loci or gene ancestrally. Therefore, there is no evidence of a gene duplication event having occurred, although equally this theory cannot be disproved.

Nanog is a member of the homeobox NK-like gene family, which is in turn a member of the ANTP class (Holland and Takahashi, 2005; Wang et al., 2003). The relationships between Nanog and other NK-like homeoboxes have been highly debated, with studies either claiming Bsx1 or Ventx to be the closest homolog (Scerbo et al., 2012; Schuff et al., 2012; Theunissen et al., 2011). Furthermore, it has been suggested that *Xenopus* Ventx has replaced the role of Nanog in this species (Scerbo et al., 2012). We have therefore investigated the relationships between Nanog, Bsx1, Ventx and other NK-like genes (Figure 6.16 and Figure B.20, Appendix page 263). The *Caenorhabditis elegans* ceh6 sequence was used to root the trees, this is a POU3 gene and not a member of the ANTP class of homeoboxes.

The Bayesian and maximum-likelihood phylogenies (Figure 6.16) both show the *Branchiostoma floridae* Vent1 sequence grouping with Nanog. This is the same result shown previously (Scerbo et al., 2012); however Figure 6.16

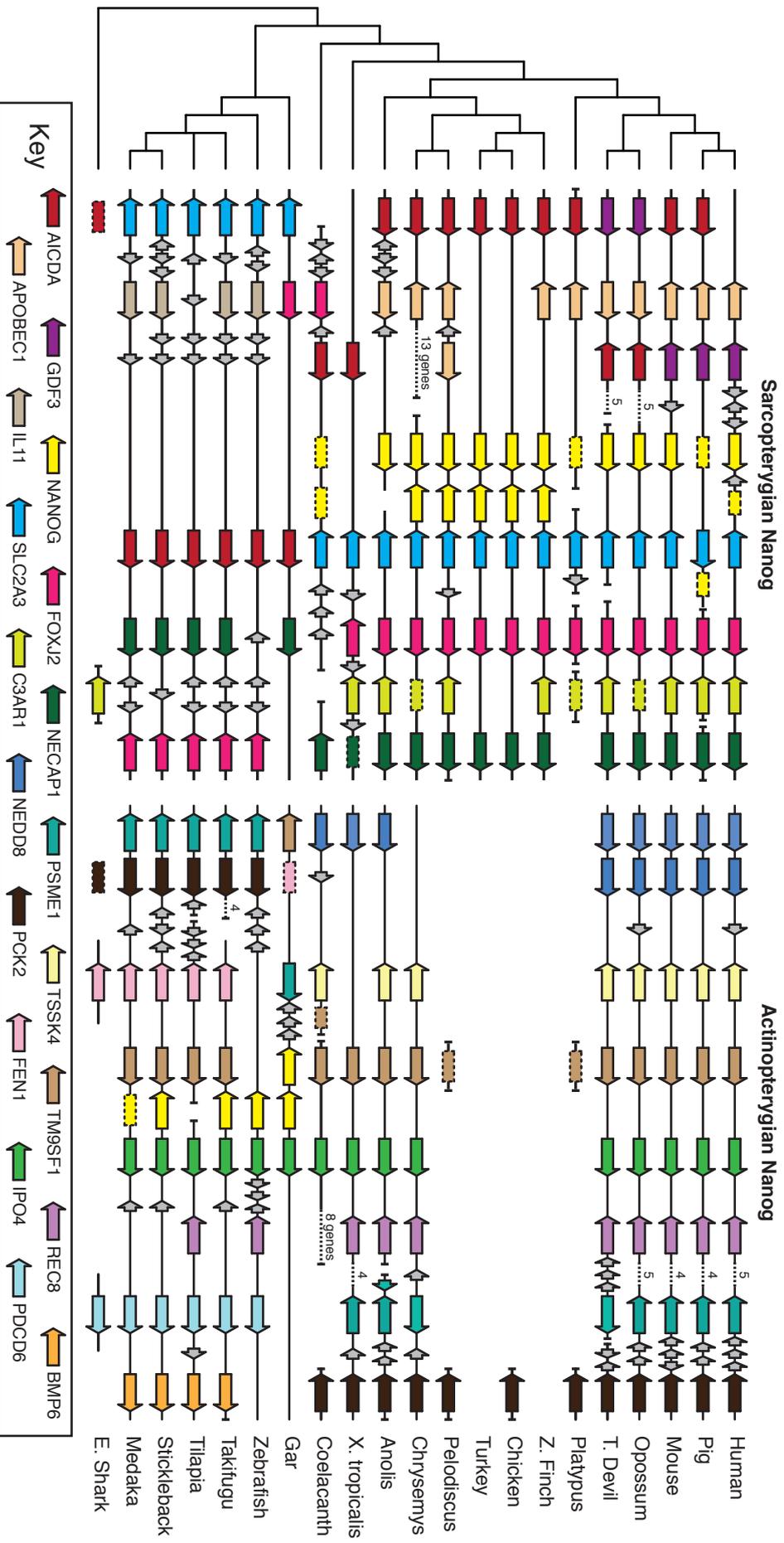


Figure 6.15: Nanog synteny in vertebrates. Two loci are shown, the genes surrounding the tetrapod Nanog and those around the actinopterygian Nanog. The format follows Figure 6.4.

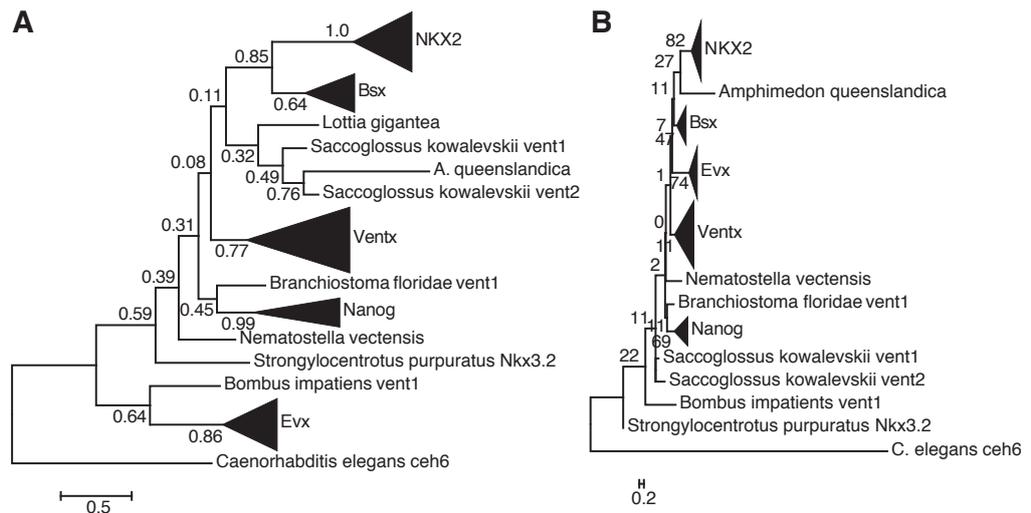


Figure 6.16: Nanog and other NK-like homeoboxes. The Bayesian (A) and ML (B) trees built using a DNA alignment are shown with the major gene clades compressed. The uncompressed trees are shown in Figure B.20 (Appendix, page 263).

shows insignificant support for this relationship. Indeed almost all of the backbone branches separating the genes are poorly supported, suggesting the relationships among these NK-like homeoboxes cannot be resolved. The ancestry of the Nanog gene therefore remains unknown as it does not group with any other NK-like gene with significant support.

To compare the relative rates of evolution for the Nanog gene we used the *Amphimedon queenslandica* NK-like gene to root the sequences from Figure 6.14. We then compared the rates of evolution between each species with differing modes of PGC specification. Figure 6.17 shows that there are fewer comparisons where the longer branch length belongs to the taxa undergoing preformation than observed in previous genes. Indeed, for several mammal and bird comparisons there is a significantly slower rate in the taxon using preformation.

The sturgeon comparisons against teleost show the same result observed in Chapter 3; i.e. there is a significantly faster rate in the teleost sequences than in sturgeon. *Takifugu rubripes* and *Tetraodon nigroviridis* have the longest branch lengths within the teleosts and are evolving significantly faster than two of the turtle sequences, one of the axolotl genes and one of the coelacanth orthologs.

Overall our analyses of the Nanog gene have identified two previously unknown copies of Nanog in the urodele *Ambystoma mexicanum* as well as single copies in the lungfish *Neoceratodus forsteri* and the sturgeon *Acipenser ruthenus*. The phylogenies and observed synteny suggest that Nanog has undergone

		Preformation									
		Taeniopygia guttata (1)	Gallus Gallus (1)	Meleagris gallopavo	Taeniopygia guttata (2)	Gallus gallus (2)	Takifugu rubripes	Tetraodon nigroviridis	Danio rerio	Oryzias latipes	
Epigenesis	Homo sapiens (1)	0.362	0.382	0.482	0.325	0.161	-1.304	-0.323	0.339	0.271	Mammals
	Homo sapiens (2)	0.326	0.358	0.416	0.298	0.158	-1.441	-0.399	0.285	0.244	
	Pan troglodytes	0.326	0.358	0.416	0.298	0.158	-1.441	-0.399	0.285	0.244	
	Sus scrofa (1)	0.171	0.251	0.317	0.164	0.002	-1.369	-0.585	0.072	0.123	
	Sus scrofa (2)	0.171	0.251	0.317	0.164	0.002	-1.369	-0.585	0.072	0.123	
	Mus musculus	0.417	0.378	0.452	0.257	0.080	-2.483	-0.895	0.225	0.150	
	Rattus norvegicus (1)	0.568	0.470	0.520	0.453	0.180	-2.574	-1.696	0.292	0.258	
	Rattus norvegicus (2)	0.529	0.465	0.515	0.464	0.156	-2.268	-1.574	0.289	0.253	
	Monodelphis domestica	0.524	0.582	0.639	0.354	0.149	-0.151	-2.274	0.425	0.638	
	Macropus eugenii	0.673	0.635	0.666	0.716	0.220	-0.282	-2.148	0.567	1.262	
	Ornithorhynchus anatinus	0.232	0.257	0.361	0.126	0.792	-1.257	-0.603	0.208	0.314	Turtles
	Anolis carolinensis	0.080	0.253	0.222	0.109	-0.003	-0.329	-0.343	0.242	0.130	
	Pelodiscus sinensis (1)	-0.184	-0.085	0.033	-0.054	-0.287	-0.599	-0.522	-0.162	-0.086	
	Chrysemys picta bellii (1)	-0.092	0.154	0.326	-0.023	-0.415	-0.574	-0.546	-0.138	-0.057	
	Pelodiscus sinensis (2)	0.303	0.068	0.181	0.124	-0.044	-0.708	-0.447	0.144	0.084	Urodèles
	Chrysemys picta bellii (2)	0.253	0.134	0.338	0.185	-0.210	-0.719	-0.804	0.067	0.080	
	Ambystoma mexicanum (1)	0.643	-0.152	-0.066	-0.187	-0.305	-0.668	-0.625	-0.215	-0.143	
	Ambystoma mexicanum (2)	-0.270	-0.048	0.051	-0.047	-0.145	-1.626	-0.795	0.511	0.019	
	Ambystoma mexicanum (3)	0.163	0.139	0.263	0.211	0.017	-0.854	-0.517	0.395	0.236	
	Neoceratodus forsteri	-0.266	0.123	0.287	0.188	-0.469	-0.796	-0.288	-0.001	0.039	
Latimeria chalumnae (1)	1.130	0.840	0.869	0.778	1.679	-0.241	-1.134	0.554	1.053		
Latimeria chalumnae (2)	-1.498	-1.271	-0.265	-1.350	-0.800	-1.402	-1.619	-0.208	-0.340		
Acipenser ruthenus	-0.346	-0.231	-0.130	-0.324	-0.439	-0.578	-0.523	-0.502	-0.289		
Lepisosteus oculatus 1	0.944	0.581	0.750	0.874	0.220	0.035	0.161	0.368	0.597		
Lepisosteus oculatus 2	0.562	0.399	0.981	0.720	0.059	-0.773	-0.701	0.196	0.489		
		Birds					Teleosts				

Figure 6.17: Relative rate test results for Nanog. The figure follows the same format as Figure 6.13. Each RRT was carried out using the *Amphimedon queenslandica* sequence as the reference.

many tandem duplications such as at the base of the archosaur and testudine lineage and within coelacanth. This corresponds with the analyses undertaken by Scerbo et al. (2014). As previously observed Nanog is missing from the *Xenopus* genome (Hellsten et al., 2010; Schuff et al., 2012), however we have also been unable to identify any orthologs in the *Rana* transcriptomes. We have been unable to identify the ancestry of the Nanog gene, no orthologs have been located basal to vertebrates and the gene does not have a clear relationship to other NK-like homeoboxes.

6.5 Conclusion

We investigated whether in-depth analysis of a few genes would demonstrate the same patterns of incongruence and changes in rate as the global analyses carried out in Chapters 3 and 5. To do this we looked at four pluripotency genes, Oct4, Sox2, Klf4 and Nanog. Each of these genes are involved in early development as well as the production and maintenance of the PGCs. Oct4 and Nanog are both specific to pluripotency while Sox2 and Klf4 have more generalized functions.

We first showed that the mammalian Oct4 gene, otherwise known as POU5F1, has been lost in many species, particularly those that have acquired preformation. The gene has been lost in birds, anurans and teleosts but retained in turtles, urodeles and sturgeon. Nanog also shows patterns of gene loss as it has been deleted in anurans and potentially teleosts. However, in the latter this may be a gene translocation rather than a duplication and subsequent loss since no species have copies at both loci. Sox2 and Klf4 show a more conserved synteny with no particular gene duplication or loss; however neither gene could be located in the turkey genome.

Analysing the relative rate of evolution between taxa undergoing epigenesis and their orthologs in species using preformation has shown many of the same patterns as in Chapter 3. In all four genes the sturgeon species is evolving significantly slower than at least one teleost. Urodeles have been observed evolving significantly slower than Anurans in the POU5F3 gene although not in Klf4 or Sox2. Interestingly, we have not observed any turtle sequences evolving significantly slower than their bird orthologs, indeed Klf4 shows the opposite result.

We have shown that patterns of gene loss appear related to the mode of PGC specification in POU5F1 and Nanog. The two genes that are specific to pluripotency have been lost in species that have acquired preformation. This suggests an extreme change of selection has occurred in these pluripotency-specific genes which relates to the mode of PGC specification.

Discussion

The two modes of PGC specification, epigenesis and preformation, differ in how the somatic and germ cells are segregated during embryo development. Preformation occurs through maternally deposited germ plasm while epigenesis occurs later in development through embryonic signalling (Ikenishi, 1998; Ohinata et al., 2009). Preformation is the derived mechanism and has evolved independently in birds, anurans and teleost fish. These taxa are associated with derived gene regulatory networks and an increase in speciation events (Crother et al., 2007; Swiers et al., 2010). Based on this information, it had been proposed that epigenesis enforces a developmental constraint on the somatic cells which was released in species that acquired preformation (Johnson et al., 2003b, 2011).

To investigate this hypothesis we studied protein-coding genes across vertebrates and whether phylogenetic incongruence or changes in the rate of evolution correlated with the mode of PGC specification. We began this analysis by taking a global view across all available vertebrate sequences, assisted by the three novel transcriptomes we had sequenced. We built 4-taxon phylogenetic trees, analysed the distance matrices, investigated the relative rate of evolution and identified gene function (Chapters 3, 4 and 5; Evans et al., 2014). We also analysed phylogenetic trees and relative rates of evolution for specific genes involved in pluripotency (Chapter 6). For these genes we also identified gene duplication and loss events using the available synteny information.

7.1 Phylogenetic Incongruence

We had observed that many previously published gene trees featuring a urodele, anuran and mammal were incongruent (Table A.1, Appendix page 221). Many of these incongruent trees grouped the two species undergoing epigenesis, urodele and mammal, together. This included the trees for *Dazl* and *Vasa*, both of which are involved in PGC specification and so it had

been suggested that the tree was reflecting a mechanistic relationship (Johnson et al., 2003a). Our first aim was therefore to assess the extent of phylogenetic incongruence. We then investigated whether these trees correlated with the mode of PGC specification.

We began this process in Chapter 3, building trees within amphibians, actinopterygians and sauropsids. We were able to build a large number of phylogenies with significant support, either by bootstrapping or the SH-test. Within amphibians and actinopterygians we saw a large proportion of incongruent topologies, the majority of which grouped either urodeles or sturgeon with mammals. The sauropsids however showed very few incongruent trees and there was no consistent bias within them. To further investigate this result we increased the size of each dataset (Chapter 5), adding complete genomes and transcriptomes. However, this continued to show very few incongruent trees within sauropsids, suggesting this was not an artefact due to low sequence numbers. One possible reason for this result is the difference in time since divergence; birds and crocodiles were thought to have diverged 219 Mya, while anurans and urodeles diverged 264 Mya, and sturgeon and teleosts diverged 312 Mya (Hedges et al., 2006). It could therefore be that there has been insufficient time since preformation evolved in birds for the sequences to differ to the extent that the tree is incongruent.

As well as increasing the size of the dataset in Chapter 5, we also included new species to our analysis. This showed that gar has a similar proportion of incongruent trees as sturgeon, and that there is a strong bias towards the Mammal-Gar topology. This suggests that gars are also undergoing epigenesis; although this requires experimental verification. We also showed a bias within the incongruent trees of coelacanth, lungfish and sharks; all of which tended to group the taxa undergoing epigenesis together. This was particularly extreme in trees featuring sharks, teleosts and mammals, wherein the majority of phylogenies showed a Mammal-Shark topology contrary to the species phylogeny.

Phylogenetic incongruence was wide reaching as when we mapped the results to mouse (Chapter 4 and Section 5.3) over 50% of the mouse genes had an incongruent phylogeny within at least one vertebrate ortholog. This demonstrates the breadth of incongruence in vertebrates, and concurs with other accounts of phylogenetic incongruence (Rokas et al., 2003b). We also discovered

that these genes were generally expressed early in development, which led us to look at the phylogeny of pluripotency factors.

Chapter 6 showed the phylogenetic trees built for Oct4, Sox2, Klf4 and Nanog were all poorly supported. None of these trees were conclusively able to show the species phylogeny or an incongruent topology. The majority of these poorly supported trees were due to short alignments over a well conserved protein domain. We were therefore unable to determine the extent of phylogenetic incongruence in pluripotency factors.

In Section 3.5 we identified the cause behind the four-taxon incongruent phylogenies, long branch attraction (LBA; Anderson and Swofford, 2004; Felsenstein, 1978; Sanderson et al., 2000). That is, when one of the sister taxa in the amphibian trees was evolving at a significantly faster rate, it was highly likely to be positioned at the base of the tree with the outgroup. This positioning is characteristic of long branch attraction (Philippe and Laurent, 1998). Furthermore we showed that if the outgroup was replaced by a sequence evolving significantly slower then the proportion of incongruent trees dropped. This suggested that the incongruent trees were largely an artefact due to differences in the rate of evolution.

7.2 The rate of evolution and PGC specification

We used the relative rate test to investigate whether sister taxa with differing modes of PGC specification were evolving at significantly different rates. This showed that a large number of genes in anurans, teleosts, snakes and birds were evolving significantly faster than in their orthologs from species undergoing epigenesis (Chapters 3 and 5). This included orthologs of Oct4, Sox2, Klf4 and Nanog (Chapter 6). However, there were some discrepancies within lepidosaurs (Section 5.2), specifically between *Anolis carolinensis* and skinks, neither of which have had the mode of PGC specification experimentally verified. The relative rate test results for coelacanth and shark also showed a generally faster rate in the species which has retained epigenesis. In this case we weren't comparing directly between sister taxa which differ in the mode of PGC specification and so this observation may be unrelated to our previous analyses.

However, the overriding observation in amphibians, actinopterygians, archosaurs and testudines was for a significantly faster rate in the species that

has acquired preformation, independent of whole genome duplications. The extent of this result was demonstrated in Section 5.3 when over 9,000 mouse genes had at least one ortholog with a significantly faster rate in the taxon that has acquired preformation. There is therefore a strong association within vertebrates between the mode of PGC specification and the relative rate of molecular evolution.

Nevertheless, we have considered whether other factors that differed between the sister taxa could explain the result. Figure 7.1 shows the relative rate test results from Section 5.2, including lepidosaurs, plotted against various morphological characteristics which have previously been linked to changes in the rate of molecular evolution (Goetting-Minesky and Makova, 2006; Nabholz et al., 2008; Thomas et al., 2010).

Table 7.1: Student’s T-test results. For each variable the epigenesis and preformation means were compared using a two-tailed Student’s T-test with unequal variances. This used the same data as shown in Figure 7.1. The significant results ($p < 0.05$) are highlighted in bold.

Variable	Mean		t-value	d.f.	p-value
	Epi.	Pre.			
Generation time (days)	2551.33	962.31	1.911	8.426	0.0906
Maximum longevity (years)	35.81	21.12	1.741	13.932	0.1037
Genome size (c-value)	11.93	1.51	2.518	11.032	0.0285
Genome size exc. Uro.	2.57	1.51	2.660	9.784	0.0243
Significantly faster rate (%)	21.93	94.76	-15.444	14.701	1.722E⁻¹⁰

Figure 7.1A and B show a weak positive correlation between generation time, maximum longevity and the rate of molecular evolution. However, there is no significant difference between the generation time and longevity values for the taxa undergoing epigenesis compared to those that have acquired preformation (Table 7.1). These factors are therefore not able to explain the bias within the relative rate test results.

We also analysed the correlation between genome size and the rate of evolution, since it is well known that the urodele and anuran genome sizes differ (Vinogradov, 1998). Indeed the urodeles clearly differentiate from all other taxa as can be seen in Figure 7.1C and D. Once again there is a weak positive correlation against the rate of evolution, even when the urodeles are excluded. Interestingly, the mean genome size for the species with epigenesis significantly differs to the average genome size in species that have acquired preformation

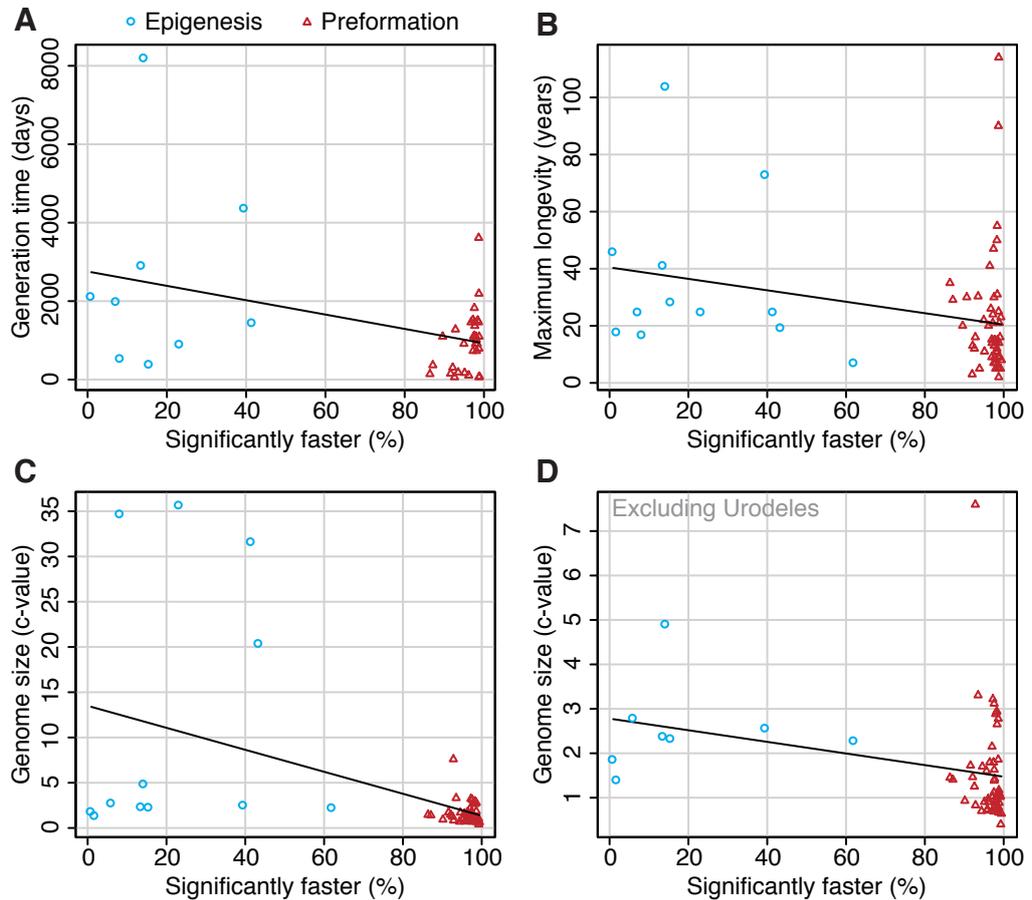


Figure 7.1: Rate of evolution correlations. Each panel shows the proportion of significant sequences evolving at a faster rate (from Section 5.2) plotted against the generation time (A), maximum longevity (B) and genome size (C and D) for each vertebrate species with the relevant information. (D) excludes the urodeles, thereby allowing a closer look at the remaining species. In each plot the line of best fit shows a weak correlation, however the epigenesis (blue) and preformation (red) species never overlap on the horizontal axis. The data was collated from AnAge (Tacutu et al., 2013) and the Genome Size Database (Gregory, 2014).

(Table 7.1). This may be worth future investigation to explore how a change in genome size, which is mostly caused by chromosomal changes or transposable elements (Petrov, 2001), relates to changes in the rate of evolution of protein-coding genes. It may be that a smaller genome allows the cell to divide faster, potentially increasing the number of cell divisions.

Although there is a slight correlation between the mode of PGC specification and genome size, the strongest correlation by far is with the proportion of sequences evolving at a significantly faster rate (Table 7.1). This suggests that the changes in rate we have observed are directly associated with the mode of PGC specification, and not another difference between sister taxa. Indeed, this was also suggested when we compared the relative rate of evolution between species which share the same mode of PGC specification (Section 5.2.1). However, the mechanism behind this association remains unknown.

The hypothesis of constraint and constraint release (Section 1.4.1; Johnson et al., 2003b, 2011), fits with the observed pattern in relative rate differences. Genes in species which have acquired preformation, and as such would have undergone a constraint release, are evolving significantly faster than their orthologs in species which have retained epigenesis. We also observed that these genes were typically expressed prior to PGC specification in mouse, but were over-represented throughout embryo development in zebrafish (Chapter 4). Considering the mode of PGC specification in these species, this also fits with the hypothesis of constraint and constraint release. Genes expressed before PGC specification in mouse are under a constraint which has been released in taxa that have acquired preformation. We also observed that genes which are specific to pluripotency (Oct4 and Nanog) have been lost in species which have acquired preformation (Chapter 6). This suggests an extreme loss of selection, or constraint in these genes. Therefore our data support the hypothesis that a release of constraint has occurred in taxa that have acquired preformation. However, there is still no direct evidence that a release in developmental constraint is the cause for the altered rate of protein-coding gene evolution.

One alternative interpretation of our results is that the mode of PGC specification is affecting the number of germ cell divisions, and therefore altering the rate of evolution through a well recognised process (Goetting-Minesky and Makova, 2006; Thomas et al., 2010). For this hypothesis to explain our data, we

would predict that the number of germ cell divisions in species using preformation to be greater than in species undergoing epigenesis.

A proxy for this kind of analysis is to observe the differences in litter/clutch sizes. Urodeles vary from 37 in *Ambystoma tigrinum* through to 725 in *Pleurodeles waltl* (Tacutu et al., 2013), while *Xenopus laevis* has been reported to have approximately 4000 eggs released per clutch (Du Preez et al., 2008). This would suggest that there is a difference in the number of germ cell divisions between anuran and urodele amphibians. However, the 13 egg clutch size for the turtle *Trachemys scripta*, does not particularly differ to the clutches of 11 and 15 eggs in the birds *Meleagris gallopavo* and *Colinus virginianus* (Tacutu et al., 2013). This suggests that a change in the number of germ cell divisions, associated with the mode of PGC specification, may not explain all of the results observed. However, clutch size is merely a proxy for the number of germ cell divisions, and so further work would be required to examine this hypothesis in more detail across all vertebrates.

7.3 Future work

The obvious progression from this project is to deduce the mechanism between the rate of evolution and the mode of PGC specification. As previously suggested this could be constraint and constraint release (Johnson et al., 2003b, 2011), a change in germ cell divisions or another unknown mechanism. To investigate these hypotheses would require a sizeable undertaking, and yet it would still be impossible to truly identify causation since evolution cannot be observed.

It may be possible to measure constraint using loss-of-function experiments (Roux and Robinson-Rechavi, 2008). Genes with severe phenotypes can be inferred as being under higher constraint than genes with milder phenotypes. Comparing these global analyses between species using preformation and epigenesis may identify the same genes as those with significant differences in the rate of evolution. The rate of non-synonymous and synonymous substitutions could also be examined (Castillo-Davis and Hartl, 2002; Davis et al., 2005; Roux and Robinson-Rechavi, 2008). Genes under constraint may have a lower proportion of non-synonymous mutations. However, to calculate these values would require the inclusion of the third codon position, which we show has a

high variance in GC bias. To observe the number of germ cell divisions that occur in vertebrates may require fate mapping, as has been done in *Caenorhabditis elegans* (Sulston et al., 1983). If, at an equivalent stage, a single PGC could be permanently dyed, this might provide a method of comparing the number of germ cell divisions between organisms.

One aspect that became clear during this project is that there is a lack of information and sequences for non-model vertebrates. In many species there is little to no information on the mode of PGC specification, for example in lepidosaurs, gar and crocodiles. In these cases a relatively simple experiment would be to analyse whether their orthologs to *Vasa* and *Dazl* localise in the oocyte. As demonstrated by the conflicting results in sturgeon, it would be vital for these experiments to observe the localisation of the endogenous RNA (Johnson et al., 2011; Saito et al., 2014).

In Chapter 5 we identified a difference in the relative rate test results between skinks and snakes, both of which are supposedly undergoing preformation (Hubert, 1985). To further explore this result we would require a more comprehensive dataset of sequences, and the two recently published snake genomes (Castoe et al., 2013; Vonk et al., 2013) could go towards this. The differences in rate results between sturgeon and gar in Chapter 6 suggests that increasing the number of sequences for non-teleost actinopterygian fish would also be worthwhile. Furthermore, there are still many groups of vertebrates we have been unable to investigate due to insufficient information being available. One key group worthy of future work would be the caecilians, the third amphibian order.

Analysing the molecular evolution of *Oct4*, *Sox2*, *Klf4* and *Nanog* in Chapter 6 identified a correlation between gene loss and the mode of PGC specification. This warrants further investigation, both in detail such as in the *Vent* gene, as well as on a global level. By identifying all orthologs, not just those with a one-to-one relationship, we may be able to deduce the scale of gene loss/duplication and whether this correlates with the mode of PGC specification. However, for this type of analysis to work we would require a complete genome for all species.

This project has highlighted the disparity in the volume of sequences between species using epigenesis and those which have acquired preformation.

All vertebrate non-mammal model species undergo preformation and there have been few sequencing projects in species which do not have germ plasm. This is particularly important since in Section 3.5 we showed that by changing the outgroup to a species using epigenesis we were able to recover the species phylogeny. This suggests that sequencing the genomes of axolotl and sturgeon would prove highly beneficial to the wider scientific community.

Bibliography

- Abdullayev, I., Kirkham, M., Björklund, Å. K., Simon, A., and Sandberg, R. (2013). A reference transcriptome and inferred proteome for the salamander *Notophthalmus viridescens*. *Experimental Cell Research*, 319(8):1187–1197.
- Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., et al. (2000). The new animal phylogeny: Reliability and implications. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9):4453–4456.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.
- Aksoy, I., Jauch, R., Chen, J., Dyla, M., Divakar, U., et al. (2013). Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm. *The EMBO Journal*, 32:938–953.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.
- Amemiya, C. T., Alföldi, J., Lee, A. P., Fan, S., Philippe, H., et al. (2013). The african coelacanth genome provides insights into tetrapod evolution. *Nature*, 496:311–316.
- Anderson, F. E. and Swofford, D. L. (2004). Should we be worried about long-branch attraction in real data sets? investigations using metazoan 18s rdna. *Molecular Phylogenetics and Evolution*, 33(2):440–451.
- Anderson, J. S. (2008). Focal review: The origin(s) of modern amphibians. *Evolutionary Biology*, 35(4):231–247.
- Anderson, J. S., Reisz, R. R., Scott, D., Fröbisch, N. B., and Sumida, S. S. (2008). A stem batrachian from the early permian of texas and the origin of frogs and salamanders. *Nature*, 453(7194):515–518.
- Arthur, W. (2001). Developmental drive: an important determinant of the direction of phenotypic evolution. *Evolution and Development*, 3(4):271–278.
- Arthur, W. (2011). *Evolution: a developmental approach*. Wiley-Blackwell.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.
- Avilion, A. A., Nicolis, S. K., Pevny, L. H., Perez, L., Vivian, N., et al. (2003). Multipotent cell lineages in early mouse development depend on Sox2 function. *Genes & Development*, 17:126–140.
- Bachvarova, R. F., Crother, B. I., and Johnson, A. D. (2009). Evolution of germ cell development in tetrapods: comparison of urodeles and amniotes. *Evolution & Development*, 11(5):603–609.
- Bachvarova, R. F., Masi, T., Drum, M., Parker, N., Mason, K., et al. (2004). Gene expression in the axolotl germ line: *Axdazl*, *Axvh*, *Axoct-4*, and *Axkit*. *Developmental Dynamics*, 231(4):871–880.
- Bascove, M. and Frippiat, J. (2010). Molecular characterization of pleurodeles waltl activation-induced cytidine deaminase. *Molecular Immunology*, 47(7-8):1640–1649.
- Bergmann, P. J. and Irschick, D. J. (2011). Vertebral evolution and the diversification of squamate reptiles. *Evolution*, 66(4):1044–1058.
- Bieker, J. J. (2001). Krüppel-like factors: Three fingers in many pies. *Journal of Biological Chemistry*, 276:34355–34358.
- Billett, F. S. and Adam, E. (1976). The structure of the mitochondrial cloud of *Xenopus laevis* oocytes. *Journal of Embryology and Experimental Morphology*, 33(6):697–710.

- Blair, J. E. and Hedges, S. B. (2005). Molecular phylogeny and divergence times of deuterostome animals. *Molecular Biology and Evolution*, 22(11):2275–2284.
- Bodner, M., Castrillo, J., Theill, L. E., Deerinck, T., Ellisman, M., et al. (1988). The pituitary-specific transcription factor GHF-1 is a homeobox-containing protein. *Cell*, 55(3):505–518.
- Booth, H. F. and Holland, P. W. (2004). Eleven daughters of nanog. *Genomics*, 84(2):229–238.
- Boswell, T., Dunn, I. C., Wilson, P. W., Joseph, N., Burt, D. W., et al. (2006). Identification of a non-mammalian leptin-like gene: Characterization and expression in the tiger salamander (*Ambystoma tigrinum*). *General and Comparative Endocrinology*, 146(2):157–166.
- Boterenbrood, E. and Nieuwkoop, P. (1973). The formation of mesoderm in urodelean amphibians. 5. its regional induction by endodermal. *Wilhelm Roux' Archiv Für Entwicklungsmechanik Der Organismen*, 173(4):319–332.
- Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., et al. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum xenoturbellida. *Nature*, 444:85–88.
- Bourlat, S. J., Nielsen, C., Lockyer, A. E., Littlewood, D. T. J., and Telford, M. J. (2003). Xenoturbella is a deuterostome that eats molluscs. *Nature*, 424:925–928.
- Bowes, J. B., Snyder, K. A., Segerdell, E., Gibb, R., Jarabek, C., et al. (2008). Xenbase: a xenopus biology and genomics resource. *Nucleic Acids Research*, 36(Database issue):D761–D767.
- Bowles, J., Schepers, G., and Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Developmental Biology*, 227(2):239–255.
- Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956.
- Braat, A. K., Zandbergen, T., Water, S. V. D., Goos, H. J. T., and Zivkovic, D. (1999). Characterization of zebrafish primordial germ cells: Morphology and early distribution of Vasa RNA. *Developmental Dynamics*, 216(2):153–167.
- Bradford, C. S., Walthers, E. A., Searcy, B. T., and Moore, F. L. (2005). Cloning, heterologous expression and pharmacological characterization of a kappa opioid receptor from the brain of the rough-skinned newt, *taricha granulosa*. *Journal of Molecular Endocrinology*, 34:809–823.
- Bradford, C. S., Walthers, E. A., Stanley, D. J., Baugh, M. M., and Moore, F. L. (2006). Delta and mu opioid receptors from the brain of a urodele amphibian, the rough-skinned newt *Taricha granulosa*: Cloning, heterologous expression, and pharmacological characterization. *General and Comparative Endocrinology*, 146(3):275–290.
- Britten, R. J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science*, 231(4744):1393–1398.
- Bromham, L. and Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics*, 4:216–224.
- Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., et al. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, 23(9):1808–1816.
- Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42(3):384–397.
- Burgess, S., Reim, G., Chen, W., Hopkins, N., and Brand, M. (2002). The zebrafish *spiel-ohne-grenzen* (spg) gene encodes the POU domain protein Pou2 related to mammalian Oct4 and is essential for formation of the midbrain and hindbrain, and for pre-gastrula morphogenesis. *Development*, 129:905–916.

- Cadinouche, M., Liversage, R., Muller, W., and Tsilfidis, C. (1999). Molecular cloning of the *Notophthalmus viridescens* radical fringe cDNA and characterization of its expression during forelimb development and adult forelimb regeneration. *Developmental Dynamics*, 214(3):259–268.
- Cameron, C. B., Garey, J. R., and Swalla, B. J. (2000). Evolution of the chordate body plan: New insights from phylogenetic analyses of deuterostome phyla. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9):4469–4474.
- Cao, Y., Siegel, D., and Knöchel, W. (2006). *Xenopus* POU factors of subclass V inhibit activin/nodal signaling during gastrulation. *Mechanisms of Development*, 123(8):614–625.
- Castillo-Davis, C. I. and Hartl, D. L. (2002). Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Molecular Biology and Evolution*, 19(5):728–735.
- Castoe, T. A., de Koning, A. P. J., Hall, K. T., Card, D. C., Schield, D. R., et al. (2013). The burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences of the United States of America*, 110(51):20645–20650.
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552.
- Castrillon, D. H., Quade, B. J., Wang, T. Y., Quigley, C., and Crum, C. P. (2000). The human VASA gene is specifically expressed in the germ cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 97(17):9585–9590.
- Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., et al. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, 113(5):643–655.
- Chang, J. T. (1996). Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, 134(2):189–215.
- Chatfield, J., O'Reilly, M.-A., Bachvarova, R. F., Ferjentsik, Z., Redwood, C., et al. (2014). Stochastic specification of primordial germ cells from mesoderm precursors in axolotl embryos. *Development*, 141:2429–2440.
- Chen, F., Mackey, A. J., Vermunt, J. K., and Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4).
- Chiquoine, A. D. (1954). The identification, origin, and migration of the primordial germ cells in the mouse embryo. *The Anatomical Record*, 118(2):135–146.
- Chiu, C., Dewar, K., Wagner, G. P., Takahashi, K., Ruddle, F., et al. (2004). Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Research*, 14(1):11–17.
- Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.
- Cowen, R., editor (2013). *History of Life*. Wiley-Blackwell, 5th edition.
- Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., et al. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, 8(5):783–786.
- Crews, L., Gates, P. B., Brown, R., Joliot, A., Foley, C., et al. (1995). Expression and activity of the newt *Msx-1* gene in relation to limb regeneration. *Proceedings of the Royal Society B*, 259(1355):161–171.
- Crother, B. I., White, M. E., and Johnson, A. D. (2007). Inferring developmental constraint and constraint release: Primordial germ cell determination mechanisms as examples. *Journal of Theoretical Biology*, 248(2):322–330.
- Dang, D. T., Pevsner, J., and Yang, V. W. (2000). The biology of the mammalian krüppel-like family of transcription factors. *The International Journal of Biochemistry & Cell Biology*, 32(11-12):1103–1121.

- Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165.
- Daugherty, C. H., Cree, A., Hay, J. M., and Thompson, M. B. (1990). Neglected taxonomy and continuing extinctions of tuatara (sphenodon). *Nature*, 347:177–179.
- Dávalos, L. M., Cirranello, A. L., Geisler, J. H., and Simmons, N. B. (2012). Understanding phylogenetic incongruence: lessons from phyllostomid bats. *Biological Reviews of the Cambridge Philosophical Society*, 87:991–1024.
- Davis, J. C., Brandman, O., and Petrov, D. A. (2005). Protein evolution in the context of drosophila development. *Journal of Molecular Evolution*, 60:774–785.
- Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, pages 345–352. Natl. Biomed. Res. Found., Washington, DC.
- De Felici, M. (2013). Origin, migration, and proliferation of human primordial germ cells. In Coticchio, G., Albertini, D. F., and Santis, L. D., editors, *Oogenesis*, chapter 2, pages 19–39. Springer, London.
- De Sousa Lopes, S. M. C., Roelen, B. A., Monteiro, R. M., Emmens, R., Lin, H. Y., et al. (2004). BMP signaling mediated by ALK2 in the visceral endoderm is necessary for the generation of primordial germ cells in the mouse embryo. *Genes & Development*, 18(15):1838–1849.
- Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10):1700–1708.
- Delarbre, C., Gallut, C., Barriel, V., Janvier, P., and Gachelin, G. (2002). Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: The comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. *Molecular Phylogenetics and Evolution*, 22(2):184–192.
- Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439:965–968.
- Dixon, J. E., Allegrucci, C., Redwood, C., Kump, K., Bian, Y., et al. (2010). Axolotl Nanog activity in mouse embryonic stem cells demonstrates that ground state pluripotency is conserved from urodele amphibians to mammals. *Development*, 137:2973–2980.
- Domazet-Lošo, T., Brajković, J., and Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, 23(11):533–539.
- Domazet-Lošo, T. and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468:815–818.
- Du Preez, L. H., Kunene, N., Everson, G. J., Carr, J. A., Giesy, J. P., et al. (2008). Reproduction, larval growth, and reproductive development in african clawed frogs (*Xenopus laevis*) exposed to atrazine. *Chemosphere*, 71(3):546–552.
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.
- Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48.
- Elinson, R. P., Sabo, M. C., Fisher, C., Yamaguchi, T., Orii, H., et al. (2011). Germ plasm in *Eleutherodactylus coqui*, a direct developing frog with large eggs. *EvoDevo*, 2(20).
- Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. (2003). Reliability of bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, 52(5):665–673.
- Evans, S. E. (2003). At the feet of the dinosaurs: the early history and radiation of lizards. *Biological Reviews of the Cambridge Philosophical Society*, 78:513–551.

- Evans, T., Wade, C. M., Chapman, F. A., Johnson, A. D., and Loose, M. (2014). Acquisition of germ plasm accelerates vertebrate evolution. *Science*, 344(6180):200–203.
- Extavour, C. G. and Akam, M. (2003). Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development*, 130(24):5869–5884.
- Eyal-Giladi, H., Ginsburg, M., and Farbarov, A. (1981). Avian primordial germ cells are of epiblastic origin. *Journal of Embryology and Experimental Morphology*, 65:139–147.
- Fairbanks, D. J. and Maughan, P. J. (2006). Evolution of the NANOG pseudogene family in the human and chimpanzee genomes. *BMC Evolutionary Biology*, 6:12.
- Farris, J. S. (1970). Methods for computing wagner trees. *Systematic Zoology*, 19(1):83–92.
- Fellah, J. S., Wiles, M. V., Charlemagne, J., and Schwager, J. (1992). Evolution of vertebrate IgM: complete amino acid sequence of the constant region of *Ambystoma mexicanum* μ chain deduced from cDNA sequences. *European Journal of Immunology*, 22(10):2595–2601.
- Feller, A. E. and Hedges, S. (1998). Molecular evidence for the early history of living amphibians. *Molecular Phylogenetics and Evolution*, 9(3):509–516.
- Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.
- Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791.
- Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416.
- Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., et al. (2013). Ensembl 2013. *Nucleic Acids Research*, 41(D1):D48–D55.
- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., et al. (2014). Ensembl 2014. *Nucleic Acids Research*, 42(D1):D749–D755.
- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., et al. (2012). Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90.
- Forey, T. (2010). A system-wide study of amphibian sequence conservation provides evidence for mechanistic constraint. Master's thesis, University of Nottingham.
- Forristall, C., Pondel, M., Chen, L. H., and King, M. L. (1995). Patterns of localization and cytoskeletal association of two vegetally localized RNAs, Vg1 and Xcat-2. *Development*, 121(1):201–208.
- Foster, P. G. and Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, 48:284–290.
- Frank, D. and Harland, R. M. (1992). Localized expression of a xenopus POU gene depends on cell-autonomous transcriptional activation and induction-dependent inactivation. *Development*, 115:439–448.
- Frankenberg, S., Pask, A., and Renfree, M. B. (2010). The evolution of class V POU domain transcription factors in vertebrates and their characterisation in a marsupial. *Developmental Biology*, 337(1):162–170.
- Frankenberg, S. and Renfree, M. B. (2013). On the origin of POU5F1. *BMC Biology*, 11:56.
- Fritsch, B. (1987). Inner ear of the coelacanth fish *Latimeria* has tetrapod affinities. *Nature*, 327:153–154.
- Frost, D. R., Grant, T., Faivovich, J., Bain, R. H., Haas, A., et al. (2006). The amphibian tree of life. *Bulletin of the American Museum of Natural History*, 297(297):8–370.
- Fry, B. G., Vidal, N., Norman, J. A., Vonk, F. J., Scheib, H., et al. (2006). Early evolution of the venom system in lizards and snakes. *Nature*, 439:584–588.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

- Fujiwara, Y., Komiya, T., Kawabata, H., Sato, M., Fujimoto, H., et al. (1994). Isolation of a dead-family protein gene that encodes a murine homolog of drosophila-vasa and its specific expression in germ-cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 91(25):12258–12262.
- Galis, F. (1999). Why do almost all mammals have seven cervical vertebrae? developmental constraints, Hox genes, and cancer. *Journal of Experimental Zoology*, 285:19–26.
- Galis, F., Dooren, T. J. M. V., Feuth, J. D., Metz, J. A. J., Witkam, A., et al. (2006). Extreme selection in humans against homeotic transformations of cervical vertebrae. *Evolution*, 60(12):2643–2654.
- Garrett-Sinha, L. A., Eberspaecher, H., Seldin, M. F., and de Crombrughe, B. (1996). A gene for a novel zinc-finger protein expressed in differentiated epithelial cells and transiently in certain mesenchymal cells. *Journal of Biological Chemistry*, 271:31384–31390.
- Gilbert, S. F. (2006). *Developmental Biology*. Sinauer Associates Inc., 8th edition edition.
- Ginsburg, M. and Eyal-Giladi, H. (1987). Primordial germ cells of the young chick blastoderm originate from the central zone of the area pellucida irrespective of the embryo-forming process. *Development*, 101:209–219.
- Ginsburg, M., Snow, M. H. L., and McLaren, A. (1990). Primordial germ cells in the mouse embryo during gastrulation. *Development*, 110:521–528.
- Goetting-Minesky, M. P. and Makova, K. D. (2006). Mammalian male mutation bias: Impacts of generation time and regional variation in substitution rates. *Journal of Molecular Evolution*, 63(4):537–544.
- Goldman, N., Anderson, J. P., and Rodrigo, A. G. (2000). Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49(4):652–670.
- Golub, R., André, S., Hassanin, A., Affaticati, P., Larijani, M., et al. (2004). Early expression of two TdT isoforms in the hematopoietic system of the mexican axolotl. *Immunogenetics*, 56(3):204–213.
- Graham, V., Khudyakov, J., Ellis, P., and Pevny, L. (2003). SOX2 functions to maintain neural progenitor identity. *Neuron*, 39(5):749–765.
- Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology*, 47(1):9–17.
- Gregory, T. (2014). Animal genome size database. <http://www.genomesize.com>.
- Gubbay, J., Collingnon, J., Koopman, P., Capel, B., Economou, A., et al. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature*, 346:245–250.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., et al. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.
- Guo, G., Yang, J., Nichols, J., Hall, J. S., Eyres, I., et al. (2009). Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development*, 136:1063–1069.
- Haldane, J. B. S. (1949). The rate of mutation of human genes. *Hereditas*, 35(S1):267–273.
- Han, M., An, J., and Kim, W. (2001). Expression patterns of Fgf-8 during development and limb regeneration of the axolotl. *Developmental Dynamics*, 220(1):40–48.
- Hashimoto, Y., Maegawa, S., Nagai, T., Yamaha, E., Suzuki, H., et al. (2004). Localized maternal factors are required for zebrafish germ cell formation. *Developmental Biology*, 268(1):152–161.
- Haston, K. M., Tung, J. Y., and Pera, R. A. R. (2009). Dazl functions in maintenance of pluripotency and genetic and epigenetic programs of differentiation in mouse primordial germ cells in vivo and in vitro. *PLoS ONE*, 4(5):e5654.
- Hasunuma, I., Sakai, T., Nakada, T., Toyoda, F., Namiki, H., et al. (2007). Molecular cloning of three types of arginine vasotocin receptor in the newt, *Cynops pyrrhogaster*. *General and Comparative Endocrinology*, 151(3):252–258.

- Heasman, J., Quarmby, J., and Wylie, C. (1984). The mitochondrial cloud of xenopus oocytes: The source of germinal granule material. *Developmental Biology*, 105(2):458–469.
- Hedges, S. B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23):2971–2972.
- Hedges, S. B., Hass, C. A., and Maxson, L. R. (1993). Relations of fish and tetrapods. *Nature*, 363:501–502.
- Hedges, S. B., Moberg, K. D., and Maxson, L. R. (1990). Tetrapod phylogeny inferred from 18S-ribosomal and 28S-ribosomal RNA sequences and a review of the evidence for amniote relationships. *Molecular Biology and Evolution*, 7(6):607–633.
- Heimberg, A. M., Cowper-Salálari, R., Sémon, M., Donoghue, P. C. J., and Peterson, K. J. (2010). microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, 107(45):19379–19383.
- Hellsten, U., Harland, R. M., Gilchrist, M. J., Hendrix, D., Jurka, J., et al. (2010). The genome of the western clawed frog *Xenopus tropicalis*. *Science*, 328(5978):633–636.
- Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22):10915–10919.
- Herpin, A., Rohr, S., Riedel, D., Kluever, N., Raz, E., et al. (2007). Specification of primordial germ cells in medaka (*Oryzias latipes*). *BMC Developmental Biology*, 7:3.
- Hickford, D. E., Frankenberg, S., Pask, A. J., Shaw, G., and Renfree, M. B. (2011). DDX4 (VASA) is conserved in germ cell development in marsupials and monotremes. *Biology of Reproduction*, 85(4):733–743.
- Hillman, S. S., Withers, P. C., Drewes, R. C., and Hillyard, S. D. (2009). *Ecological and Environmental Physiology of Amphibians*, volume 1 of *Ecological and Environmental Physiology Series*. Oxford University Press, New York.
- Hinkley, C. S., Martin, J. F., Leibham, D., and Perry, M. (1992). Sequential expression of multiple POU proteins during amphibian early development. *Molecular and Cellular Biology*, 12(2):638–649.
- Hoegg, S., Brinkmann, H., Taylor, J. S., and Meyer, A. (2004). Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *Journal of Molecular Evolution*, 59(2):190–203.
- Holland, P. W. and Takahashi, T. (2005). The evolution of homeobox genes: Implications for the study of brain development. *Brain Research Bulletin*, 66:484–490.
- Houston, D. W. and King, M. L. (2000). A critical role for *Xdazl*, a germ plasm-localized RNA, in the differentiation of primordial germ cells in xenopus. *Development*, 127(3):447–456.
- Houston, D. W., Zhang, J., Maines, J. Z., Wasserman, S. A., and King, M. L. (1998). A xenopus DAZ-like gene encodes an RNA component of germ plasm and is a functional homologue of *drosophila boule*. *Development*, 125(2):171–180.
- Huang, C. and Peng, J. (2005). Evolutionary conservation and diversification of Rh family genes and proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15512–15517.
- Huang, X. Q. and Madan, A. (1999). CAP3: A DNA sequence assembly programs. *Genome Research*, 9(9):868–877.
- Hubert, J. (1985). Origin and development of oocytes. In Gans, C., Billett, F. S., and Maderson, P. F. A., editors, *Biology of the Reptilia, Development A*, volume 14. New York, John Wiley & Sons.
- Hudson, C. and Woodland, H. R. (1998). *Xpat*, a gene expressed specifically in germ plasm and primordial germ cells of *Xenopus laevis*. *Mechanisms of Development*, 73(2):159–168.

- Huelsenbeck, J. P. and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology Evolution and Systematics*, 28:437–466.
- Hughes, S. and Mouchiroud, D. (2001). High evolutionary rates in nuclear genes of squamates. *Journal of Molecular Evolution*, 53(1):70–76.
- Hurley, I. A., Mueller, R. L., Dunn, K. A., Schmidt, E. J., Friedman, M., et al. (2007). A new time-scale for ray-finned fish evolution. *Proceedings of the Royal Society B*, 274(1609):489–498.
- Hutchins, A. P., Choo, S. H., Mistri, T. K., Rahmani, M., Woon, C. T., et al. (2013). Co-motif discovery identifies an Esrrb-Sox2-DNA ternary complex as a mediator of transcriptional differences between mouse embryonic and epiblast stem cells. *Stem Cells*, 31(2):269–281.
- Ikenishi, K. (1998). Germ plasm in *Caenorhabditis elegans*, *Drosophila* and *Xenopus*. *Development Growth & Differentiation*, 40:1–10.
- Ikenishi, K. and Tanaka, T. (1997). Involvement of the protein of *Xenopus* vasa homolog (*Xenopus* vasa-like gene 1, XVLG1) in the differentiation of primordial germ cells. *Development Growth & Differentiation*, 39(5):625–633.
- Ikenishi, K., Tanaka, T. S., and Komiyama, T. (1996). Spatio-temporal distribution of the protein of *Xenopus* vasa homologue (*Xenopus* vasa-like gene 1, XVLG1) in embryos. *Development Growth & Differentiation*, 38:527–535.
- Illmensee, K. and Mahowald, A. P. (1974). Transplantation of posterior polar plasm in *Drosophila*. induction of germ cells at the anterior pole of the egg. *Proceedings of the National Academy of Sciences of the United States of America*, 71(4):1016–1020.
- Inoue, J. G., Miya, M., Lam, K., Tay, B.-H., Danks, J. A., et al. (2010). Evolutionary origin and phylogeny of the modern holocephalans (chondrichthyes: Chimaeriformes): A mitogenomic perspective. *Molecular Biology and Evolution*, 27(11):2576–2586.
- Inoue, J. G., Miya, M., Tsukamoto, K., and Nishida, M. (2003). Basal actinopterygian relationships: a mitogenomic perspective on the phylogeny of the ‘ancient fish’. *Molecular Phylogenetics and Evolution*, 26(1):110–120.
- Iwabe, N., Hara, Y., Kumazawa, Y., Shibamoto, K., Saito, Y., et al. (2005). Sister group relationship of turtles to the bird-crocodylian clade revealed by nuclear dna-coded proteins. *Molecular Biology and Evolution*, 22(4):810–813.
- Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957.
- James-Zorn, C., Ponferrada, V. G., Jarabek, C. J., Burns, K. A., Segerdell, E. J., et al. (2013). Xenbase: expansion and updates of the *Xenopus* model organism databases. *Nucleic Acids Research*, 41(D1):D865–D870.
- Janke, A., Xu, X., and Arnason, U. (1997). The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proceedings of the National Academy of Sciences of the United States of America*, 94(4):1276–1281.
- Johnson, A. D., Bachvarova, R. F., Drum, M., and Masi, T. (2001). Expression of axolotl DAZL RNA, a marker of germ plasm: Widespread maternal RNA and onset of expression in germ cells approaching the gonads. *Developmental Biology*, 234(2):402–415.
- Johnson, A. D., Crother, B., White, M. E., Patient, R., Bachvarova, R. F., et al. (2003a). Regulative germ cell specification in axolotl embryos: a primitive trait conserved in the mammalian lineage. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 358(1436):1371–1379.
- Johnson, A. D., Drum, M., Bachvarova, R. F., Masi, T., White, M. E., et al. (2003b). Evolution of predetermined germ cells in vertebrate embryos: implications for macroevolution. *Evolution & Development*, 4(4):414–431.

- Johnson, A. D., Richardson, E., Bachvarova, R. F., and Crother, B. I. (2011). Evolution of the germ line-soma relationship in vertebrate embryos. *Reproduction*, 141:291–300.
- Jones, D. T., Taylor, W. R., and Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences*, 8(3):275–282.
- Jones, M. E. H., Anderson, C. L., Hipsley, C. A., Müller, J., Evans, S. E., et al. (2013). Integration of molecules and new fossils supports a triassic origin for lepidosauria (lizards, snakes, and tuatara). *BMC Evolutionary Biology*, 13:208.
- Kagami, H., Tagami, T., Matsubara, Y., Harumi, T., Hanada, H., et al. (1997). The developmental origin of primordial germ cells and the transmission of the donor-derived gametes in mixed-sex germline chimeras to the offspring in the chicken. *Molecular Reproduction and Development*, 48(4):501–510.
- Kamachi, Y., Uchikawa, M., Collignon, J., Lovell-Badge, R., and Kondoh, H. (1998). Involvement of Sox1, 2 and 3 in the early and subsequent molecular events of lens inductions. *Development*, 125:2521–2532.
- Kastler, S., Honold, L., Luedeke, M., Kuefer, R., Möller, P., et al. (2010). POU5F1P1, a putative cancer susceptibility gene, is overexpressed in prostatic carcinoma. *Prostate*, 70(6):666–674.
- Katsu, Y., Kohno, S., Oka, T., Mitsui, N., Tooi, O., et al. (2006). Molecular cloning of estrogen receptor alpha (ER α ; ESR1) of the japanese giant salamander, andrias japonicus. *Molecular and Cellular Endocrinology*, 257-258:84–94.
- Katz, J. P., Perreault, N., Goldstein, B. G., Lee, C. S., Labosky, P. A., et al. (2002). The zinc-finger transcription factor Klf4 is required for terminal differentiation of goblet cells in the colon. *Development*, 129:2619–2628.
- Kehler, J., Tolkunova, E., Koschorz, B., Pesce, M., Gentile, L., et al. (2004). Oct4 is required for primordial germ cell survival. *EMBO Reports*, 5:1078–1083.
- Kiemiec-Tyburczy, K. M., Watts, R. A., and Arnold, S. J. (2011). Characterization of two putative cytokine receptors, gp130 and ciliary neurotrophic factor receptor, from terrestrial salamanders. *Genes and Genetic Systems*, 86(2):131–137.
- Killian, J. K., Buckley, T. R., Stewart, N., Munday, B. L., and Jirtle, R. L. (2001). Marsupials and eutherians reunited: genetic evidence for the theria hypothesis of mammalian evolution. *Mammalian Genome*, 12(7):513–517.
- Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B., and Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Developmental Dynamics*, 203:253–310.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–626.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *Journal of Molecular Evolution*, 16(2):111–120.
- Kimura, M. and Ohta, T. (1971). On the rate of molecular evolution. *Journal of Molecular Evolution*, 1(1):1–17.
- King, B. L., Gillis, J. A., Carlisle, H. R., and Dahn, R. D. (2011). A natural deletion of the HoxC cluster in elasmobranch fishes. *Science*, 334(6062):1517.
- Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29(2):170–179.
- Ko, C., Chesnel, A., Mazerbourg, S., Kuntz, S., Flament, S., et al. (2008). Female-enriched expression of ER α during gonad differentiation of the urodele amphibian *Pleurodeles waltl*. *General and Comparative Endocrinology*, 156(2):234–245.
- Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39:309–338.

- Koprunner, M., Thisse, C., Thisse, B., and Raz, E. (2001). A zebrafish nanos-related gene is essential for the development of primordial germ cells. *Genes & Development*, 15(21):2877–2885.
- Kosaka, K., Kawakami, K., Sakamoto, H., and Inoue, K. (2007). Spatiotemporal localization of germ plasm RNAs during zebrafish oogenesis. *Mechanisms of Development*, 124(4):279–289.
- Koshihara, K., Kuroiwa, A., Yamamoto, H., Tamura, K., and Ide, H. (1998). Expression of Msx genes in regenerating and developing limbs of axolotl. *Journal of Experimental Zoology*, 282(6):703–714.
- Kumazawa, Y. and Nishida, M. (1999). Complete mitochondrial dna sequences of the green turtle and blue-tailed mole skink: Statistical evidence for archosaurian affinity of turtles. *Molecular Biology and Evolution*, 16(6):784–792.
- Kurimoto, K., Yabuta, Y., Ohinata, Y., Shigeta, M., Yamanaka, K., et al. (2008). Complex genome-wide transcription dynamics orchestrated by Blimp1 for the specification of the germ cell lineage in mice. *Genes & Development*, 22(12):1617–1635.
- Kuwana, T. (1993). Migration of avian primordial germ cells toward the gonadal anlage. *Development Growth & Differentiation*, 35(3):237–243.
- Lai, F., Singh, A., and King, M. L. (2012). Xenopus Nanos1 is required to prevent endoderm gene expression and apoptosis in primordial germ cells. *Development*, 139:1476–1486.
- Lanfear, R., Kokko, H., and Eyre-Walker, A. (2013). Population size and the rate of evolution. *Trends in Ecology and Evolution*.
- Lariviere, K., MacEachern, L., Greco, V., Majchrzak, G., Chiu, S., et al. (2002). GAD65 and GAD67 isoforms of the glutamic acid decarboxylase gene originated before the divergence of cartilaginous fishes. *Molecular Biology and Evolution*, 19(12):2325–2329.
- Laurens, V., Chapusot, C., del Rosario Ordonez, M., Bentrari, F., Padros, M. R., et al. (2001). Axolotl MHC class II β chain: predominance of one allele and alternative splicing of the β -1 domain. *European Journal of Immunology*, 31(2):506–515.
- Laurin, M. and Reisz, R. R. (1995). A reevaluation of early amniote phylogeny. *Zoological Journal of the Linnean Society*, 113(2):165–223.
- Lawson, K. A., Dunn, N. R., Roelen, B. A., Zeinstra, L. M., Davis, A. M., et al. (1999). Bmp4 is required for the generation of primordial germ cells in the mouse embryo. *Genes & Development*, 13(4):424–436.
- Lawson, K. A. and Hage, W. J. (1994). Clonal analysis of the origin of primordial germ cells in the mouse. In Marsh, J. and Goode, J., editors, *Ciba Foundation Symposium 182 - Germline Development*, volume 182, pages 68–91. John Wiley & Sons, Ltd., Chichester, UK.
- Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.
- Lee, M. S. (2000). Soft anatomy, diffuse homoplasy, and the relationships of lizards and snakes. *Zoologica Scripta*, 29(2):101–130.
- Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., et al. (2005). The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Research*, 33:D71–D74.
- Lemey, P., Salemi, M., and Vandamme, A.-M., editors (2009). *The Phylogenetic Handbook*. Cambridge University Press, second edition.
- Lewis, S. L., Khoo, P., de Young, R. A., Steiner, K., Wilcock, C., et al. (2008). Dkk1 and Wnt3 interact to control head morphogenesis in the mouse. *Development*, 135:1791–1801.
- Li, W. Z. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

- Liang, D., Shen, X. X., and Zhang, P. (2013). One thousand two hundred ninety nuclear genes from a genome-wide survey support lungfishes as the sister group of tetrapods. *Molecular Biology and Evolution*, 30(8):1803–1807.
- Lin, Y. and Page, D. C. (2005). Dazl deficiency leads to embryonic arrest of germ cell development in XY C57BL/6 mice. *Developmental Biology*, 288(2):309–316.
- Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11(4):605–612.
- Loose, M. and Patient, R. (2004). A genetic regulatory network for xenopus mesoderm formation. *Developmental Biology*, 271(2):467–478.
- Looso, M., Michel, C. S., Konzer, A., Bruckskotten, M., Borchardt, T., et al. (2012). Spiked-in pulsed in vivo labeling identifies a new member of the CCN family in regenerating newt hearts. *Journal of Proteome Research*, 11(9):4693–4704.
- Lunde, K., Belting, H.-G., and Driever, W. (2004). Zebrafish pou5f1/pou2, homolog of mammalian Oct4, functions in the endoderm specification cascade. *Current Biology*, 14(1):48–55.
- Luo, Z., Cifelli, R. L., and Kielan-Jaworowska, Z. (2001). Dual origin of tribosphenic mammals. *Nature*, 409:53–57.
- Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155.
- MacArthur, H., Bubunenko, M., Houston, D. W., and King, M. L. (1999). Xcat RNA is a translationally sequestered germ plasm component in xenopus. *Mechanisms of Development*, 84(1-2):75–88.
- Machado, R. J., Moore, W., Hames, R., Houliston, E., Chang, P., et al. (2005). Xenopus Xpat protein is a major component of germ plasm and may function in its organisation and positioning. *Developmental Biology*, 287(2):289–300.
- Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.
- Maddison, W. P. and Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55(1):21–30.
- Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., et al. (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell*, 1(1):55–70.
- Maisey, J. G. (1986). Heads and tails: A chordate phylogeny. *Cladistics*, 2(4):201–256.
- Maki, N., Suetsugu-Maki, R., Tarui, H., Agata, K., Rio-Tsonis, K. D., et al. (2009). Expression of stem cell pluripotency factors during regeneration in newts. *Developmental Dynamics*, 238(6):1613–1616.
- Marracci, S., Michelotti, V., Casola, C., Giacomini, C., and Raghianti, M. (2011). Daz and Pumilio-like genes are asymmetrically localized in Pelophylax (Rana) oocytes and are expressed during early spermatogenesis. *Journal of Experimental Zoology*, 316B(5):330–338.
- Martin, A. P. and Palumbi, S. R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, 90(9):4087–4091.
- Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nature Cell Biology*, 9:625–635.
- Maurizii, M. G., Cavaliere, V., Gamberi, C., Lasko, P., Gargiulo, G., et al. (2009). Vasa protein is localized in the germ cells and in the oocyte-associated pyriform follicle cells during early oogenesis in the lizard Podarcis sicula. *Development Genes and Evolution*, 219(7):361–367.
- McLean, L. A., Zia, S., Gorin, F. A., and Cala, P. M. (1999). Cloning and expression of the Na⁺/H⁺ exchanger from amphiuma RBCs: resemblance to mammalian NHE1. *American Journal of Physiology Cell Physiology*, 276(5):C1025–C1037.

- Minelli, A., editor (2008). *Perspectives in Animal Phylogeny & Evolution*. Oxford University Press, UK.
- Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., et al. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, 113(5):631–642.
- Miyazaki, K., Uchiyama, K., Imokawa, Y., and Yoshizato, K. (1996). Cloning and characterization of cDNAs for matrix metalloproteinases of regenerating newt limbs. *Proceedings of the National Academy of Sciences of the United States of America*, 93:6819–6824.
- Modrell, M. S. and Baker, C. V. H. (2012). Evolution of electrosensory ampullary organs: conservation of *Eya4* expression during lateral line development in jawed vertebrates. *Evolution and Development*, 14(3):277–285.
- Møllgård, K., Jespersen, A., Lutterodt, M., Andersen, C. Y., Høyer, P., et al. (2010). Human primordial germ cells migrate along nerve fibers and schwann cells from the dorsal hind gut mesentery to the gonadal ridge. *Molecular Human Reproduction*, 16(9):621–631.
- Morrison, G. M. and Brickman, J. M. (2006). Conserved roles for Oct4 homologues in maintaining multipotency during early vertebrate development. *Development*, 133:2011–2022.
- Mosquera, L., Forristall, C., Zhou, Y., and King, M. L. (1993). A mRNA localized to the vegetal cortex of xenopus oocytes encodes a protein with a nanos-like zinc finger domain. *Development*, 117:377–386.
- Mounaji, K., Erraiss, N.-E., Iddar, A., Wegnez, M., Serrano, A., et al. (2002). Glyceraldehyde-3-phosphate dehydrogenase from the newt *Pleurodeles waltl*. protein purification and characterization of a GapC genes. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 131(3):411–421.
- Mullin, N. P., Yates, A., Rowe, A. J., Nijmeijer, B., Colby, D., et al. (2008). The pluripotency rheostat Nanog functions as a dimer. *Biochemical Journal*, 411:227–231.
- Muse, S. V. and Weir, B. S. (1992). Testing for equality of evolutionary rates. *Genetics*, 132(1):269–276.
- Nabholz, B., Glémin, S., and Galtier, N. (2008). Strong variations of mitochondrial mutation rate across mammals - the longevity hypothesis. *Molecular Biology and Evolution*, 25(1):120–130.
- Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304.
- Nakatake, Y., Fukui, N., Iwamatsu, Y., Masui, S., Takahashi, K., et al. (2006). Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. *Molecular and Cellular Biology*, 26(20):7772–7782.
- Nath, K., Boorech, J. L., Beckham, Y. M., Burns, M. M., and Elinson, R. P. (2005). Status of RNAs, localized in *Xenopus laevis* oocytes, in the frogs *Rana pipiens* and *Eleutherodactylus coqui*. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304B(1):28–39.
- Nath, K. and Elinson, R. P. (2007). RNA of AmVegT, the axolotl orthologue of the xenopus meso-endodermal determinant, is not localized in the oocytes. *Gene Expression Patterns*, 7(1-2):197–201.
- Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., et al. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, 109(34):13698–13703.
- Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., et al. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*, 95(3):379–391.
- Nieuwkoop, P. (1974). Experimental investigations on the origin and determination of the germ cells, and on the development of the lateral plates and germ ridges in urodeles. *Archives Néerlandaises de Zoologie*, 8(1-4):1–205.

- Niwa, H., Miyazaki, J., and Smith, A. G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genetics*, 24:372–376.
- Niwa, H., Sekita, Y., Tsend-Ayush, E., and Grützner, F. (2008). Platypus Pou5f1 reveals the first steps in the evolution of trophoderm differentiation and pluripotency in mammals. *Evolution and Development*, 10(6):671–682.
- Norén, M. and Jondelius, U. (1997). Xenoturbella's molluscan relatives... *Nature*, 390:31–32.
- Ohinata, Y., Ohta, H., Shigeta, M., Yamanaka, K., Wakayama, T., et al. (2009). A signaling principle for the specification of the germ cell lineage in mice. *Cell*, 137:571–584.
- Ohinata, Y., Payer, B., O'Carroll, D., Ancelin, K., Ono, Y., et al. (2005). Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature*, 436:207–213.
- Ohta, T. (1987). Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution*, 26(1-2):1–6.
- Ohta, T. and Kimura, M. (1971). On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution*, 1:18–25.
- Oisi, Y., Ota, K. G., Kuraku, S., Fujimoto, S., and Kuratani, S. (2013). Craniofacial development of hagfishes and the evolution of vertebrates. *Nature*, 493:175–180.
- Okamoto, K., Okazawa, H., Okuda, A., Sakai, M., Muramatsu, M., et al. (1990). A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells. *Cell*, 60(3):461–472.
- Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. *Nature*, 448:313–317.
- Onichtchouk, D. (2012). Pou5f1/Oct4 in pluripotency control: Insights from zebrafish. *Genesis*, 50:75–85.
- Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583.
- Panopoulou, G., Hennig, S., Groth, D., Krause, A., Poustka, A. J., et al. (2003). New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Research*, 13(6):1056–1066.
- Pearse, 2nd, R. V., Drolet, D. W., Kalla, K. A., Hooshmand, F., Bermingham, Jr., J. R., et al. (1997). Reduced fertility in mice deficient for the POU protein sperm-1. *Proceedings of the National Academy of Sciences of the United States of America*, 94:7555–7560.
- Peng, J. X., Xie, J. L., Zhou, L., Hong, Y. H., and Gui, J. F. (2009). Evolutionary conservation of Dazl genomic organization and its continuous and dynamic distribution throughout germline development in gynogenetic gibel carp. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 312B(8):855–871.
- Perea-Gomez, A., Vella, F. D., Shawlot, W., Oulad-Abdelghani, M., Chazaud, C., et al. (2002). Nodal antagonists in the anterior visceral endoderm prevent the formation of multiple primitive streaks. *Developmental Cell*, 3(5):745–756.
- Pereira, L. A., Wong, M. S., Lim, S. M., Stanley, E. G., and Elefanty, A. G. (2012). The Mix family of homeobox genes - key regulators of mesendoderm formation during vertebrate development. *Developmental Biology*, 367(2):163–177.
- Peterson, K. J. (2004). Isolation of Hox and Parahox genes in the hemichordate *Ptychodera flava* and the evolution of deuterostome Hox genes. *Molecular Phylogenetics and Evolution*, 31(3):1208–1215.
- Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, 17(1):23–28.
- Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., et al. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, 9(3):e1000602.
- Philippe, H. and Laurent, J. (1998). How good are deep phylogenetic trees? *Current Opinion in Genetics & Development*, 8(6):616–623.

- Phillips, M. J. and Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution*, 28(2):171–185.
- Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679.
- Posada, D. and Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitutions. *Bioinformatics*, 14(9):814–818.
- Pough, H. F., Janis, C. M., and Heiser, J. B. (2009). *Vertebrate Life*. Pearson Education, San Francisco, London, 8th edition.
- Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453:1064–1071.
- Pyron, R. A. and Wiens, J. J. (2011). A large-scale phylogeny of amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, 61(2):543–583.
- Rasmussen, A.-S., Janke, A., and Arnason, U. (1998). The mitochondrial DNA molecule of the hagfish (*Myxine glutinosa*) and vertebrate phylogeny. *Journal of Molecular Evolution*, 46:382–388.
- Rest, J. S., Ast, J. C., Austin, C. C., Waddell, P. J., Tibbetts, E. A., et al. (2003). Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Molecular Phylogenetics and Evolution*, 29(2):289–297.
- Robb, L., Hartley, L., Begley, C. G., Brodnicki, T. C., Copeland, N. G., et al. (2000). Cloning, expression analysis, and chromosomal localization of murine and human homologues of a xenopus *Mix* genes. *Developmental Dynamics*, 219(4):497–504.
- Robinson, M., Gouy, M., Gautier, C., and Mouchiroud, D. (1998). Sensitivity of the relative-rate test to taxonomic sampling. *Molecular Biology and Evolution*, 15(9):1091–1098.
- Rodda, D. J., Chew, J.-L., Lim, L.-H., Loh, Y.-H., Wang, B., et al. (2005). Transcriptional regulation of Nanog by OCT4 and SOX2. *The Journal of Biological Chemistry*, 280:24731–24737.
- Rokas, A., King, N., Finnerty, J., and Carroll, S. B. (2003a). Conflicting phylogenetic signals at the base of the metazoan tree. *Evolution and Development*, 5(4):346–359.
- Rokas, A., Williams, B. L., King, N., and Carroll, S. B. (2003b). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., et al. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.
- Ronquist, F., van der Mark, P., and Huelsenbeck, J. P. (2009). Bayesian phylogenetic analyses using MrBayes. In Lemey, P., Salemi, M., and Vandamme, A.-M., editors, *The Phylogenetic Handbook*, chapter 7, pages 210–266. Cambridge University Press, second edition.
- Rosner, M. H., Vigano, M. A., Ozato, K., Timmons, P. M., Poirie, F., et al. (1990). A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryos. *Nature*, 345:686–692.
- Roux, J. and Robinson-Rechavi, M. (2008). Developmental constraints on vertebrate genome evolution. *PLoS Genetics*, 4(12):e1000311.
- Rowland, B. D., Bernard, R., and Peeper, D. S. (2005). The KLF4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene. *Nature Cell Biology*, 7:1074–1082.
- Ruggiu, M., Speed, R., Taggart, M., McKay, S. J., Kilanowski, F., et al. (1997). The mouse *Dazl* gene encodes a cytoplasmic protein essential for gametogenesis. *Nature*, 389:73–77.
- Saffman, E. E. and Lasko, P. (1999). Germline development in vertebrates and invertebrates. *Cellular and Molecular Life Sciences*, 55(8-9):1141–1163.

- Safi, R., Begue, A., Hänni, C., Stehelin, D., Tata, J. R., et al. (1997). Thyroid hormone receptor genes of neotenic amphibians. *Journal of Molecular Evolution*, 44(6):595–604.
- Saito, A., Kano, Y., Suzuki, M., Tomura, H., Takeda, J., et al. (2002). Sequence analysis and expressional regulation of messenger RNAs encoding beta subunits of follicle-stimulating hormone and luteinizing hormone in the red-bellied newt, *Cynops pyrrhogaster*. *Biology of Reproduction*, 66(5):1299–1309.
- Saito, T., Pšenička, M., Goto, R., Adachi, S., Inoue, K., et al. (2014). The origin and migration of primordial germ cells in sturgeons. *PLoS ONE*, 9(2):e86861.
- Saitou, M., Barton, S. C., and Surani, M. A. (2002). A molecular programme for the specification of germ cell fate in mice. *Nature*, 418(6895):293–300.
- Saitou, M. and Yamaji, M. (2010). Germ cell specification in mice: signaling, transcription regulation, and epigenetic consequences. *Reproduction*, 139:931–942.
- Saitou, M. and Yamaji, M. (2012). Primordial germ cells in mice. *Cold Spring Harbor Perspectives in Biology*, 4:a008375.
- Saitou, N. and Nei, M. (1987). The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.
- Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497:327–331.
- Sammut, B., Pasquier, L. D., Ducoroy, P., Laurens, V., Marcuz, A., et al. (1999). Axolotl MHC architecture and polymorphism. *European Journal of Immunology*, 29(9):2897–2907.
- Sanderson, M. J. and Shaffer, H. B. (2002). Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics*, 33:49–72.
- Sanderson, M. J., Wojciechowski, M. F., Hu, J.-M., Khan, T. S., and Brady, S. G. (2000). Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution*, 17(5):782–797.
- Sato, M., Kimura, T., Kurokawa, K., Fujita, Y., Abe, K., et al. (2002). Identification of PGC7, a new gene expressed specifically in preimplantation embryos and germ cells. *Mechanisms of Development*, 113(1):91–94.
- Sauka-Spengler, T., Germot, A., Shi, D., and Mazan, S. (2002). Expression patterns of an Otx2 and an Otx5 orthologue in the urodele *Pleurodeles waltl*: implications on the evolutionary relationships between the balancers and cement gland in amphibians. *Development Genes and Evolution*, 212(8):380–387.
- Savage, R. M. and Danilchik, M. V. (1993). Dynamics of germ plasm localization and its inhibition by ultraviolet irradiation in early cleavage xenopus embryos. *Developmental Biology*, 157(2):371–382.
- Scerbo, P., Girardot, F., Vivien, C., Markov, G. V., Luxardi, G., et al. (2012). Ventx factors function as Nanog-like guardians of developmental potential in xenopus. *PLoS ONE*, 7(5):e36855.
- Scerbo, P., Markov, G. V., Vivien, C., Kodjabachian, L., Demeneix, B., et al. (2014). On the origin and evolutionary history of NANOG. *PLoS ONE*, 9(1):e85104.
- Schaerlinger, B. and Frippiat, J.-P. (2008). IgX antibodies in the urodele amphibian *Ambystoma mexicanum*. *Developmental and Comparative Immunology*, 32(8):908–915.
- Schepers, G. E., Teasdale, R. D., and Koopman, P. (2002). Twenty pairs of Sox: Extent, homology, and nomenclature of the mouse and human Sox transcription factor gene families. *Developmental Cell*, 3(2):167–170.
- Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–504.
- Schmidt, H. A. and von Haeseler, A. (2009). Phylogenetic inference using maximum likelihood methods. In Lemey, P., Salemi, M., and Vandamme, A.-M., editors, *The Phylogenetic Handbook*, chapter 6, pages 181–209. Cambridge University Press, second edition.

- Schöler, H. R., Dressler, G. R., Balling, R., Rohdewohld, H., and Gruss, P. (1990a). Oct-4: a germline-specific transcription factor mapping to the mouse t-complex. *EMBO Journal*, 9(7):2185–2195.
- Schöler, H. R., Hatzopoulos, A. K., Balling, R., Suzuki, N., and Gruss, P. (1989). A family of octamer-specific proteins present during mouse embryogenesis: evidence for germline-specific expression of an Oct factors. *The EMBO Journal*, 8(9):2543–2550.
- Schöler, H. R., Ruppert, S., Suzuki, N., Chowdhury, K., and Gruss, P. (1990b). New type of POU domain in germ line-specific protein Oct-4. *Nature*, 344:435–439.
- Schuff, M., Siegel, D., Philipp, M., Bundschu, K., Heymann, N., et al. (2012). Characterization of danio rerio Nanog and functional comparison to xenopus Vents. *Stem Cells and Development*, 21(8):1225–1238.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Searcy, B. T., Bradford, C. S., Thompson, R. R., Filtz, T. M., and Moore, F. L. (2011). Identification and characterization of mesotocin and V1a-like vasotocin receptors in a urodele amphibian, *Taricha granulosa*. *General and Comparative Endocrinology*, 170(1):131–143.
- Segre, J. A., Bauer, C., and Fuchs, E. (1999). Klf4 is a transcription factor required for establishing the barrier function of the skin. *Nature Genetics*, 22:356–360.
- Seki, Y., Yamaji, M., Yabuta, Y., Sano, M., Shigeta, M., et al. (2007). Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. *Development*, 134:2627–2638.
- Seligman, J. and Page, D. C. (1998). The Dazh gene is expressed in male and female embryonic gonads before germ cell sex differentiation. *Biochemical and Biophysical Research Communications*, 245(3):878–882.
- Shan, Y. and Gras, R. (2011). 43 genes support the lungfish-coelacanth grouping related to the closest living relative of tetrapods with the bayesian method under the coalescence model. *BMC Research Notes*, 4:49.
- Shedlock, A. M., Botka, C. W., Zhao, S., Shetty, J., Zhang, T., et al. (2007). Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8):2767–2772.
- Shields, J. M., Christy, R. J., and Yang, V. W. (1996). Identification and characterization of a gene encoding a gut-enriched krüppel-like factor expressed during growth arrest. *Journal of Biological Chemistry*, 271:20009–20017.
- Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.
- Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16(8):1114–1116.
- Shimodaira, H. and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247.
- Silva, J., Nichols, J., Theunissen, T. W., Guo, G., van Oosten, A. L., et al. (2009). Nanog is the gateway to the pluripotent ground state. *Cell*, 138(4):722–737.
- Simon, H.-G., Nelson, C., Goff, D., Laufer, E., Morgan, B. A., et al. (1995). Differential expression of myogenic regulatory genes and Msx-1 during dedifferentiation and redifferentiation of regenerating amphibian limbs. *Developmental Dynamics*, 202:1–12.
- Simon, H.-G. and Oppenheimer, S. (1996). Advanced mRNA differential display: isolation of a new differentially regulated myosin heavy chain-encoding gene in amphibian limb regeneration. *Gene*, 172(2):175–181.
- Sinner, D., Kirilenko, P., Rankin, S., Wei, E., Howard, L., et al. (2006). Global analysis of the transcriptional network controlling xenopus endoderm formation. *Development*, 133(10):1955–1966.

- Sites, Jr., J. W., Reeder, T. W., and Wiens, J. J. (2011). Phylogenetic insights on evolutionary novelties in lizards and snakes: Sex, birth, bodies, niches, and venom. *Annual Review of Ecology Evolution and Systematics*, 42:227–244.
- Slack, J. M. W., Darlington, B. G., Heath, J. K., and Godsave, S. F. (1987). Mesoderm induction in early xenopus embryos by heparin-binding growth factors. *Nature*, 326:187–200.
- Smith, J. C. and Slack, J. M. W. (1983). Dorsalization and neural induction: properties of the organizer in *Xenopus laevis*. *Journal of Embryology and Experimental Morphology*, 78:299–317.
- Smith, J. J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., et al. (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature Genetics*, 45:415–421.
- Sone, K., Takahashi, T. C., Takabatake, Y., Takeshima, K., and Takabatake, T. (1999). Expression of Pve novel T-box genes and brachyury during embryogenesis, and in developing and regenerating limbs and tails of newts. *Development Growth & Differentiation*, 41:321–333.
- Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., et al. (2002). The Bioperl Toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618.
- Stark, D. R., Gates, P. B., Brockes, J. P., and Ferretti, P. (1998). Hedgehog family member is expressed throughout regenerating and developing limbs. *Developmental Dynamics*, 212(3):352–363.
- Stewart, R., Rascón, C. A., Tian, S., Nie, J., Barry, C., et al. (2013). Comparative RNA-seq analysis in the unsequenced axolotl: The oncogene burst highlights early gene expression in the blastema. *PLoS Computational Biology*, 9(3):e1002936.
- Strimmer, K. and Rambaut, A. (2002). Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society of London B*, 269:137–142.
- Strome, S. and Wood, W. B. (1982). Immunofluorescence visualization of germ-line-specific cytoplasmic granules in embryos, larvae, and adults of *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the United States of America*, 79(5):1558–1562.
- Studier, J. A. and Keppler, K. J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5(6):729–731.
- Sulston, J. E., Schierenberg, E., White, J. G., and Thomson, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*, 100:64–119.
- Susaki, K., Kaneko, J., Yamano, Y., Nakamura, K., Inami, W., et al. (2009). Musashi-1, an RNA-binding protein, is indispensable for survival of photoreceptors. *Experimental Eye Research*, 88(3):347–355.
- Suzuki, M., Kubokawa, K., Nagasawa, H., and Urano, A. (1995). Sequence analysis of vasotocin cDNAs of the lamprey, *Lampetra japonica*, and the hagfish, *Eptatretus burgeri*: evolution of cyclostome vasotocin precursors. *Journal of Molecular Endocrinology*, 14:67–77.
- Swiers, G., Chen, Y.-H., Johnson, A. D., and Loose, M. (2010). A conserved mechanism for vertebrate mesoderm specification in urodele amphibians and mammals. *Developmental Biology*, 343(1-2):138–152.
- Swofford, D. L. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods)*. Sinauer Associates, Sunderland, Massachusetts. Version 4.0b10 for Unix.
- Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., et al. (2013). Human ageing genomic resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Research*, 41:D1027–D1033.
- Tada, H., Mochii, M., Orii, H., and Watanabe, K. (2012). Ectopic formation of primordial germ cells by transplantation of the germ plasm: Direct evidence for germ cell determinant in xenopus. *Developmental Biology*, 371(1):86–93.

- Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135(2):599–607.
- Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., et al. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872.
- Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676.
- Takahashi, S., Onuma, Y., Yokota, C., Westmoreland, J. J., Asashima, M., et al. (2006). Nodal-related gene *Xnr5* is amplified in the xenopus genomes. *Genesis*, 44(7):309–321.
- Takeda, H., Matsuzaki, T., Oki, T., Miyagawa, T., and Amanuma, H. (1994). A novel POU domain gene, zebrafish *pou2*: expression and roles of two alternatively spliced twin products in early development. *Genes & Development*, 8:45–59.
- Takehita, H., Yasuda, T., Iida, R., Nakajima, T., Mori, S., et al. (2001). Amphibian DNases I are characterized by a C-terminal end with a unique, cysteine-rich stretch and by the insertion of a serine residue into the Ca²⁺-binding site. *Biochemical Journal*, 357:473–480.
- Takezaki, N., Figueroa, F., Zaleska-Rutczynska, Z., and Klein, J. (2003). Molecular phylogeny of early vertebrates: Monophyly of the agnathans as revealed by sequences of 35 genes. *Molecular Biology and Evolution*, 20(2):287–292.
- Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4):564–577.
- Tam, P. P. L. and Loebel, D. A. F. (2007). Gene function in mouse embryogenesis: get set for gastrulation. *Nature Reviews Genetics*, 8:363–381.
- Tam, P. P. L. and Zhou, S. X. (1996). The allocation of epiblast cells to ectodermal and germ-line lineages is influenced by the position of the cells in the gastrulating mouse embryo. *Developmental Biology*, 178(1):124–132.
- Tamori, Y., Iwai, T., Mita, K., and Wakahara, M. (2004). Spatio-temporal expression of a DAZ-like gene in the japanese newt *Cynops pyrrhogaster* that has no germ plasm. *Development Genes and Evolution*, 214(12):615–627.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., et al. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28(10):2731–2739.
- Tanaka, S. S., Toyooka, Y., Akasu, R., Katoh-Fukui, Y., Nakahara, Y., et al. (2000). The mouse homolog of drosophila *Vasa* is required for the development of male germ cells. *Genes & Development*, 14:841–853.
- Tanimura, N., Saito, M., Ebisuya, M., Nishida, E., and Ishikawa, F. (2013). Stemness-related factor *Sall4* interacts with transcription factors Oct-3/4 and Sox2 and occupies Oct-Sox elements in mouse embryonic stem cells. *Journal of Biological Chemistry*, 288:5027–5038.
- Tantin, D. (2013). Oct transcription factors in development and stem cells: insights and mechanisms. *Development*, 140:2857–2866.
- Tapia, N., Reinhardt, P., Duemmler, A., Wu, G., Araúzo-Bravo, M. J., et al. (2012). Reprogramming to pluripotency is an ancient trait of vertebrate Oct4 and Pou2 proteins. *Nature Communications*, 3:1279.
- Taylor, J. S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Research*, 13:382–390.
- Theiler, K. (1989). *The House Mouse - Atlas of Embryonic Development*. Springer-Verlag.

- Theunissen, T. W., Costa, Y., Radzishewska, A., van Oosten, A. L., Laval, F., et al. (2011). Reprogramming capacity of Nanog is functionally conserved in vertebrates and resides in a unique homeodomain. *Development*, 138:4853–4865.
- Theusch, E. V., Brown, K. J., and Pelegri, F. (2006). Separate pathways of RNA recruitment lead to the compartmentalization of the zebrafish germ plasm. *Developmental Biology*, 292(1):129–141.
- Thomas, J. A., Welch, J. J., Lanfear, R., and Bromham, L. (2010). A generation time effect on the rate of molecular evolution in invertebrates. *Molecular Biology and Evolution*, 27(5):1173–1180.
- Ton-That, H., Kaestner, K. H., Shields, J. M., Mahatanankoon, C. S., and Yang, V. W. (1997). Expression of the gut-enriched krüppel-like factor gene during development and intestinal tumorigenesis. *FEBS Letters*, 419(2-3):239–243.
- Tourasse, N. J. and Li, W. H. (1999). Performance of the relative-rate test under nonstationary models of nucleotide substitution. *Molecular Biology and Evolution*, 16(8):1068–1078.
- Toyooka, Y., Tsunekawa, N., Takahashi, Y., Matsui, Y., Satoh, M., et al. (2000). Expression and intracellular localization of mouse Vasa-homologue protein during germ cell development. *Mechanisms of Development*, 93(1-2):139–149.
- Trueb, L. and Cloutier, R. (1991). A phylogenetic investigation of the inter- and intrarelationships of the lissamphibia (amphibia:temnospondyli). In Trueb, L. and Schultz, H.-P., editors, *Origins of the Higher Group of Tetrapods: Controversy and Consensus*, chapter 8, pages 223–313. Cornell University Press.
- Tsunekawa, N., Naito, M., Sakai, Y., Nishida, T., and Noce, T. (2000). Isolation of chicken vasa homolog gene and tracing the origin of primordial germ cells. *Development*, 127(12):2741–2750.
- Turbeville, J. M., Schulz, J. R., and Raff, R. A. (1994). Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology. *Molecular Biology and Evolution*, 11(4):648–655.
- Uchiyama, M., Kumano, T., Konno, N., Yoshizawa, H., and Matsuda, K. (2011). Ontogeny of ENaC expression in the gills and the kidneys of the japanese black salamander (*Hynobius nigrescens* Stejneger). *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 316B(2):135–145.
- Venkatarama, T., Lai, F., Luo, X., Zhou, Y., Newman, K., et al. (2010). Repression of zygotic gene expression in the xenopus germline. *Development*, 137:651–660.
- Venkatesh, B., Lee, A. P., Ravi, V., Maurya, A. K., Lian, M. M., et al. (2014). Elephant shark genome provides unique insights into gnathostome evolution. *Nature*, 505:174–179.
- Vidal, N. and Hedges, S. B. (2009). The molecular evolutionary tree of lizards, snakes, and amphisbaenians. *Comptes Rendus Biologies*, 332(2-3):129–139.
- Villiard, E., Brinkmann, H., Moiseeva, O., Mallette, F. A., Ferbeyre, G., et al. (2007). Urodele p53 tolerates amino acid changes found in p53 variants linked to human cancer. *BMC Evolutionary Biology*, 7:180.
- Vincent, S. D., Dunn, N. R., Sciammas, R., Shapiro-Shalef, M., Davis, M. M., et al. (2005). The zinc finger transcriptional repressor Blimp1/Prdm1 is dispensable for early axis formation but is required for specification of primordial germ cells in the mouse. *Development*, 132:1315–1325.
- Vinogradov, A. E. (1998). Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationships. *Cytometry*, 31:100–109.
- Vonk, F. J., Casewell, N. R., Henkel, C. V., Heimberg, A. M., Jansen, H. J., et al. (2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proceedings of the National Academy of Sciences of the United States of America*, 110(51):20651–20656.

- Wada, H. and Satoh, N. (1994). Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18s rDNA. *Proceedings of the National Academy of Sciences of the United States of America*, 91(5):1801–1804.
- Walthers, E. A. and Moore, F. L. (2005). Cloning proenkephalin from the brain of a urodele amphibian (*Taricha granulosa*) using a DOR-specific primer in a 3'RACE reaction. *General and Comparative Endocrinology*, 142(3):364–370.
- Wang, J., Levasseur, D. N., and Orkin, S. H. (2008). Requirement of Nanog dimerization for stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences of the United States of America*, 105(17):6326–6331.
- Wang, S.-H., Tsai, M.-S., Chiang, M.-F., and Li, H. (2003). A novel NK-type homeobox gene, ENK (early embryo specific NK), preferentially expressed in embryonic stem cells. *Gene Expression Patterns*, 3(1):99–103.
- Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., et al. (2013). The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plans. *Nature Genetics*, 45(6):701–706.
- Weidinger, G., Wolke, U., Kopranner, M., Klinger, M., and Raz, E. (1999). Identification of tissues and patterning events required for distinct steps in early migration of zebrafish primordial germ cells. *Development*, 126:5295–5307.
- Welch, J. J., Bininda-Emonds, O. R., and Bromham, L. (2008). Correlates of substitution rate variation in mammalian protein-coding sequences. *BMC Evolutionary Biology*, 8(53).
- Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., et al. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature*, 448:318–324.
- West, J. A., Viswanathan, S. R., Yabuuchi, A., Cunniff, K., Takeuchi, A., et al. (2009). A role for Lin28 in primordial germ-cell development and germ-cell malignancy. *Nature*, 460:909–913.
- Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–699.
- Whitfield, T., Heasman, J., and Wylie, C. (1993). XLPOU-60, a xenopus POU-domain mRNA, is oocyte-specific from very early stages of oogenesis, and localised to presumptive mesoderm and ectoderm in the blastula. *Developmental Biology*, 155(2):361–370.
- Whitfield, T. T., Heasman, J., and Wylie, C. C. (1995). Early embryonic expression of XLPOU-60, a xenopus POU-domain proteins. *Developmental Biology*, 169(2):759–769.
- Whittington, P. M. and Dixon, K. E. (1975). Quantitative studies of germ plasm and germ cells during early embryogenesis of xenopus laevis. *Journal of Embryology and Experimental Morphology*, 33(1):57–74.
- Wiens, J. J., Hutter, C. R., Mulcahy, D. G., Noonan, B. P., Townsend, T. M., et al. (2012). Resolving the phylogeny of lizards and snakes (squamata) with extensive sampling of genes and species. *Biology Letters*, 8(6):1043–1046.
- Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7273–7280.
- Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713.
- Wylie, C. C. and Heasman, J. (1976). The formation of the gonadal ridge in *Xenopus laevis* I. a light and transmission electron microscope study. *Journal of Embryology and Experimental Morphology*, 35(1):125–138.
- Yamaguchi, T., Taguchi, A., Watanabe, K., and Orii, H. (2013). DEADSouth protein localizes to germ plasm and is required for the development of primordial germ cells in *Xenopus laevis*. *Biology Open*, 2:191–199.

- Yamaji, M., Seki, Y., Kurimoto, K., Yabuta, Y., Yuasa, M., et al. (2008). Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nature Genetics*, 40:1016–1022.
- Yang, W., Qi, Y., Bi, K., and Fu, J. (2012). Toward understanding the genetic basis of adaptation to high-elevation life in poikilothermic species: A comparative transcriptomic analysis of two ranid frogs, *Rana chensinensis* and *R. kukunoris*. *BMC Genomics*, 13:588.
- Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.
- Yeom, Y. I., Fuhrmann, G., Ovitt, C. E., Brehm, A., Ohbo, K., et al. (1996). Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development*, 122:881–894.
- Yoon, C., Kawakami, K., and Hopkins, N. (1997). Zebrafish vasa homologue RNA is localized to the cleavage planes of 2- and 4-cell-stage embryos and is expressed in the primordial germ cells. *Development*, 124:3157–3166.
- Yu, F., Li, J., Chen, H., Fu, J., Ray, S., et al. (2011). Krüppel-like factor 4 (KLF4) is required for maintenance of breast cancer stem cells and for cell migration and invasion. *Oncogene*, 30:2161–2172.
- Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920.
- Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995). Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes & Development*, 9:2635–2645.
- Yusuf, D., Butland, S. L., Swanson, M. I., Bolotin, E., Ticoll, A., et al. (2012). The transcription factor encyclopedia. *Genome Biology*, 13:R24.
- Zardoya, R. and Meyer, A. (1996). Evolutionary relationships of the coelacanth, lungfishes, and tetrapods based on the 28S ribosomal RNA genes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(11):5449–5454.
- Zardoya, R. and Meyer, A. (2001). On the origin of and phylogenetic relationships among living amphibians. *Proceedings of the National Academy of Sciences of the United States of America*, 98(13):7380–7383.
- Zelazowska, M., Kilarski, W., Bilinski, S. M., Podder, D. D., and Kloc, M. (2007). Balbiani cytoplasm in oocytes of a primitive fish, the sturgeon *Acipenser gueldenstaedtii*, and its potential homology to the balbiani body (mitochondrial cloud) of *Xenopus laevis* oocytes. *Cell and Tissue Research*, 329(1):137–145.
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18(6):292–298.
- Zhang, J. and King, M. (1996). *Xenopus* VegT RNA is localized to the vegetal cortex during oogenesis and encodes a novel T-box transcription factor involved in mesodermal patterning. *Development*, 122:4119–4129.
- Zhang, P., Zhou, H., Chen, Y.-Q., Liu, Y.-F., and Qu, L.-H. (2005). Mitogenomic perspectives on the origin and phylogeny of living amphibians. *Systematic Biology*, 54(3):391–400.
- Zhang, X., Ma, Y., Liu, X., Zhou, Q., and Wang, X.-J. (2013). Evolutionary and functional analysis of the key pluripotency factor Oct4 and its family proteins. *Journal of Genetics and Genomics*, 40(8):399–412.
- Zharkikh, A. and Li, W.-H. (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. four taxa with a molecular clock. *Molecular Biology and Evolution*, 9(6):1119–1147.
- Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L., and Kuehn, M. R. (1993). Nodal is a novel TGF- β -like gene expressed in the mouse node during gastrulation. *Nature*, 361:543–547.

Zhou, Y. and King, M. L. (1996). Localization of Xcat-2 RNA, a putative germ plasm component, to the mitochondrial cloud in xenopus stage I oocytes. *Development*, 122:2947–2953.

Appendices

APPENDIX A

Additional Tables

Table A.1: Amphibian Gene Trees from Literature Review

Gene	Tree Topology	Method and Comments	Support (%)	Reference
AID	Mammal-Urodele	NJ	-	(Bascove and Frippiat, 2010)
Brachyury	Mammal-Urodele	Higgins-Sharp algorithm	96.4	(Sone et al., 1999)
c-Myc	Mammal-Anuran	UPGMA	-	(Maki et al., 2009)
CCN1	Mammal-Anuran	No methods printed	-	(Looso et al., 2012)
CCN2	Species Phylogeny	No methods printed	-	(Looso et al., 2012)
CCN3	Species Phylogeny	No methods printed	-	(Looso et al., 2012)
CCN4	Species Phylogeny	No methods printed	-	(Looso et al., 2012)
C μ 4 (IgM)	Species Phylogeny	TREEALIGN software	-	(Fellah et al., 1992)
Dazl	Mammal-Urodele	No methods printed, diagram only	-	(Johnson et al., 2003a)
DNase1	Species Phylogeny	NJ	-	(Takeshita et al., 2001)
EnaC α	Species Phylogeny	NJ (100 boot reps)	100	(Uchiyama et al., 2011)
Eomes	Species Phylogeny	Higgins-Sharp algorithm	96.1	(Sone et al., 1999)
ER α	Species Phylogeny	NJ (JTT, 1000 boot reps)	74	(Katsu et al., 2006)
ER α	Species Phylogeny	NJ (500 boot reps)	100	(Ko et al., 2008)
Eya4	Mammal-Anuran	Bayesian	84	(Modrell and Baker, 2012)
fgf8	Species Phylogeny	MegAlign software	-	(Han et al., 2001)
fgn	Species Phylogeny	MP (100 bootstrap replicates)	69	(Cadinouche et al., 1999)
FSH	Mammal-Urodele	unweighted pair-group method with arithmetic mean	-	(Saito et al., 2002)

Table A.1 – continued from previous page

Gene	Tree Topology	Method and Comments	Support (%)	Reference
GAD(65)	Species Phylogeny	ML (JTT)	66	(Lariviere et al., 2002)
GAPDH	Mammal-Urodele	NJ/ML (1000 boot reps/quartet puzzling)	59/69	(Mounaji et al., 2002)
gp130	Species Phylogeny	ME (poisson, 5000 boot reps)	<70	(Kiemnec-Tyburczy et al., 2011)
IGHM	Species Phylogeny	NJ (1000 boot reps)	83	(Schaerlinger and Frippiat, 2008)
Ihh	Mammal-Urodele	PileUp software	-	(Stark et al., 1998)
Klf4	Mammal-Anuran	UPGMA	-	(Maki et al., 2009)
Leptin	Species Phylogeny	non-synonymous substitutions, unweighted pair group method	-	(Boswell et al., 2006)
LH	Species Phylogeny	unweighted pair-group method with arithmetic mean	-	(Saito et al., 2002)
MHC-I	Mammal-Urodele	NJ (1000 boot reps)	87.3	(Sammut et al., 1999)
MHC-II β 1	Mammal-Anuran	NJ (1000 boot reps)	59.3	(Laurens et al., 2001)
MHC-II β 2	Species Phylogeny	NJ (1000 boot reps)	50.2	(Laurens et al., 2001)
Mix	Mammal-Urodele	NJ (JTT)	<50	(Swiers et al., 2010)
MMP13	Species Phylogeny	UPGMA	-	(Miyazaki et al., 1996)
Msi1	Mammal-Anuran	NJ	-	(Susaki et al., 2009)
Msx-1	Species Phylogeny	PileUp software	-	(Crews et al., 1995)
Msx-1	Species Phylogeny	UPGMA	-	(Koshiba et al., 1998)
Msx-1	Mammal-Anuran	J.Hein method with structural residue weight table	-	(Simon et al., 1995)
MTR	Species Phylogeny	NJ (1000 boot reps)	100	(Searcy et al., 2011)
NHE1	Mammal-Urodele	MP (1000 boot reps)	68	(McLean et al., 1999)
Nodal-2	Mammal-Urodele	NJ (JTT)	51	(Swiers et al., 2010)
Oct4	Mammal-Urodele	No methods printed, diagram only	-	(Johnson et al., 2003a)
Oct4	Species Phylogeny	UPGMA	-	(Maki et al., 2009)
PENK	Species Phylogeny	MP	93	(Walthers and Moore, 2005)
POMC	Species Phylogeny	MP	82	(Walthers and Moore, 2005)
DOR	Species Phylogeny	Bayesian	75	(Bradford et al., 2006)
DOR	Species Phylogeny	MP	91	(Bradford et al., 2005)
MOR	Species Phylogeny	Bayesian	100	(Bradford et al., 2006)
MOR	Species Phylogeny	MP	93	(Bradford et al., 2005)

Table A.1 – continued from previous page

Gene	Tree Topology	Method and Comments	Support (%)	Reference
ORL	Species Phylogeny	Bayesian	83	(Bradford et al., 2006)
ORL	Species Phylogeny	MP	57	(Bradford et al., 2005)
Otx2	Mammal-Urodele	ML (JTT, 2000 boot reps)	-	(Sauka-Spengler et al., 2002)
Otx5	Species Phylogeny	ML (JTT, 2000 boot reps)	-	(Sauka-Spengler et al., 2002)
p53	Species Phylogeny	ML (WAG+G8, 100 boot reps)	64	(Villiard et al., 2007)
RhAG	Species Phylogeny	Bayesian, 1st and 3rd positions converted to R/Y, 2nst+4G+I for 1st and 3rd, GTR+4G+I for 2nd.	100	(Huang and Peng, 2005)
Shh	Mammal-Urodele	PileUp software	-	(Stark et al., 1998)
skeletal myosin	Mammal-Urodele	Higgins-Sharp algorithm	61.4	(Simon and Oppenheimer, 1996)
Sox2	Mammal-Anuran	UPGMA	-	(Maki et al., 2009)
Tbx2	Mammal-Urodele	Higgins-Sharp algorithm	93.6	(Sone et al., 1999)
Tbx6	Mammal-Urodele	Higgins-Sharp algorithm	60.2	(Sone et al., 1999)
TDT	Mammal-Urodele	MP and Bayesian (GTR+G+I)	70	(Golub et al., 2004)
Thr- α	Mammal-Urodele	NJ (1000 boot reps)	26	(Safi et al., 1997)
Thr- β	Mammal-Anuran	NJ (1000 boot reps)	52	(Safi et al., 1997)
V1 VTR	Mammal-Urodele	NJ. Very small branch length	-	(Hasunuma et al., 2007)
V2 VTR	Species Phylogeny	NJ	-	(Hasunuma et al., 2007)
V2 VTR	Species Phylogeny	NJ (1000 boot reps)	100	(Searcy et al., 2011)
Vasa	Mammal-Urodele	No methods printed, diagram only	-	(Johnson et al., 2003a)

Table A.2: Query Sequence Dataset. The number of ESTs, mRNAs and cDNAs downloaded and after processing are shown in the accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.3: Trees with maximum distance >999. The rates are shown for each tree with a maximum corrected distance greater than 999. The G-T rate in all cases is set to 1 and therefore not shown. The Seq No. column refers to the arbitrary identifier assigned to each gene, this value is used to extract data from the MySQL database (Section 2.1).

Seq No.	Rates				
	A-C	A-G	A-T	C-G	C-T
5630	1.0000	27.5426	1.0000	1.0000	19630.4668
6143	1.0000	4.5670	1.0000	1.0000	60554.2773
6936	1.0000	6.5646	1.0000	1.0000	41180.2266
7112	1.0000	1.0022	0.0001	0.0001	32339.4746
8639	1.0000	79.0817	1.0000	1.0000	7455.5898
9213	1.0000	4.3720	1.0000	1.0000	69718.2109
9941	1.0000	10.0223	1.0000	1.0000	37.6817
11619	3.3×10^7	148.9520	563.9587	2987.5310	148.9520
180040	1.0000	3.1314	1.0000	1.0000	24667.8301
180482	1.0000	27.4805	1.0000	1.0000	6559.0132
181004	1.0000	5.0×10^{12}	1.6×10^8	1.6×10^8	9.0×10^7
998386	1.0000	30559.2656	1.0000	1.0000	1795.4271
1001687	1.0000	31443.2832	1.0000	1.0000	4.1350
1001896	1.0000	12346.6426	1.0000	1.0000	63.3616
1003657	1.0000	7.1326	1.0000	1.0000	82557.1797
1004296	1.0000	17153.7129	0.0001	0.0001	0.6897
1004322	1.0000	1.4346	1.0000	1.0000	42956.1992
1237710	1.0000	1.9827	1.0000	1.0000	33865.7461
1241662	1.0000	4149.3135	1.0000	1.0000	20343.7422
1254139	1.0000	1.9827	1.0000	1.0000	33865.7461
1255972	1.0000	7.7704	1.0000	1.0000	122989.9688
1255976	1.0000	3.9242	1.0000	1.0000	50701.6016
1295337	1.0000	10.6857	1.0000	1.0000	14275.1943
1306844	1.6×10^4	12550.3213	0.3699	0.2124	1.5×10^8
1307019	1.0000	1.3147	0.0001	0.0001	22427.3984
1309425	1.0000	7.0183	1.0000	1.0000	110662.9688
1309508	1.0000	10.7987	1.0000	1.0000	55477.2344
1313719	1.0000	5.1689	1.0000	1.0000	57405.7422
1339987	1.0000	10.7987	1.0000	1.0000	55477.2344

Table A.4: Amphibian results with refined quality parameters.

Species	Total	Species Phylogeny	Mammal- Urodele	Mammal- Anuran
<i>Rana catesbeiana</i>	539	82	147	25
<i>Rana chensinensis</i>	110	22	25	4
<i>Rana pirica</i>	23	2	8	2
<i>Xenopus laevis</i>	2070	370	532	127
<i>Xenopus tropicalis</i>	2655	523	669	169
<i>Ambystoma mexicanum</i>	1130	215	197	93
<i>Ambystoma tigrinum</i>	1063	206	190	103
<i>Andrias davidianus</i>	25	6	5	3
<i>Cynops pyrrhogaster</i>	1259	260	236	120
<i>Desmognathus ocoee</i>	2	0	0	1
<i>Notophthalmus viridescens</i>	245	59	40	31
<i>Pleurodeles waltl</i>	59	14	14	6
Total	9180	1759	2063	684

Table A.5: Actinopterygian four-taxon tree results significant by bootstrapping. The number of trees including and excluding the transcriptome are shown in the accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.6: Actinopterygian four-taxon tree results significant by the SH test. The number of trees including and excluding the transcriptome are shown in the accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.7: Archosaur and Testudine bootstrap results.

Species	Total	Species Phylogeny	Mammal- Croc/Turtle	Mammal- Bird
<i>Anas platyrhynchos</i>	43	26	1	1
<i>Colinus virginianus</i>	2	1	0	0
<i>Columba livia</i>	6	3	0	0
<i>Gallus gallus</i>	267	124	10	21
<i>Lagopus lagopus scotica</i>	36	17	0	6
<i>Lonchura striata domestica</i>	42	21	4	1
<i>Lophonetta specularoides</i>	14	4	1	0
<i>Meleagris gallopavo</i>	227	102	8	19
<i>Taeniopygia guttata</i>	247	123	14	14
<i>Alligator mississippiensis</i>	140	70	8	7
<i>Trachemys scripta</i>	58	30	2	3
Total	1082	521	48	72

Table A.8: Archosaur and Testudine SH test results.

Species	Total	Species Phylogeny	Mammal-Croc/Turtle	Mammal-Bird
<i>Anas platyrhynchos</i>	43	5	0	0
<i>Colinus virginianus</i>	2	0	0	0
<i>Columba livia</i>	6	0	0	0
<i>Gallus gallus</i>	267	28	0	2
<i>Lagopus lagopus scotica</i>	36	4	0	0
<i>Lonchura striata domestica</i>	42	4	1	0
<i>Lophonetta specularoides</i>	14	0	0	0
<i>Meleagris gallopavo</i>	227	20	0	2
<i>Taeniopygia guttata</i>	247	19	1	1
<i>Alligator mississippiensis</i>	140	12	0	1
<i>Trachemys scripta</i>	58	4	0	1
Total	1082	96	2	7

Table A.9: Lepidosaur Bootstrap results.

Species	Total	Species Phylogeny	Mammal-Lizard	Mammal-Snake
<i>Anolis carolinensis</i>	492	367	7	7
<i>Anolis sagrei</i>	9	5	0	0
<i>Gekko japonicus</i>	64	43	2	2
<i>Bothrops alternatus</i>	53	42	0	0
<i>Bothrops atrox</i>	15	11	0	0
<i>Bothrops insularis</i>	5	4	0	0
<i>Bothrops jararaca</i>	24	12	1	3
<i>Bungarus multicinctus</i>	36	23	1	0
<i>Deinagkistrodon acutus</i>	103	79	4	0
<i>Echis carinatus sochureki</i>	47	33	0	0
<i>Echis coloratus</i>	45	36	0	0
<i>Echis ocellatus</i>	42	33	0	1
<i>Echis pyramidum leakeyi</i>	26	17	1	1
<i>Lachesis muta</i>	58	46	1	2
<i>Micrurus corallinus</i>	52	35	0	3
<i>Naja atra</i>	36	23	0	1
<i>Philodryas olfersii</i>	42	30	2	1
<i>Rhabdophis tigrinus</i>	20	13	0	1
<i>Sistrurus catenatus edwardsi</i>	18	12	1	1
Total	1187	864	20	23

Table A.10: Lepidosaur SH test results.

Species	Total	Species Phylogeny	Mammal-Lizard	Mammal-Snake
<i>Anolis carolinensis</i>	492	138	1	0
<i>Anolis sagrei</i>	9	2	0	0
<i>Gekko japonicus</i>	64	16	1	0
<i>Bothrops alternatus</i>	53	17	0	0
<i>Bothrops atrox</i>	15	4	0	0
<i>Bothrops insularis</i>	5	0	0	0
<i>Bothrops jararaca</i>	24	4	0	0
<i>Bungarus multicinctus</i>	36	4	0	0
<i>Deinagkistrodon acutus</i>	103	20	0	0
<i>Echis carinatus sochureki</i>	47	15	0	0
<i>Echis coloratus</i>	45	16	0	0
<i>Echis ocellatus</i>	42	15	0	0
<i>Echis pyramidum leakeyi</i>	26	6	0	0
<i>Lachesis muta</i>	58	12	0	0
<i>Micrurus corallinus</i>	52	12	0	0
<i>Naja atra</i>	36	8	0	0
<i>Philodryas olfersii</i>	42	7	0	1
<i>Rhabdophis tigrinus</i>	20	3	0	0
<i>Sistrurus catenatus edwardsi</i>	18	4	0	0
Total	1187	303	2	1

Table A.11: Amphibian RRT results.

Species	Total	Significant	Significantly	
			Slower	Faster
<i>Rana catesbeiana</i>	1450	429	55	374
<i>Rana chensinensis</i>	136	36	9	27
<i>Rana pirica</i>	43	19	0	19
<i>Xenopus laevis</i>	6482	2048	222	1826
<i>Xenopus tropicalis</i>	8677	2850	340	2510
<i>Ambystoma mexicanum</i>	5505	2056	1893	163
<i>Ambystoma tigrinum</i>	1435	340	200	140
<i>Andrias davidianus</i>	31	7	4	3
<i>Cynops pyrrhogaster</i>	1738	472	362	110
<i>Desmognathus ocoee</i>	2	0	0	0
<i>Notophthalmus viridescens</i>	288	61	30	31
<i>Pleurodeles waltl</i>	79	23	13	10

Table A.12: Actinopterygian relative rate test results. The table is on the accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.13: Archosaur and Testudine RRT results.

Species	Total	Significant	Significantly	
			Slower	Faster
<i>Anas platyrhynchos</i>	73	20	7	13
<i>Carpodacus mexicanus</i>	1	0	0	0
<i>Colinus virginianus</i>	2	0	0	0
<i>Columba livia</i>	12	3	1	2
<i>Gallus gallus</i>	422	95	35	60
<i>Lagopus lagopus scotica</i>	67	11	4	7
<i>Lonchura striata domestica</i>	74	23	5	18
<i>Lophonetta specularoides</i>	26	2	0	2
<i>Meleagris gallopavo</i>	386	75	29	46
<i>Taeniopygia guttata</i>	409	100	31	69
<i>Alligator mississippiensis</i>	236	48	29	19
<i>Trachemys scripta</i>	87	22	19	3

Table A.14: Lepidosaur RRT results.

Species	Total	Significant	Significantly	
			Slower	Faster
<i>Anolis carolinensis</i>	812	189	153	36
<i>Anolis sagrei</i>	11	1	0	1
<i>Gekko japonicus</i>	111	25	19	6
<i>Bothrops alternatus</i>	95	32	5	27
<i>Bothrops atrox</i>	23	9	0	9
<i>Bothrops insularis</i>	14	2	1	1
<i>Bothrops jararaca</i>	43	10	2	8
<i>Bungarus multicinctus</i>	64	13	1	12
<i>Deinagkistrodon acutus</i>	190	44	9	35
<i>Echis carinatus sochureki</i>	71	14	3	11
<i>Echis coloratus</i>	94	12	3	9
<i>Echis ocellatus</i>	78	12	4	8
<i>Echis pyramidum leakeyi</i>	45	7	2	5
<i>Lachesis muta</i>	108	14	3	11
<i>Micrurus corallinus</i>	100	24	10	14
<i>Naja atra</i>	62	12	5	7
<i>Philodryas olfersii</i>	79	17	4	13
<i>Rhabdophis tigrinus</i>	41	10	2	8
<i>Sistrurus catenatus edwardsi</i>	30	10	1	9

Table A.15: Amphibian bootstrapped trees using either Teleost or Sturgeon outgroup. The Amphibian tree topologies significant by bootstrapping are shown according to the relative rate test result (n.s. = not significant, N/A = result not available).

Faster species by RRT			Teleost Outgroup			Sturgeon outgroup		
Amphibian	Actinopterygian	Total	Species Phylogeny	Mammal-Urodele	Mammal-Anuran	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
n.s.	n.s.	3471	953	430	228	855	340	364
n.s.	Teleost	2753	638	409	222	949	212	188
n.s.	Sturgeon	47	16	5	4	15	2	7
n.s.	N/A	1682	395	228	154	488	164	140
Anuran	n.s.	1090	276	276	37	284	194	60
Anuran	Teleost	1577	395	345	42	571	196	56
Anuran	Sturgeon	14	3	3	0	3	5	0
Anuran	N/A	646	170	148	17	229	108	18
Urodele	n.s.	180	43	8	35	42	8	53
Urodele	Teleost	123	31	11	14	41	11	10
Urodele	Sturgeon	5	0	1	2	0	0	3
Urodele	N/A	101	24	12	20	24	9	19

Table A.16: Amphibian SH-test trees using either Teleost or Sturgeon outgroup. The Amphibian tree topologies significant by the SH-test are shown according to the relative rate test result (n.s. = not significant, N/A = result not available).

Faster species by RRT			Teleost Outgroup			Sturgeon outgroup		
Amphibian	Actinopterygian	Total	Species Phylogeny	Mammal-Urodele	Mammal-Anuran	Species Phylogeny	Mammal-Urodele	Mammal-Anuran
n.s.	n.s.	3471	117	13	6	86	13	12
n.s.	Teleost	2753	69	15	4	116	5	6
n.s.	Sturgeon	47	1	1	0	1	0	0
n.s.	N/A	1682	56	11	3	65	5	5
Anuran	n.s.	1090	38	21	4	25	18	5
Anuran	Teleost	1577	50	34	0	98	7	2
Anuran	Sturgeon	14	0	0	0	0	0	0
Anuran	N/A	646	28	15	0	32	4	0
Urodele	n.s.	180	3	0	9	4	0	10
Urodele	Teleost	123	4	2	1	6	0	2
Urodele	Sturgeon	5	0	0	0	0	0	3
Urodele	N/A	101	3	0	6	1	0	6

Table A.17: Mouse Gene Ontology terms over-represented in Mammal-Epigenesis bootstrapped trees compared to Species Phylogeny trees. Table can be found on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.18: Mouse Gene Ontology terms over-represented in Preformation-faster genes compared to genes with no significant difference in the rate of evolution. Table is in accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.19: Mouse Gene Ontology terms over-represented in genes with no significant difference in rate of evolution compared to those with a faster rate in the taxa undergoing preformation. Table is located in accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.20: Zebrafish Gene Ontology terms over-represented in Mammal-Epigenesis bootstrapped trees compared to Species Phylogeny trees. Table is shown on accompanying disk and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.21: Zebrafish Gene Ontology terms over-represented in Preformation-faster genes compared to genes with no significant difference in the rate of evolution. Table is on the accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.22: Zebrafish Gene Ontology terms over-represented in genes with no significant difference in rate of evolution compared to those with a faster rate in the taxa undergoing preformation. Table is located on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.23: Preformation-faster mouse gene expression locations. For each location the proportion of Preformation-faster genes was compared to the average (73.0242%) using the Chi-squared test (1d.f.). Results are sorted according to the Chi-squared p-value. Only those locations with >20 total genes are shown. Table is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.24: Bootstrapped Mammal-Epigenesis mouse gene expression locations. For each location the proportion of Mammal-Epigenesis genes were compared to the average (37.4783%) using the Chi-squared test (1d.f.). Results are sorted according to the Chi-squared p-value. Only locations with >20 total genes are shown. Table is included on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.25: SH-test Mammal-Epigenesis mouse gene expression locations. For each location the proportion of Mammal-Epigenesis genes were compared to the average (18.9453%) using the Chi-squared test (1d.f.). Results are sorted according to the Chi-squared p-value. Only locations with >20 total genes are shown. Table is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.26: Preformation-faster zebrafish gene expression locations. For each location the proportion of Preformation-faster genes was compared to the average (71.0142%) using the Chi-squared test (1d.f.). Results are sorted according to the Chi-squared p-value. Only those locations with >20 total genes are shown. Table in on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.27: Bootstrapped Mammal-Epigenesis zebrafish gene expression locations. For each location the proportion of Mammal-Epigenesis genes were compared to the average (37.8832%) using the Chi-squared test (1d.f.). Results are sorted according to the Chi-squared p-value. Only locations with >20 total genes are shown. Table is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.28: SH-test Mammal-Epigenesis zebrafish gene expression locations. For each location the proportion of Mammal-Epigenesis genes were compared to the average (17.1974%) using the Chi-squared test (1d.f.). Results are sorted according to the Chi-squared p-value. Only locations with >20 total genes are shown. Table is included in accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.29: Expanded Amphibian bootstrap results.

Species	Total	Species Phylogeny	Mammal- Urodele	Mammal- Anuran
<i>Rana catesbeiana</i>	894	209	151	56
<i>Rana chensinensis</i>	3856	954	700	194
<i>Rana kukunoris</i>	4187	1070	741	211
<i>Rana pirica</i>	28	6	6	3
<i>Xenopus laevis</i>	4318	1220	680	261
<i>Xenopus tropicalis</i>	6169	1873	921	349
<i>Ambystoma mexicanum</i>	4958	1447	626	302
<i>Ambystoma tigrinum</i>	928	192	133	71
<i>Andrias davidianus</i>	19	5	3	0
<i>Cynops pyrrhogaster</i>	1119	282	194	104
<i>Desmognathus ocoee</i>	2	0	0	1
<i>Notophthalmus viridescens</i>	5759	1804	733	373
<i>Pleurodeles waltl</i>	54	15	5	3
Total	32 291	9077	4893	1928

Table A.30: Expanded Amphibian SH-test results.

Species	Total	Species Phylogeny	Mammal- Urodele	Mammal- Anuran
<i>Rana catesbeiana</i>	894	24	12	1
<i>Rana chensinensis</i>	3856	104	46	11
<i>Rana kukunoris</i>	4187	123	53	8
<i>Rana pirica</i>	28	0	0	0
<i>Xenopus laevis</i>	4318	177	47	6
<i>Xenopus tropicalis</i>	6169	312	57	10
<i>Ambystoma mexicanum</i>	4958	201	44	17
<i>Ambystoma tigrinum</i>	928	15	7	3
<i>Andrias davidianus</i>	19	1	0	0
<i>Cynops pyrrhogaster</i>	1119	28	11	6
<i>Desmognathus ocoee</i>	2	0	0	1
<i>Notophthalmus viridescens</i>	5759	263	58	24
<i>Pleurodeles waltl</i>	54	3	0	0
Total	32 291	1251	335	87

Table A.31: Expanded Actinopterygian bootstrap results. The number of trees are shown in the accompanying disk and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.**Table A.32: Expanded Actinopterygian SH-test results.** Table is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.**Table A.33: Expanded Archosaur and Testudine bootstrap results.**

Species	Total	Species Phylogeny	Mammal- Croc/Turtle	Mammal- Bird
<i>Anas platyrhynchos</i>	216	87	10	17
<i>Carpodacus mexicanus</i>	1	1	0	0
<i>Colinus virginianus</i>	10	5	0	0
<i>Columba livia</i>	42	24	0	2
<i>Gallus gallus</i>	7065	3791	293	413
<i>Lagopus lagopus scotica</i>	327	139	26	17
<i>Lonchura striata domestica</i>	974	427	59	44
<i>Lophonetta specularoides</i>	74	37	2	1
<i>Meleagris gallopavo</i>	6556	3554	271	362
<i>Taeniopygia guttata</i>	6747	3463	290	410
<i>Alligator mississippiensis</i>	148	76	8	10
<i>Chrysemys picta bellii</i>	6878	3721	275	362
<i>Pelodiscus sinensis</i>	6758	3593	295	349
<i>Trachemys scripta</i>	60	32	4	3
Total	35 856	18 950	1533	1990

Table A.34: Expanded Archosaur and Testudine SH-test results.

Species	Total	Species Phylogeny	Mammal-Croc/Turtle	Mammal-Bird
<i>Anas platyrhynchos</i>	216	17	0	1
<i>Carpodacus mexicanus</i>	1	0	0	0
<i>Colinus virginianus</i>	10	0	0	0
<i>Columba livia</i>	42	8	0	0
<i>Gallus gallus</i>	7065	1111	8	47
<i>Lagopus lagopus scotica</i>	327	20	0	1
<i>Lonchura striata domestica</i>	974	60	1	1
<i>Lophonetta specularoides</i>	74	5	0	0
<i>Meleagris gallopavo</i>	6556	1036	3	26
<i>Taeniopygia guttata</i>	6747	955	5	28
<i>Alligator mississippiensis</i>	148	11	0	1
<i>Chrysemys picta bellii</i>	6878	1091	7	32
<i>Pelodiscus sinensis</i>	6758	995	7	26
<i>Trachemys scripta</i>	60	4	0	1
Total	35 856	5313	31	164

Table A.35: Expanded Lepidosaur bootstrap results.

Species	Total	Species Phylogeny	Mammal-Epi.	Mammal-Pre.
<i>Anolis carolinensis</i>	6085	5171	68	48
<i>Anolis sagrei</i>	28	20	0	0
<i>Eublepharis macularius</i>	5673	4864	66	35
<i>Gekko japonicus</i>	375	255	29	6
<i>Bothrops alternatus</i>	59	43	1	1
<i>Bothrops atrox</i>	14	11	0	0
<i>Bothrops insularis</i>	8	5	0	1
<i>Bothrops jararaca</i>	28	16	0	1
<i>Bungarus multicinctus</i>	40	26	0	0
<i>Carlia rubrigularis</i>	5596	4755	44	45
<i>Deinagkistrodon acutus</i>	39	29	0	0
<i>Echis carinatus sochureki</i>	54	40	1	0
<i>Echis coloratus</i>	49	33	0	2
<i>Echis ocellatus</i>	46	34	1	2
<i>Echis pyramidum leakeyi</i>	30	20	1	1
<i>Lachesis muta</i>	70	47	3	1
<i>Lampropholis coggeri</i>	5151	4344	53	38
<i>Micrurus corallinus</i>	63	40	2	3
<i>Naja atra</i>	39	29	0	1
<i>Philodryas olfersii</i>	41	28	2	2
<i>Rhabdophis tigrinus</i>	20	13	0	1
<i>Saproscincus basiliscus</i>	5452	4591	62	48
<i>Sistrurus catenatus edwardsi</i>	22	15	0	2
Total	28 982	24 429	333	238

Table A.36: Expanded Lepidosaur SH-test results.

Species	Total	Species Phylogeny	Mammal- Epi.	Mammal- Pre.
<i>Anolis carolinensis</i>	6085	3204	6	0
<i>Anolis sagrei</i>	28	8	0	0
<i>Eublepharis macularius</i>	5673	2983	1	1
<i>Gekko japonicus</i>	375	88	19	0
<i>Bothrops alternatus</i>	59	15	0	0
<i>Bothrops atrox</i>	14	3	0	0
<i>Bothrops insularis</i>	8	1	0	0
<i>Bothrops jararaca</i>	28	3	0	0
<i>Bungarus multicinctus</i>	40	11	0	0
<i>Carlia rubrigularis</i>	5596	2927	0	1
<i>Deinagkistrodon acutus</i>	39	9	0	0
<i>Echis carinatus sochureki</i>	54	16	0	0
<i>Echis coloratus</i>	49	15	0	0
<i>Echis ocellatus</i>	46	16	0	0
<i>Echis pyramidum leakeyi</i>	30	6	0	0
<i>Lachesis muta</i>	70	18	0	0
<i>Lampropholis coggeri</i>	5151	2514	1	0
<i>Micrurus corallinus</i>	63	9	0	0
<i>Naja atra</i>	39	8	0	0
<i>Philodryas olfersii</i>	41	7	0	1
<i>Rhabdophis tigrinus</i>	20	5	0	0
<i>Saproscincus basiliscus</i>	5452	2796	1	3
<i>Sistrurus catenatus edwardsi</i>	22	7	0	0
Total	28 982	14 669	28	6

Table A.37: Expanded Amphibian RRT results.

Species	Total	Significant	Significantly	
			Slower	Faster
<i>Rana catesbeiana</i>	1360	403	29	374
<i>Rana chensinensis</i>	5809	1960	119	1841
<i>Rana kukunoris</i>	6167	2158	115	2043
<i>Rana pirica</i>	40	17	1	16
<i>Xenopus laevis</i>	6755	2155	140	2015
<i>Xenopus tropicalis</i>	9527	3243	273	2970
<i>Ambystoma mexicanum</i>	7505	2675	2454	221
<i>Ambystoma tigrinum</i>	1502	361	211	150
<i>Andrias davidianus</i>	30	6	3	3
<i>Cynops pyrrhogaster</i>	1792	490	376	114
<i>Desmognathus ocoee</i>	2	0	0	0
<i>Notophthalmus viridescens</i>	8854	3205	2973	232
<i>Pleurodeles waltl</i>	82	23	13	10

Table A.38: Expanded Actinopterygian relative rate test results. Table is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.39: Expanded Archosaur RRT results.

Species	Total	Significant	Significantly	
			Slower	Faster
<i>Anas platyrhynchos</i>	372	124	16	108
<i>Carpodacus mexicanus</i>	7	1	0	1
<i>Colinus virginianus</i>	14	3	0	3
<i>Columba livia</i>	64	22	3	19
<i>Gallus gallus</i>	10 069	3512	329	3183
<i>Lagopus lagopus scotica</i>	497	115	12	103
<i>Lonchura striata domestica</i>	1521	611	14	597
<i>Lophonetta specularoides</i>	147	32	3	29
<i>Meleagris gallopavo</i>	9372	3515	277	3238
<i>Taeniopygia guttata</i>	9511	3703	275	3428
<i>Alligator mississippiensis</i>	235	48	29	19
<i>Chrysemys picta bellii</i>	9739	3712	3487	225
<i>Pelodiscus sinensis</i>	9613	3023	2451	572
<i>Trachemys scripta</i>	87	22	19	3

Table A.40: Expanded Lepidosaur RRT results.

Species	Total	Significant	Significantly	
			Slower	Faster
<i>Anolis carolinensis</i>	8749	1291	490	801
<i>Anolis sagrei</i>	52	5	0	5
<i>Eublepharis macularius</i>	8178	1130	954	176
<i>Gekko japonicus</i>	567	102	69	33
<i>Bothrops alternatus</i>	97	33	1	32
<i>Bothrops atrox</i>	22	8	0	8
<i>Bothrops insularis</i>	18	2	0	2
<i>Bothrops jararaca</i>	42	9	0	9
<i>Bungarus multicinctus</i>	62	13	0	13
<i>Carlia rubrigularis</i>	8154	1046	377	669
<i>Deinagkistrodon acutus</i>	199	48	5	43
<i>Echis carinatus sochureki</i>	75	14	0	14
<i>Echis coloratus</i>	98	11	0	11
<i>Echis ocellatus</i>	83	13	2	11
<i>Echis pyramidum leakeyi</i>	49	9	1	8
<i>Lachesis muta</i>	110	13	1	12
<i>Lampropholis coggeri</i>	7449	947	306	641
<i>Micrurus corallinus</i>	103	22	7	15
<i>Naja atra</i>	68	14	1	13
<i>Philodryas olfersii</i>	81	12	2	10
<i>Rhabdophis tigrinus</i>	42	11	1	10
<i>Saprosincus basiliscus</i>	7815	1017	339	678
<i>Sistrurus catenatus edwardsi</i>	34	14	1	13

Table A.41: Lepidosaur RRT between each species results.
The table is shown on the accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.42: Amphibian RRT results between and within orders.

Species	Total	Between Orders			Within Orders		
		Significant	Slower	Faster	Significant	Slower	Faster
<i>Rana catesbeiana</i>	1091	303	23	280	92	30	62
<i>Rana chensinensis</i>	4564	1536	95	1441	144	66	78
<i>Rana kukunoris</i>	4741	1653	89	1564	166	76	90
<i>Rana pirica</i>	28	11	0	11	5	0	5
<i>Xenopus laevis</i>	4824	1550	100	1450	713	176	537
<i>Xenopus tropicalis</i>	5860	1942	158	1784	810	610	200
<i>Ambystoma mexicanum</i>	5658	2037	1885	152	631	390	241
<i>Ambystoma tigrinum</i>	1108	266	150	116	284	21	263
<i>Andrias davidianus</i>	21	3	1	2	3	0	3
<i>Cynops pyrrhogaster</i>	1262	339	268	71	151	53	98
<i>Desmognathus ocoee</i>	1	0	0	0	0	0	0
<i>Notophthalmus viridescens</i>	5969	2180	2033	147	544	283	261
<i>Pleurodeles waltl</i>	55	15	11	4	8	0	8

Table A.43: Expanded Actinopterygian RRT results between and within orders. Table is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

Table A.44: Archosaur RRT results between and within orders.

Species	Total	Between Orders			Within Orders		
		Significant	Slower	Faster	Significant	Slower	Faster
<i>Anas platyrhynchos</i>	314	106	15	91	83	19	64
<i>Carpodacus mexicanus</i>	5	1	0	1	1	0	1
<i>Colinus virginianus</i>	8	1	0	1	3	0	3
<i>Columba livia</i>	60	21	2	19	12	3	9
<i>Gallus gallus</i>	7871	2674	245	2429	1414	951	463
<i>Lagopus lagopus scotica</i>	377	86	10	76	52	6	46
<i>Lonchura striata domestica</i>	1218	501	12	489	397	15	382
<i>Lophonetta specularoides</i>	115	25	3	22	18	9	9
<i>Meleagris gallopavo</i>	7326	2681	201	2480	1392	419	973
<i>Taeniopygia guttata</i>	8112	3124	211	2913	1568	485	1083
<i>Alligator mississippiensis</i>	8	1	0	1	1	1	0
<i>Chrysemys picta bellii</i>	7917	3025	2868	157	1729	1505	224
<i>Pelodiscus sinensis</i>	7889	2390	1995	395	1674	245	1429
<i>Trachemys scripta</i>	76	18	16	2	8	3	5

Table A.45: Coelacanth, Lungfish and Shark bootstrap results. Type 1.

Species	Total	Species Phylogeny	Mammal- Epi.	Mammal- Pre.
<i>Latimeria chalumnae</i>	2944	1516	192	137
<i>Neoceratodus forsteri</i>	2188	1034	156	100
<i>Protopterus aethiopicus</i>	169	59	20	10
<i>Protopterus annectens</i>	2595	1223	146	147
<i>Chiloscyllium plagiosum</i>	36	1	10	0
<i>Leucoraja erinacea</i>	1049	159	265	90
<i>Scyliorhinus canicula</i>	913	127	233	84
<i>Squalus acanthias</i>	852	128	195	76
<i>Torpedo californica</i>	328	51	90	25

Table A.46: Coelacanth, Lungfish and Shark SH-test results. Type 1.

Species	Total	Species Phylogeny	Mammal- Epi.	Mammal- Pre.
<i>Latimeria chalumnae</i>	2944	473	10	13
<i>Neoceratodus forsteri</i>	2188	275	8	4
<i>Protopterus aethiopicus</i>	169	11	2	0
<i>Protopterus annectens</i>	2595	311	11	15
<i>Chiloscyllium plagiosum</i>	36	0	1	0
<i>Leucoraja erinacea</i>	1049	13	35	6
<i>Scyliorhinus canicula</i>	913	10	24	2
<i>Squalus acanthias</i>	852	9	18	3
<i>Torpedo californica</i>	328	1	5	0

Table A.47: Coelacanth, Lungfish and Shark bootstrap results. Type 2.

Species	Total	Species Phylogeny	Mammal- Epi.	Mammal- Pre.
<i>Latimeria chalumnae</i>	2526	1815	86	33
<i>Neoceratodus forsteri</i>	1884	1256	86	26
<i>Protopterus aethiopicus</i>	143	69	9	3
<i>Protopterus annectens</i>	2238	1457	89	49
<i>Chiloscyllium plagiosum</i>	34	8	5	4
<i>Leucoraja erinacea</i>	867	354	104	26
<i>Scyliorhinus canicula</i>	749	312	94	28
<i>Squalus acanthias</i>	684	266	88	21
<i>Torpedo californica</i>	259	95	32	12

Table A.48: Coelacanth, Lungfish and Shark SH-test results. Type 2.

Species	Total	Species Phylogeny	Mammal- Epi.	Mammal- Pre.
<i>Latimeria chalumnae</i>	2526	807	7	0
<i>Neoceratodus forsteri</i>	1884	499	6	2
<i>Protopterus aethiopicus</i>	143	22	1	0
<i>Protopterus annectens</i>	2238	565	4	2
<i>Chiloscyllium plagiosum</i>	34	1	0	0
<i>Leucoraja erinacea</i>	867	89	12	0
<i>Scyliorhinus canicula</i>	749	67	8	0
<i>Squalus acanthias</i>	684	48	5	1
<i>Torpedo californica</i>	259	10	4	0

Table A.49: Coelacanth, Lungfish and Shark RRT results.

Species	Compared against epigenesis spe.				Compared against preformation spe.			
	Total	Significant	Slower	Faster	Total	Significant	Slower	Faster
<i>Latimeria chalumnae</i>	1080	201	77	124	1232	243	118	125
<i>Neoceratodus forsteri</i>	991	127	113	14	1172	183	165	18
<i>Protopterus aethiopicus</i>	89	11	6	5	100	8	2	6
<i>Protopterus annectens</i>	1059	91	54	37	1235	122	98	24
<i>Chiloscyllium plagiosum</i>	33	5	1	4	37	6	0	6
<i>Leucoraja erinacea</i>	941	131	21	110	1075	150	64	86
<i>Scyliorhinus canicula</i>	827	94	23	71	952	109	59	50
<i>Squalus acanthias</i>	784	80	26	54	913	100	45	55
<i>Torpedo californica</i>	294	49	4	45	344	54	12	42

Table A.50: Number of gene copies for the syntenic region surrounding Sox2.

Species	SOX2	FXR1	CCDC39	TTC14	ATP11B	DCUN1D1	MCCCI	DNAJC19
Human	1	1	1	1	1	1	1	1
Pig	1	1	1	1	1	1	1	1
Mouse	1	1	1	1	1	1	1	1
Opossum	1	1	1	1	1	1	1	1
Tasmanian Devil	1	1	1	1	1	1	1	1
Platypus	1	1	1	1	1	1	1	1
Zebra Finch	1	1	1	1	1	1	1	1
Chicken	1	1	1	1	1	1	1	1
Turkey	0	1	1	1	1	1	1	0
Chinese Turtle	1	1	1	1	1	1	1	1
Painted Turtle	1	1	1	1	1	1	1	1
Anolis	1	1	1	1	1	1	1	1
Python	1	1	1	1	1	1	1	1
Xenopus tropicalis	1	1	1	1	1	2	1	1
Xenopus laevis	1	2	1	1	1	1	1	1
Axolotl	1	2	1	1	1	1	1	1
Lungfish	1	3	1	1	0	2	3	1
Coelacanth	1	1	1	1	1	1	1	1
Sturgeon	0	1	1	1	1	1	1	1
Spotted Gar	1	1	1	1	1	1	1	1
Zebrafish	1	1	1	1	1	1	1	1
Fugu	2	1	2	1	1	1	1	1
Tilapia	1	1	1	1	1	1	1	1
Stickleback	1	1	1	1	1	1	1	1
Medaka	1	1	1	1	1	1	1	1
Elephant Shark	1	1	1	1	1	1	1	1
Little Skate	1	2	1	1	1	1	1	1
Lamprey	1	1	0	1	1	0	0	1

Table A.51: Number of gene copies for the syntenic region surrounding KLF4.

Species	KLF4	RAD23B	ZNF462	TMEM38B
Human	1	1	2	1
Pig	1	1	1	1
Mouse	1	1	1	1
Opossum	1	1	1	1
Tasmanian Devil	1	1	1	1
Platypus	1	2	1	1
Zebra Finch	1	1	1	1
Chicken	2	1	1	1
Turkey	0	1	1	1
Chinese Turtle	1	1	1	1
Painted Turtle	1	1	1	1
Anolis	1	1	2	1
Python	1	1	1	1
Xenopus tropicalis	1	1	1	1
Xenopus laevis	2	1	1	2
Axolotl	2	1	1	1
Lungfish	2	3	5	2
Coelacanth	1	1	1	1
Sturgeon	1	2	1	0
Spotted Gar	1	1	1	1
Zebrafish	1	1	1	1
Fugu	1	1	1	1
Tilapia	1	1	1	1
Stickleback	1	2	1	1
Medaka	1	1	1	1
Elephant Shark	1	2	2	1
Little Skate	3	1	1	0
Lamprey	1	1	0	0

Table A.52: Number of gene copies for the syntenic region surrounding NANOG.

Species	AICDA	APOBEC1	NANOG	SLC2A3	FOXP2	C3AR1	NECAP1	PCK2	FEN1	TM9SF1	IPO4	PDCD6
Human	1	1	2	1	1	1	1	1	3	1	1	1
Pig	1	1	3	1	2	1	1	1	1	1	1	1
Mouse	1	1	1	1	1	1	1	1	1	1	1	1
Opossum	1	1	1	1	1	1	1	1	1	1	1	1
Tasmanian Devil	1	1	1	1	1	1	1	1	1	1	1	1
Platypus	1	1	1	2	3	1	0	1	1	2	0	1
Zebra Finch	1	1	2	1	1	1	1	0	1	0	0	1
Chicken	1	0	2	1	1	1	1	1	1	0	0	1
Turkey	1	0	2	1	2	1	1	0	1	0	0	1
Chinese Turtle	1	2	2	1	1	1	1	1	1	1	0	1
Painted Turtle	1	1	2	1	1	1	1	1	2	1	0	1
Anolis	1	1	1	1	1	1	1	1	1	1	1	1
Python	1	0	0	1	2	2	1	1	1	1	0	2
Xenopus tropicalis	1	0	0	1	1	1	2	2	1	1	1	1
Xenopus laevis	1	0	0	1	1	1	1	1	2	1	1	1
Axolotl	0	0	2	1	1	1	1	4	1	1	0	3
Lungfish	0	0	1	1	1	0	1	1	1	1	0	1
Coelacanth	2	0	2	1	1	3	1	1	2	2	1	1
Sturgeon	0	0	1	1	2	0	2	1	1	1	1	2
Spotted Gar	1	0	1	1	1	1	1	0	1	1	1	1
Zebrafish	1	0	1	2	1	1	1	1	1	1	1	1
Fugu	1	0	1	2	1	1	2	1	1	1	1	1
Tilapia	1	0	0	2	0	5	2	1	1	2	1	1
Stickleback	1	0	1	2	1	1	2	2	2	1	1	1
Medaka	1	0	1	2	1	0	2	1	1	1	1	1
Elephant Shark	1	0	0	0	0	1	0	1	4	0	0	1
Little Skate	0	0	0	1	0	1	0	1	1	1	0	1
Lamprey	0	0	0	0	1	0	0	1	2	1	0	1

APPENDIX B

Additional Figures

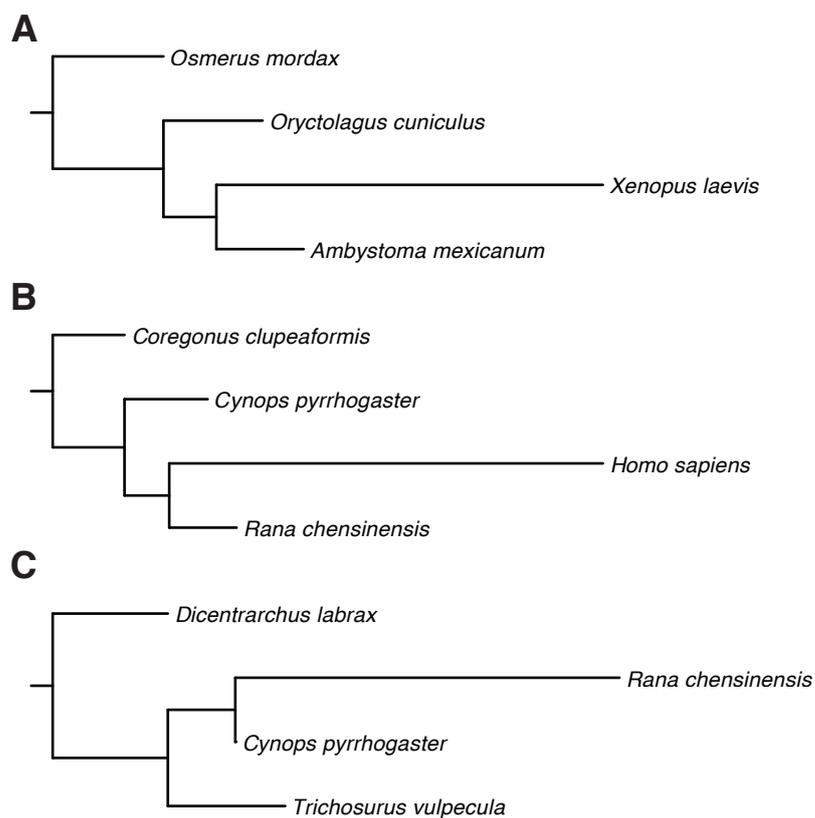


Figure B.1: Example trees with long branch lengths. Each of these trees features an excessively long branch length, producing an unreliable tree. The maximum distances were 5.79 (A), 4.91 (B) and 40.11 (C).

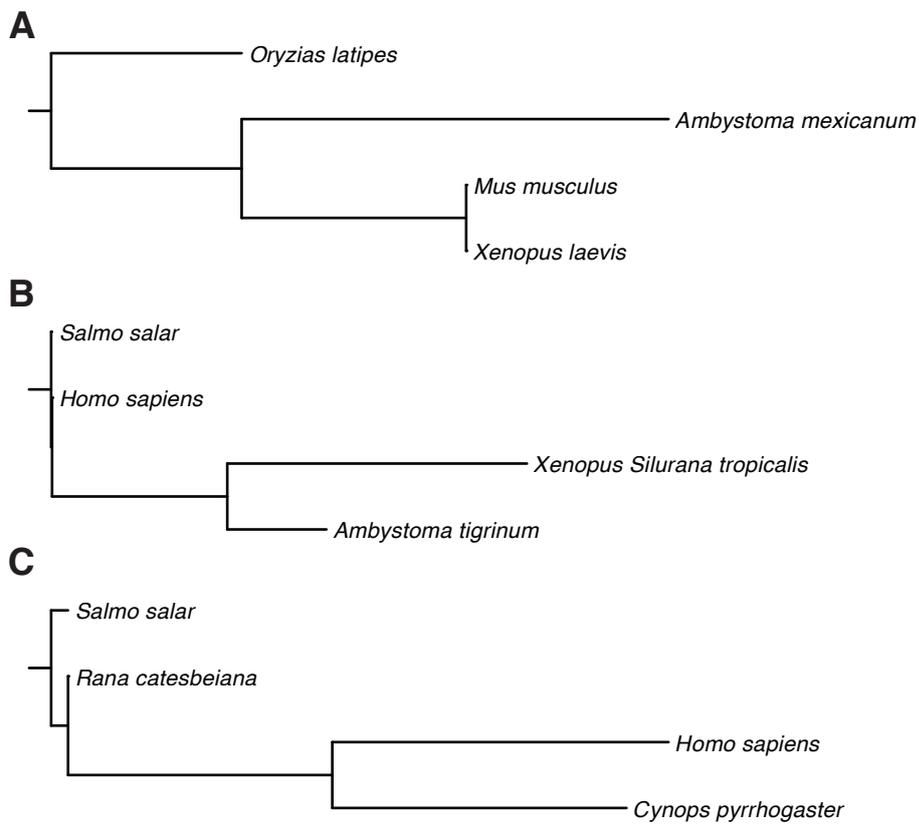


Figure B.2: Example trees with short branch lengths. Each of these trees features one or two extremely short branch lengths, producing an unreliable tree. The minimum uncorrected distances were 0 (A), 0.002 (B) and 0.018 (C).

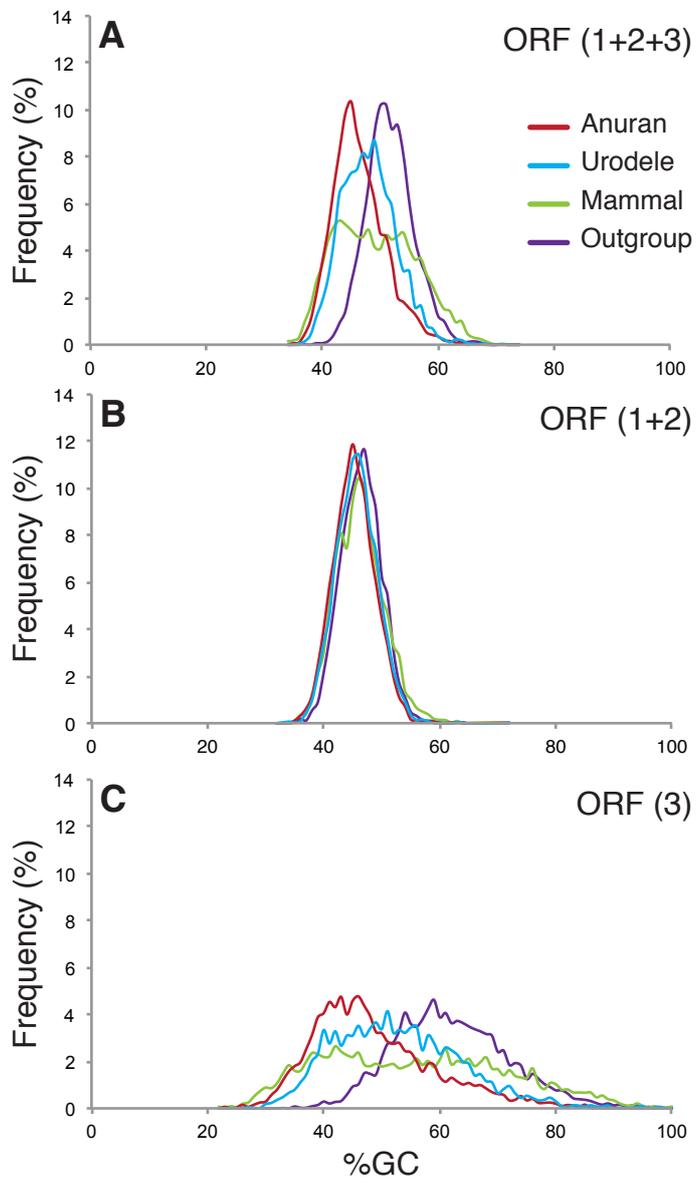


Figure B.3: The %GC for using the different codon positions. Using all the trees that we were able to build, this figure shows the frequency of each %GC for the four sequences used in the alignments. There is a large variance within the third codon compared to the first and second.

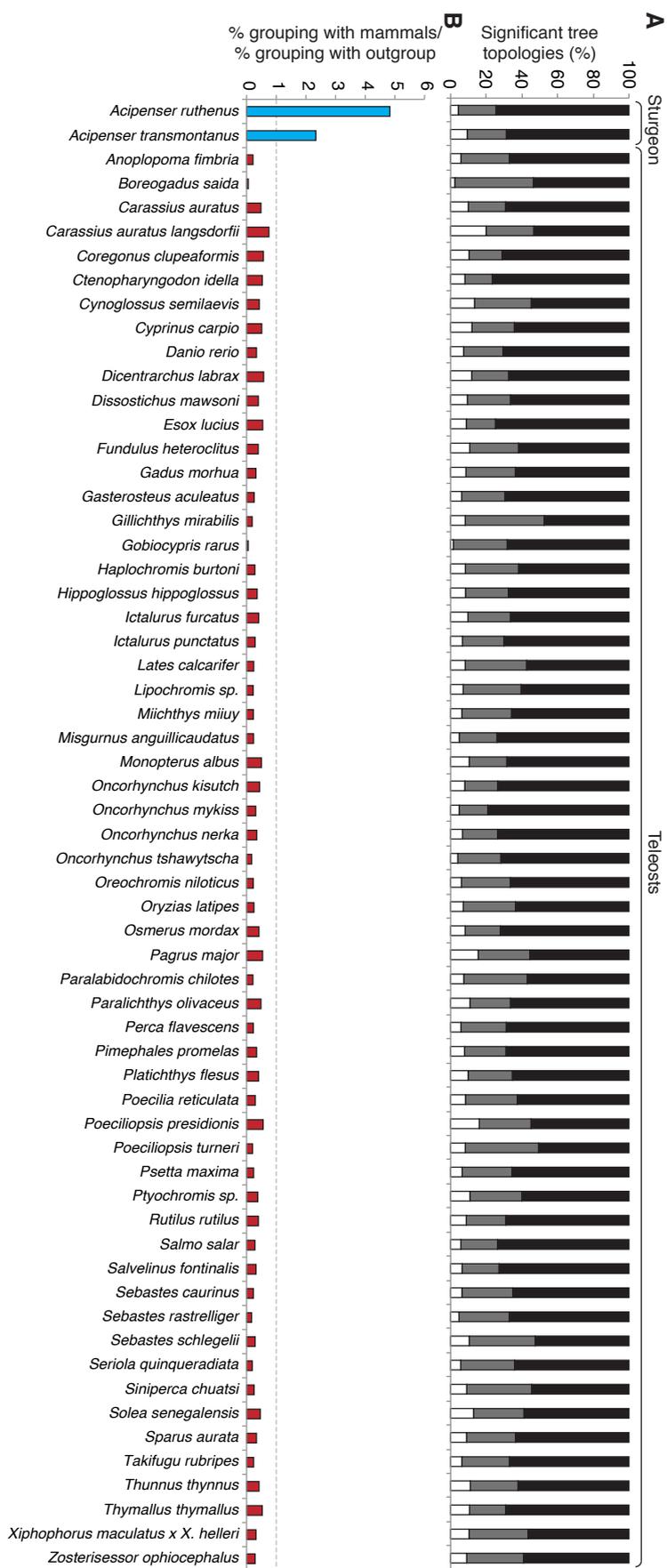


Figure B4: The Actinopterygian bootstrap results, including the sturgeon transcriptome. (A) shows the proportion of species phylogeny (black), Mammal-Sturgeon (grey) and Mammal-Teleost (white) topologies. **(B)** shows the likelihood of each species grouping with mammals when the tree is incongruent. All actinopterygian species with more than 20 significant trees are shown.

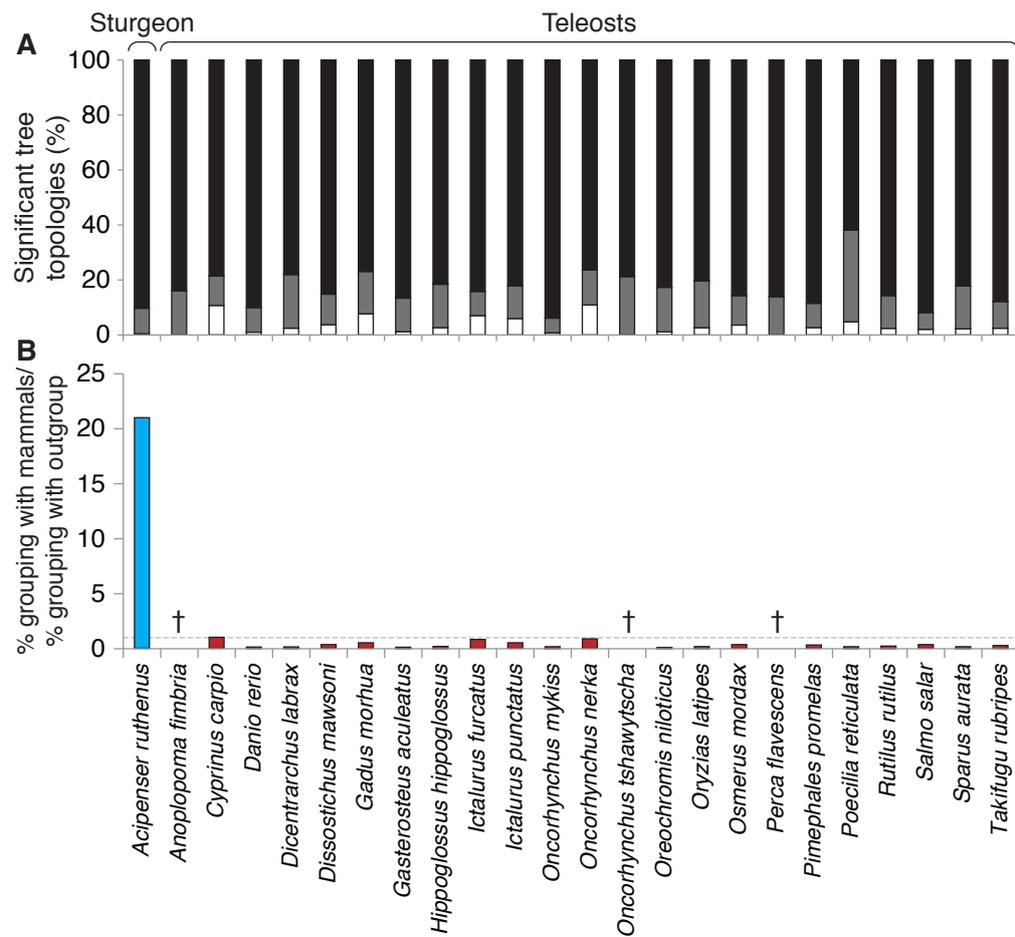


Figure B.5: The Actinopterygian SH test results, including the sturgeon transcriptome. The results are presented as in Figure B.4. All actinopterygian species with more than 20 significant trees are shown. Three species (†) have no incongruent trees in which they are grouped with Mammals.

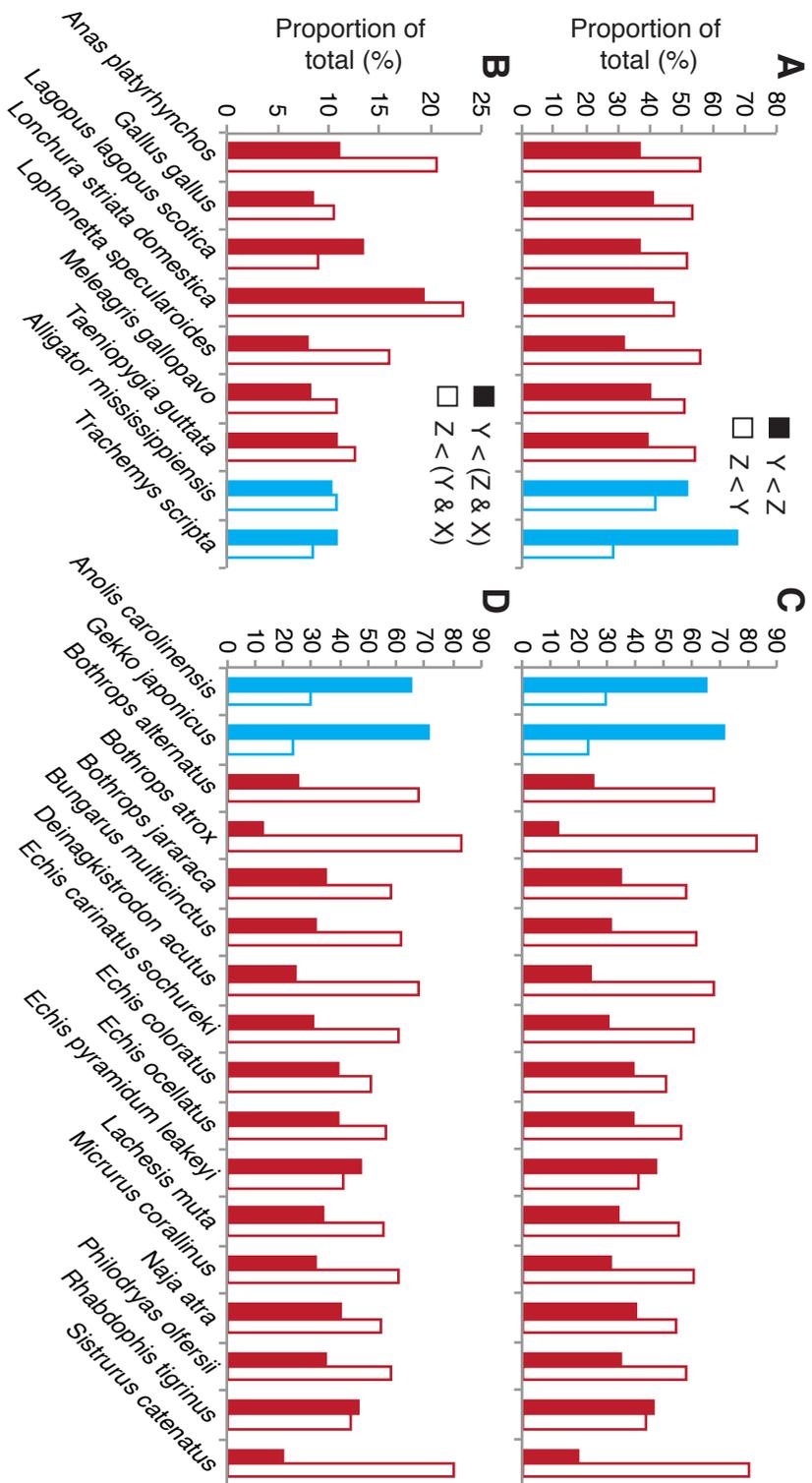


Figure B.6: The sauropsid distance matrix results. The Archosaur and Testudines (A and B) and Lepidosaur (C and D) species are coloured according to the mode of PGC specification, blue for epigenesis (Crocodile, Turtle and Lizards), red for preformation (Birds and Snakes). The top panels (A and C) show the proportion of sequences where the query-mammal distance is less than the sister-mammal distance in filled bars (Y < Z), and the opposite in clear bars (Z < Y). The bottom panels (B and D) show the proportion of results where the query-mammal distance is the smallest overall in filled bars (Y < (Z & X)), and the proportion where the sister-mammal distance is smallest overall in clear bars (Z < (Y & X)).

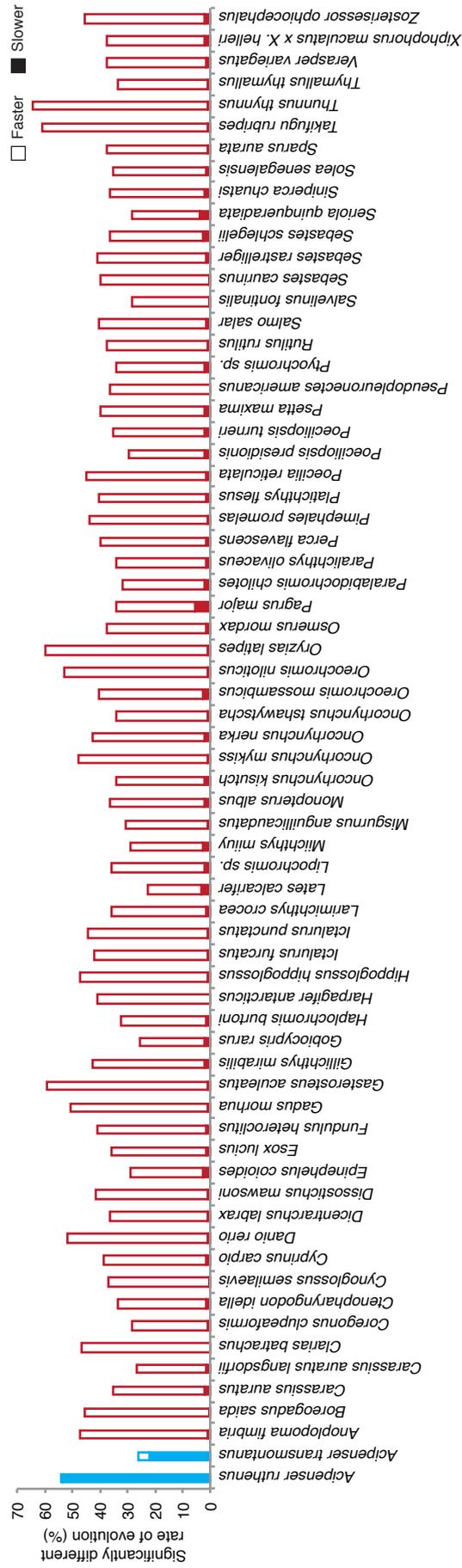


Figure B.7: Actinopterygian relative rate results. The results are presented as in Figure 3.32 but for all actinopterygian species with more than 20 significant results. The filled bars show the proportion of sequences evolving significantly slower, the clear bars show the proportion evolving significantly faster. Species are coloured by mode of PGC specification, epigenesis (blue) and preformation (red).

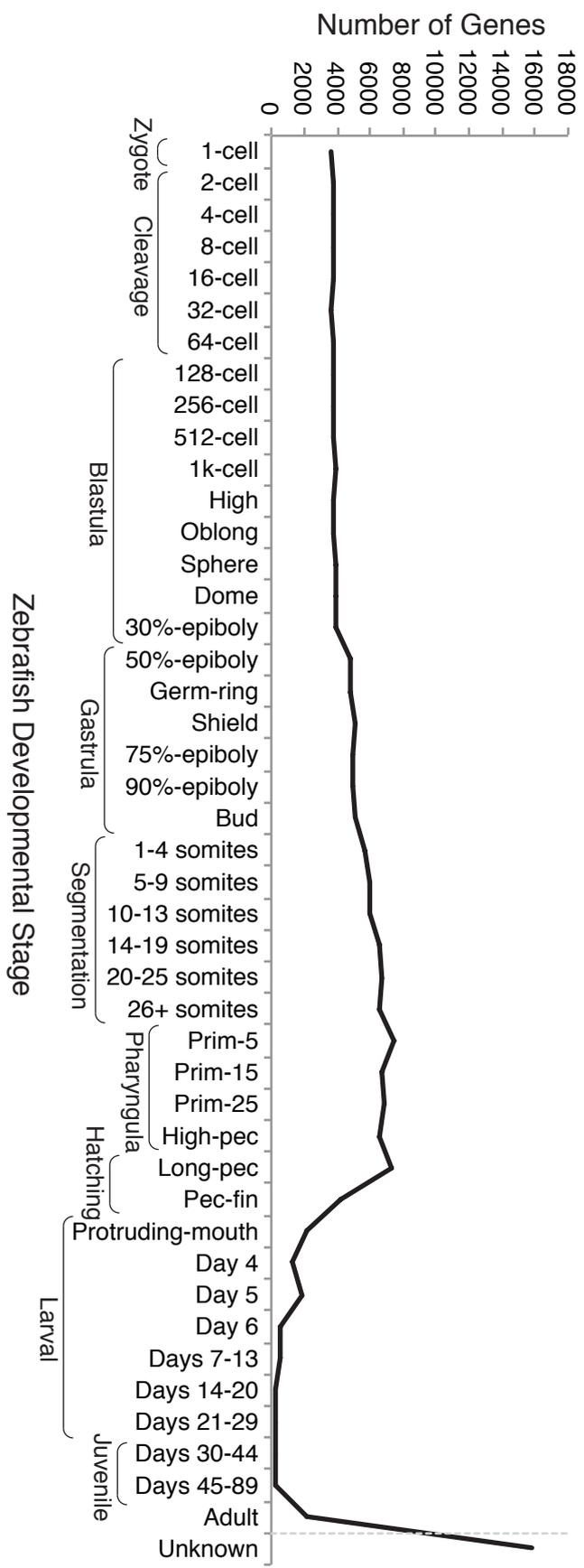


Figure B.8: Number of Zebrafish genes expressed at each stage of development. For each stage of zebrafish development the number of genes known to be expressed are plotted. The number of genes with no known expression are also included.

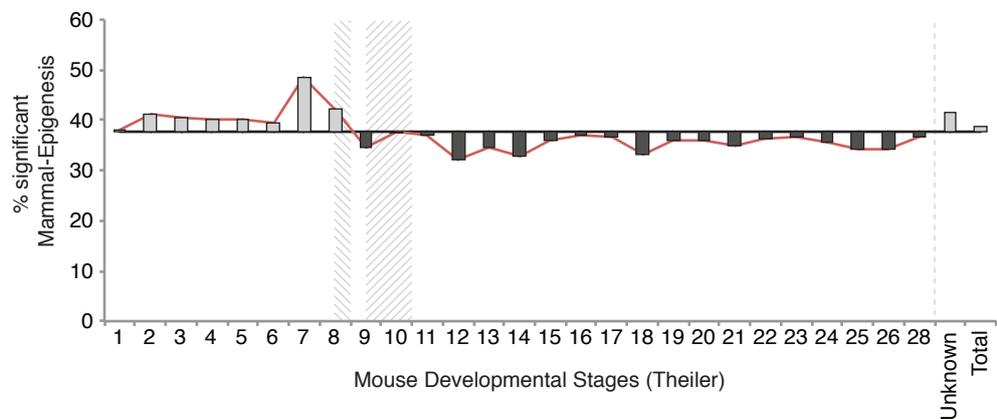


Figure B.9: The proportion of mouse genes with a Mammal-Epigenesis result expressed at each developmental stage. For each stage of development the proportion of expressed genes with orthologs that produced a Mammal-Epigenesis topology significant by bootstrapping are shown. The horizontal line represents the average for all genes with known expression information. The value at each stage is compared to the average for a significant difference (Chi-squared test; $*p < 0.05$; 1 d.f.). The stages of PGC induction (late TS 8) and gastrulation (TS 9.5-10) are hashed.

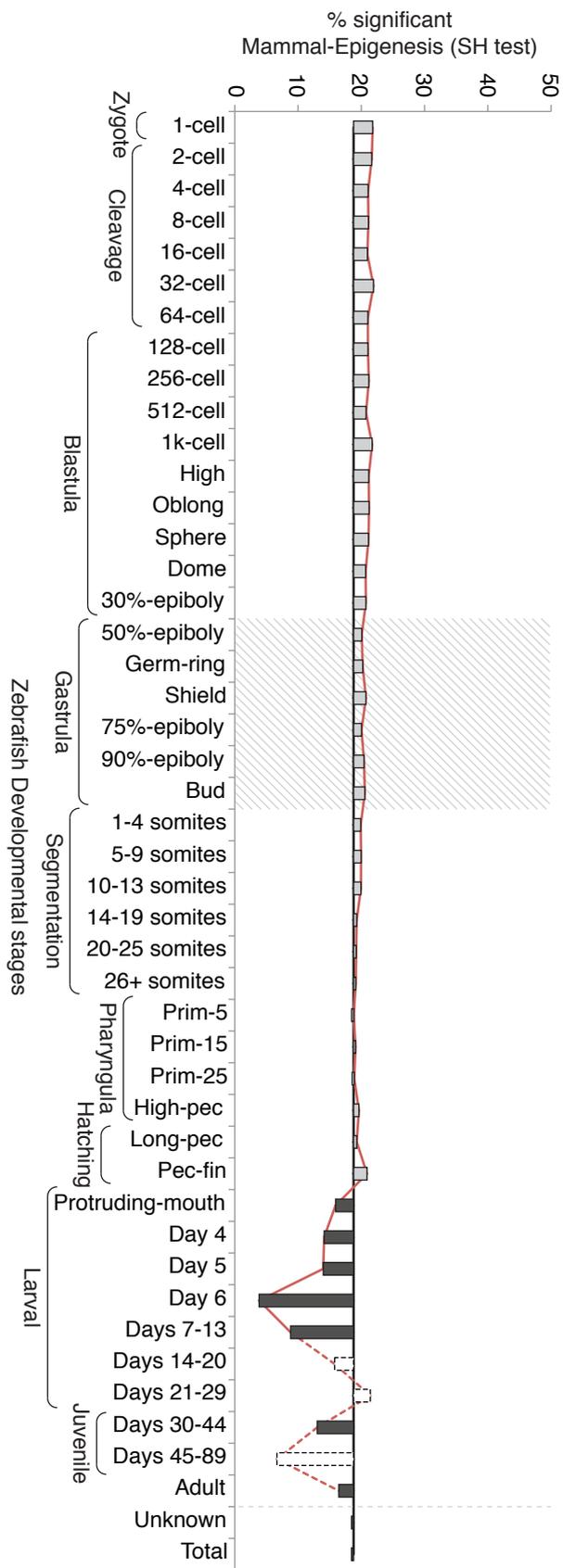


Figure B.10: The proportion of zebrafish genes with a Mammal-Epigenesis SH-test result expressed at each developmental stage. This graph is the equivalent to Figure 4.8 but for zebrafish genes instead of mouse. The difference between the proportion at each stage and the average is deemed significant using the Chi-squared test (* $p < 0.05$, ** $p < 0.01$, * $p < 0.001$; 1 d.f.). Zebrafish gastrulation is hashed. Stages with fewer than 20 total genes are shown as dashed outlines.**

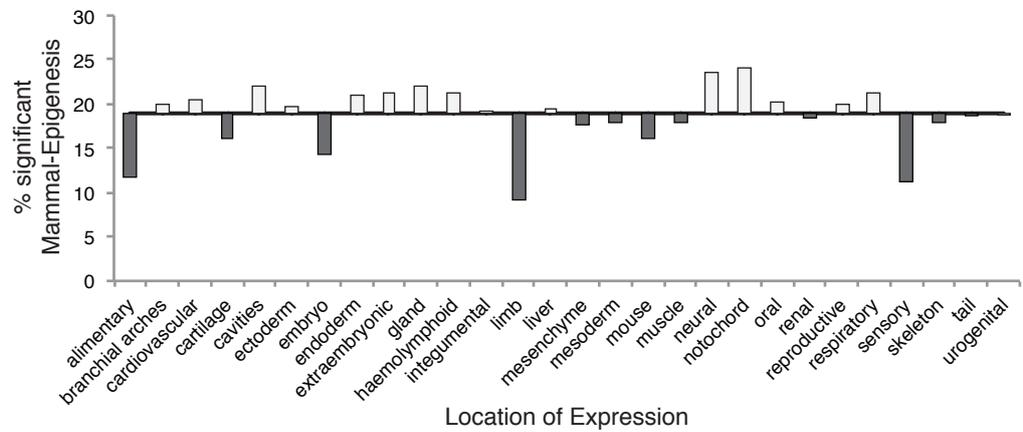


Figure B.11: Proportion of Mammal-Epigenesis trees significant by the SH-test expressed in each mouse higher-tier location. For each higher-tier location within the mouse embryo, the proportion of genes with a Mammal-Epigenesis result significant by the SH-test is shown relative to the average (solid line). The difference between the two is judged to be significant using the Chi-squared test.

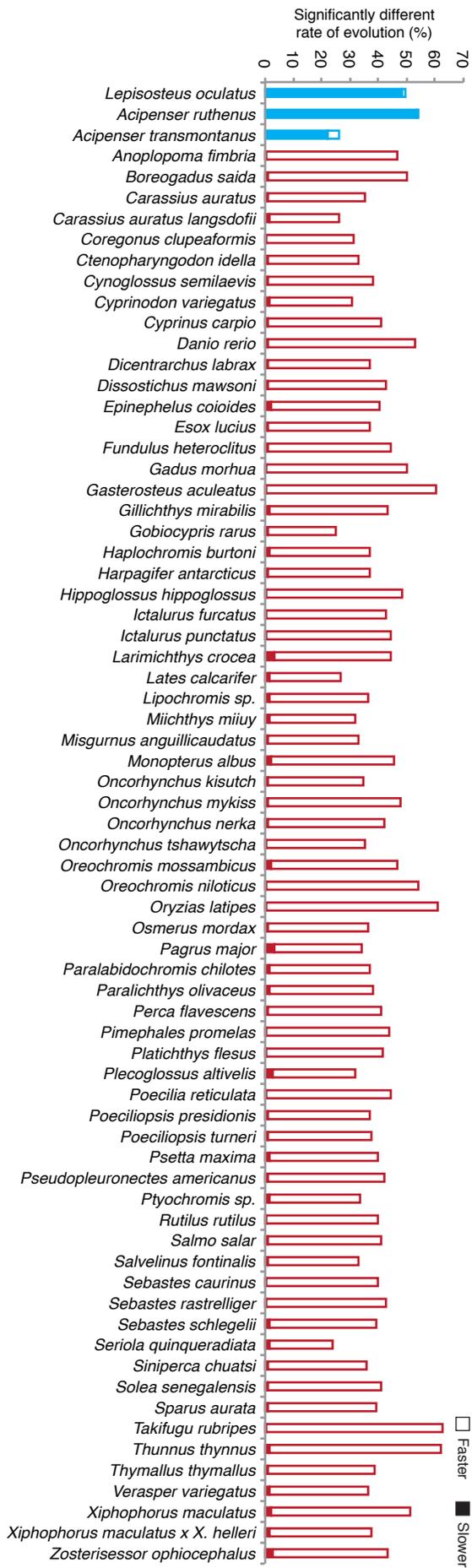


Figure B.12: Expanded Actinopterygian relative rate results. The results are presented as in Figure 5.7 but for all actinopterygian species with more than 20 significant results. The filled bars show the proportion of sequences evolving significantly slower, the clear bars show the proportion evolving significantly faster. Species are coloured by mode of PGC specification, epigenesis (blue) and preformation (red).

Figure B.13: POU gene family phylogeny. The bayesian (A) and maximum-likelihood (B) trees were created using the protein sequences from a few selected species. Both trees are rooted on the POU6 *Amphimedon queenslandica* sequence. Figure is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

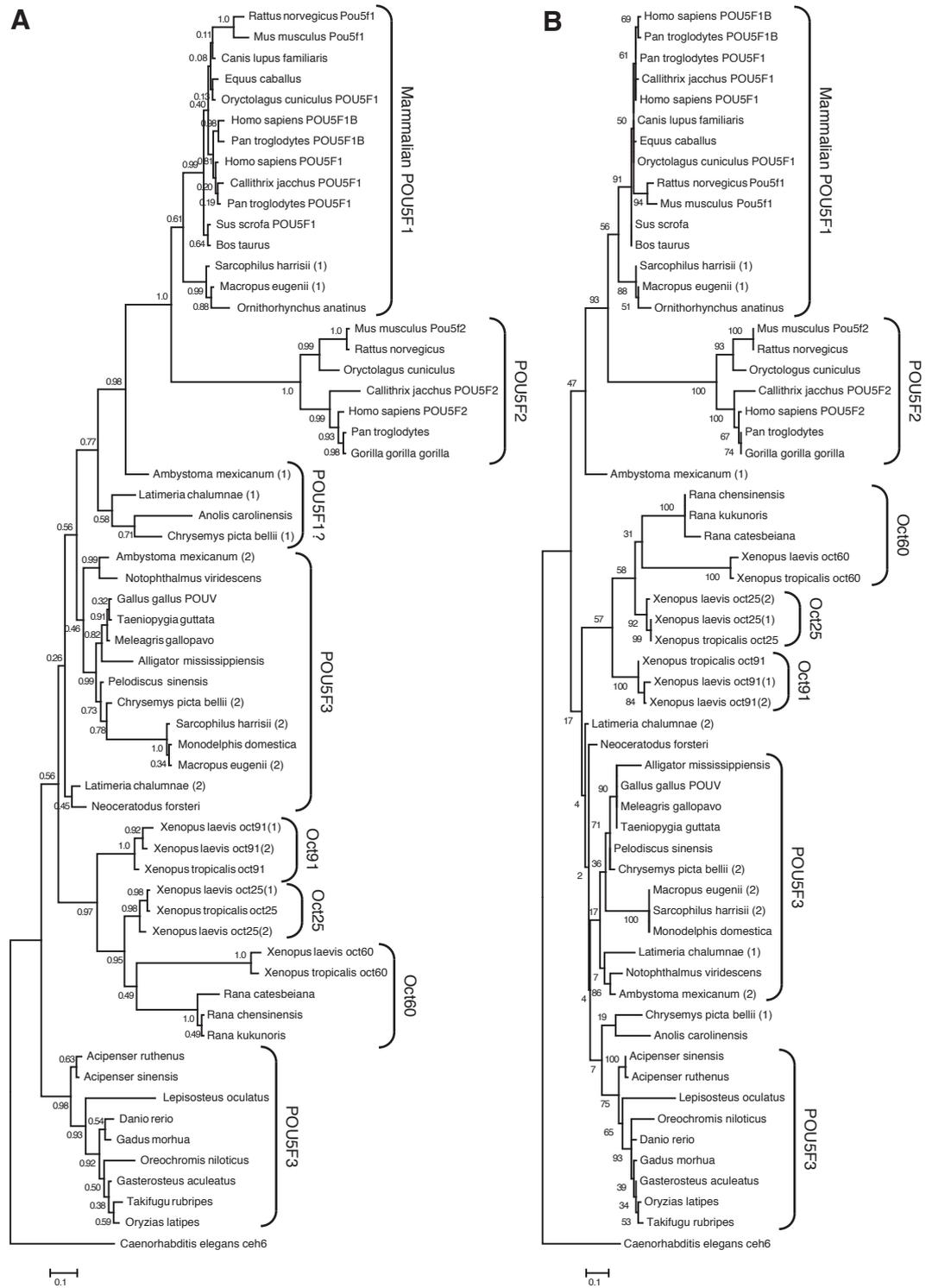


Figure B.14: POU5 phylogenies. The bayesian (A) and maximum-likelihood (B) trees were created using the full length protein sequences. Both trees are rooted on the *Caenorhabditis elegans* ceh6 gene.

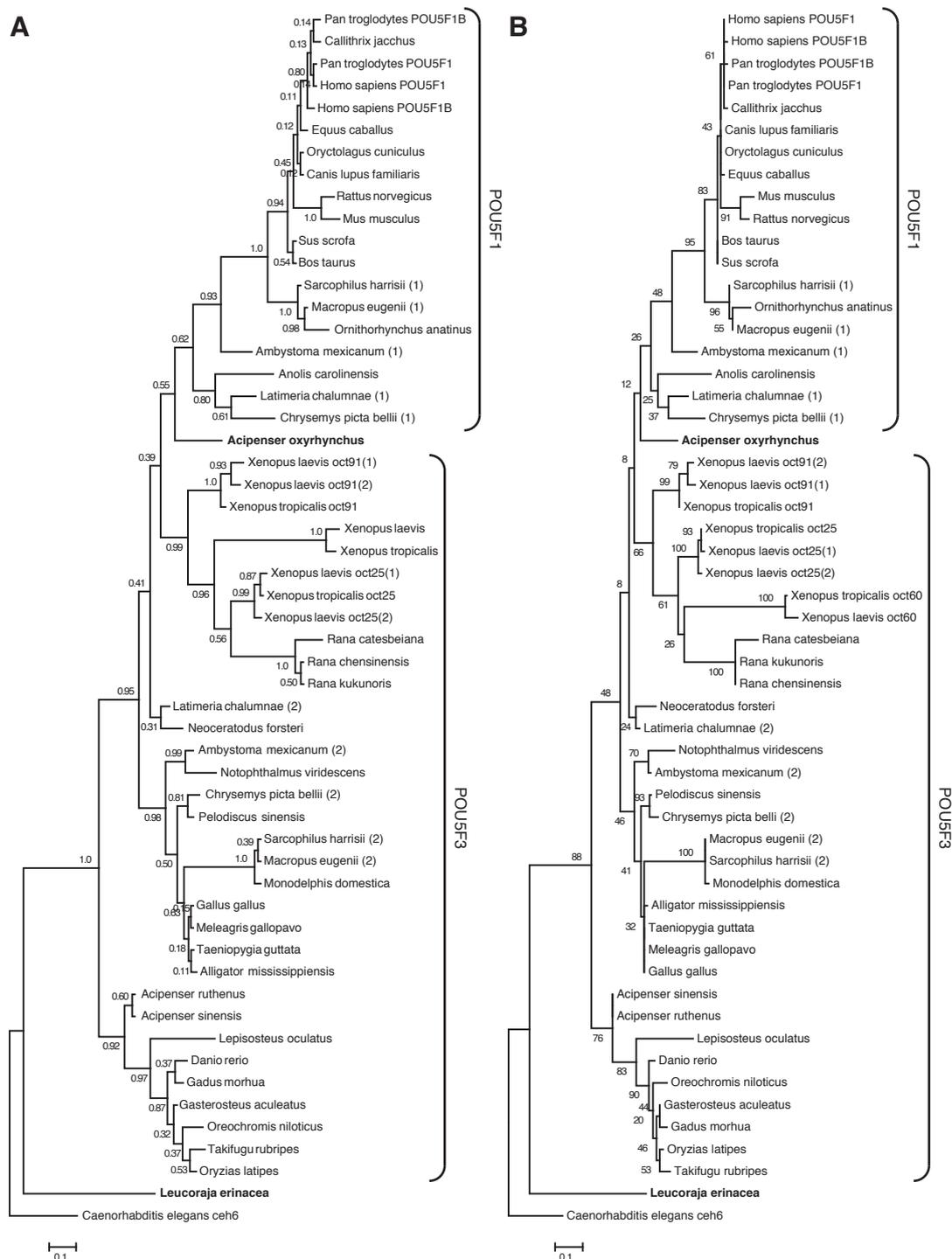


Figure B.15: POU5 phylogenies including Sturgeon and Skate. The bayesian (A) and maximum-likelihood (B) trees were built using the sequences in Figure 6.3 but excluding the POU5F2 gene. To this we added the Sturgeon sequence from Johnson et al., 2003a and a contig from the *Leucoraja erinacea* transcriptome, both of which are shown in bold font.

Figure B.16: DNA phylogeny of the genes Sox1, Sox2 and Sox3. The bayesian (A) and maximum-likelihood (B) trees were created using full length sequences. The trees are rooted on the sea anemone *Nematostella vectensis*. Figure is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

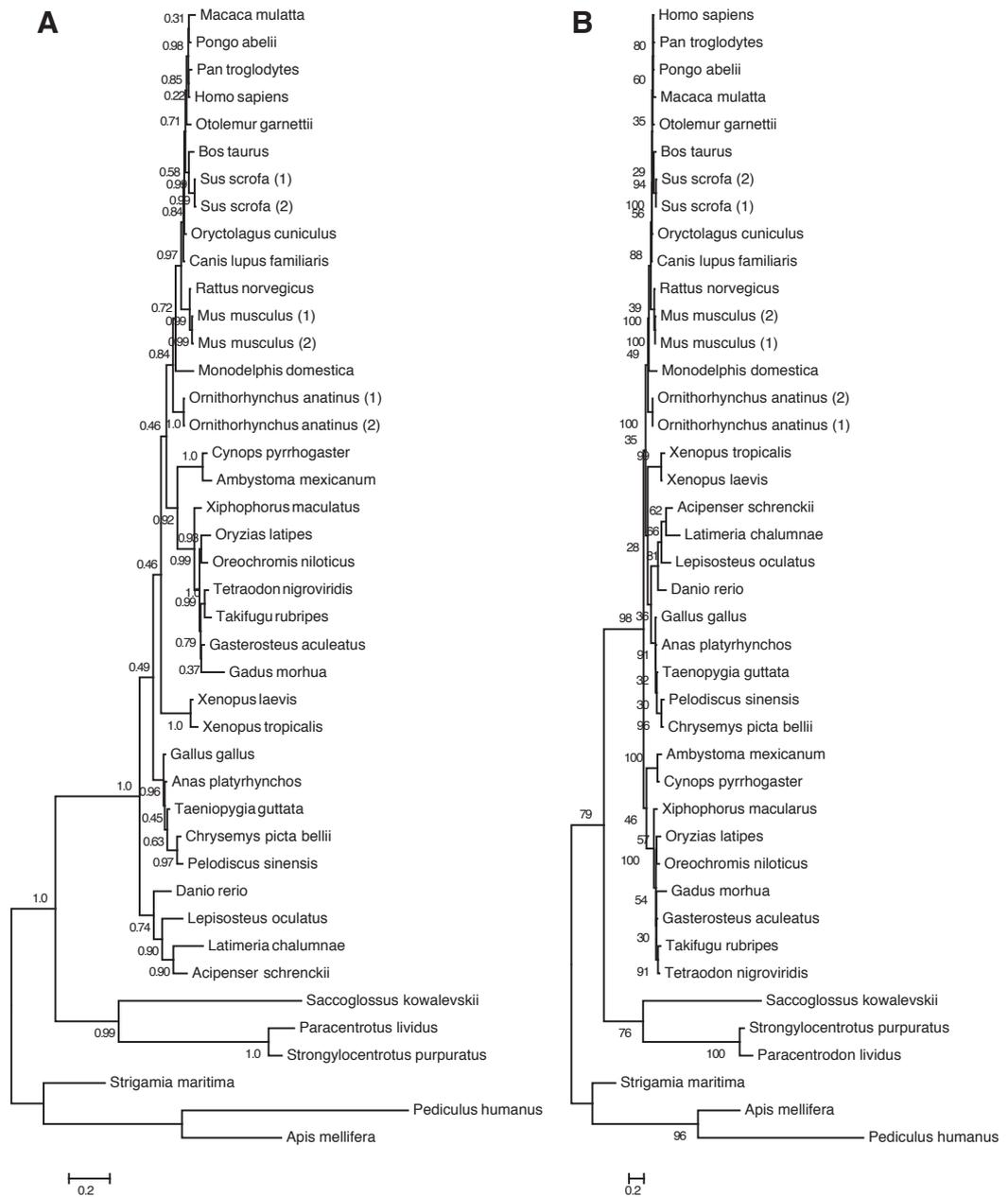


Figure B.17: DNA phylogeny of Sox2. The bayesian (A) and maximum-likelihood (B) trees were created using full length sequences from the Sox2 gene. Both trees are rooted on the protostome species.

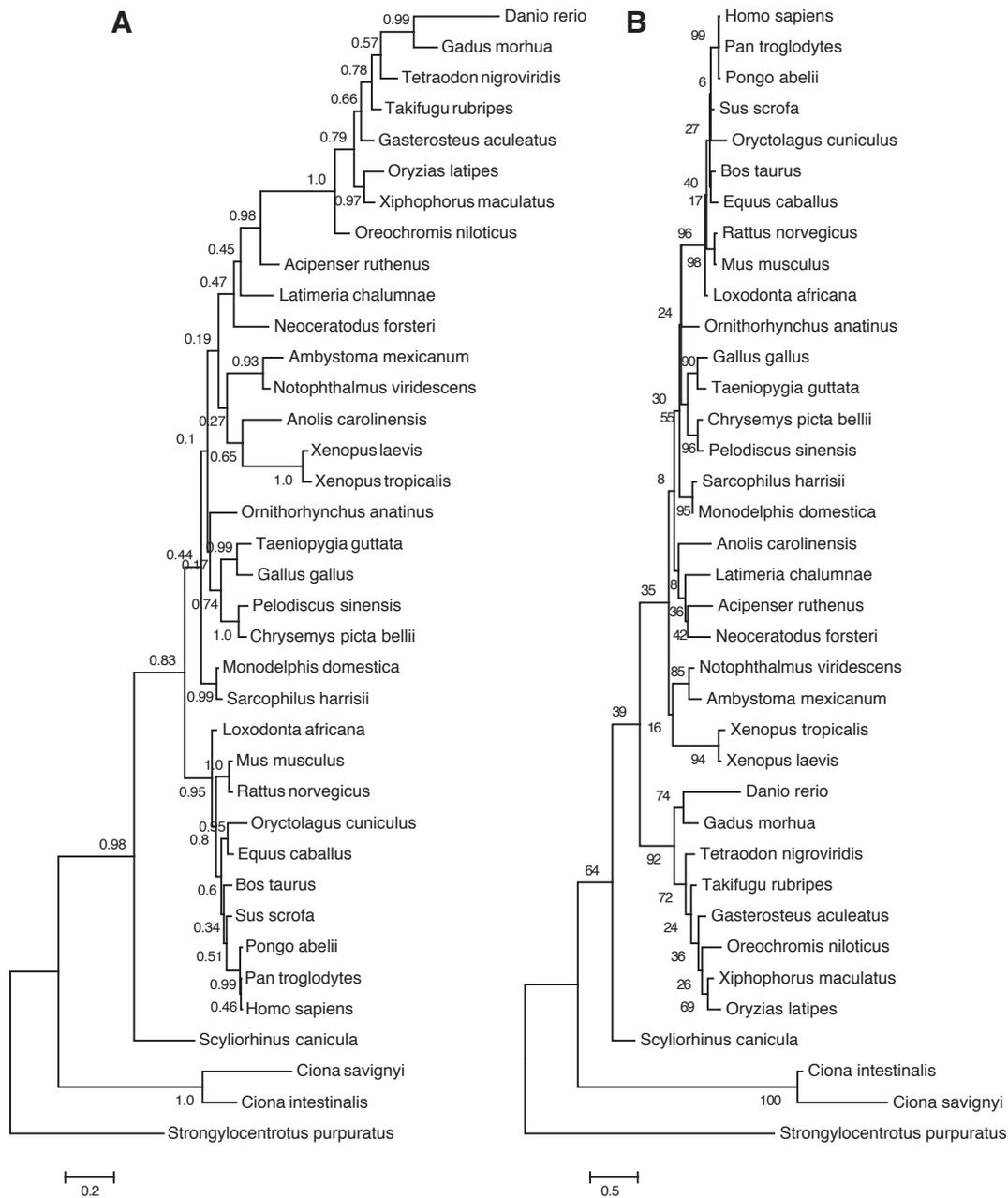


Figure B.18: KLF4 Phylogeny. The bayesian (A) and ML (B) alignments were created using full length DNA sequences and rooted on the echinoderm *Strongylocentrotus purpuratus*.

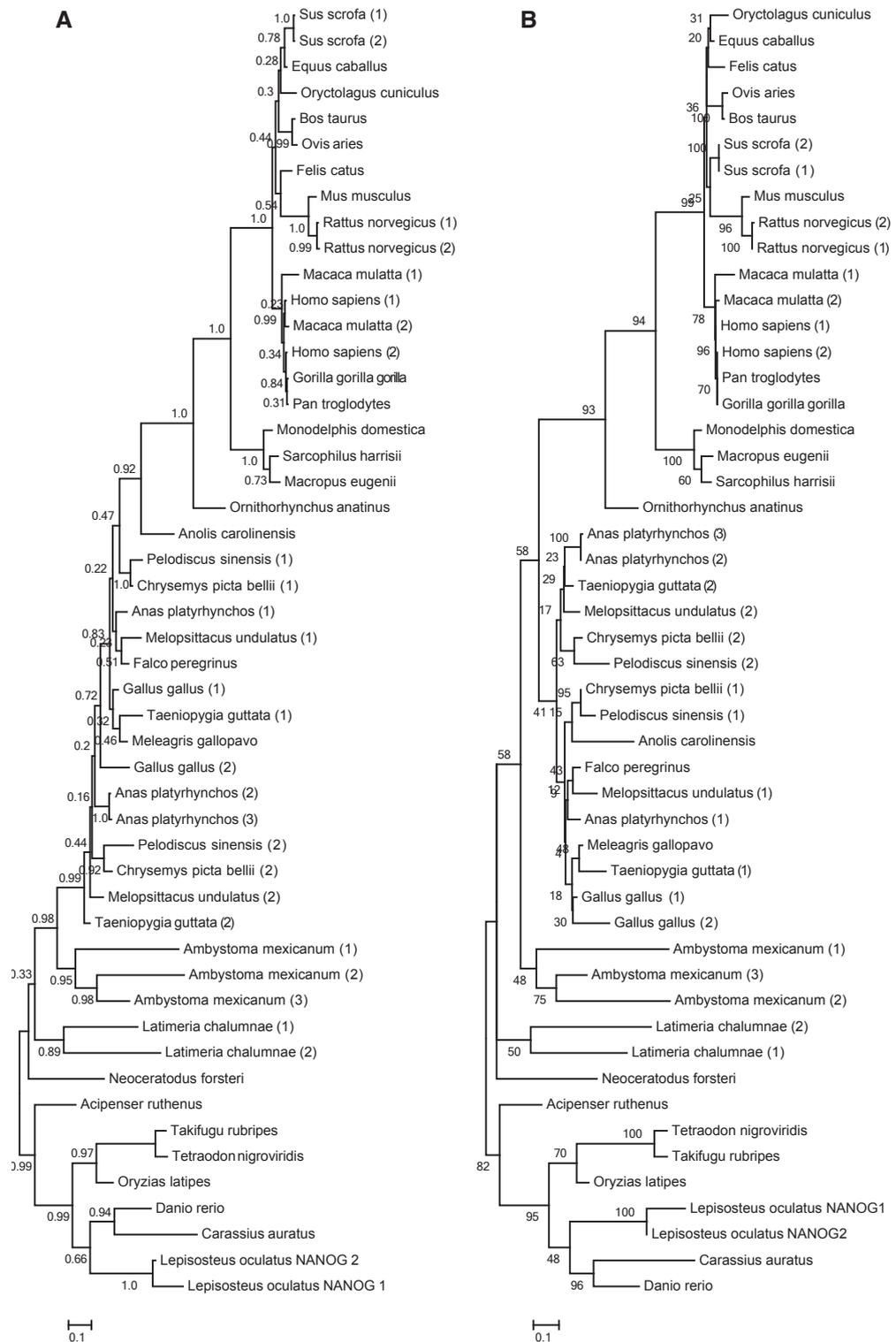


Figure B.19: Nanog DNA Phylogeny. The bayesian (A) and ML (B) alignments were created using full length DNA sequences and rooted on the Actinopterygian species.

Figure B.20: Nanog and other NK-like homeoboxes DNA phylogeny. The bayesian (A) and ML (B) trees are shown, each is rooted on the POU3 *Caenorhabditis elegans* sequence. Figure is on accompanying CD-ROM and at <http://dx.doi.org/10.6084/m9.figshare.1267447>.

13. G. Barnea *et al.*, *Science* **304**, 1468 (2004).
14. M. Gossen, H. Bujard, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5547–5551 (1992).
15. J. A. Gogos, J. Osborne, A. Nemes, M. Mendelsohn, R. Axel, *Cell* **103**, 609–620 (2000).
16. C. R. Yu *et al.*, *Neuron* **42**, 553–566 (2004).
17. M. Q. Nguyen, Z. Zhou, C. A. Marks, N. J. P. Ryba, L. Belluscio, *Cell* **131**, 1009–1017 (2007).
18. A. Fleischmann *et al.*, *Neuron* **60**, 1068–1081 (2008).
19. M. Q. Nguyen, C. A. Marks, L. Belluscio, N. J. P. Ryba, *J. Neurosci.* **30**, 9271–9279 (2010).
20. B. M. Shykind *et al.*, *Cell* **117**, 801–815 (2004).
21. A. Tsuboi *et al.*, *J. Neurosci.* **19**, 8409–8418 (1999).

22. A. Vassalli, A. Rothman, P. Feinstein, M. Zapotocky, P. Mombaerts, *Neuron* **35**, 681–696 (2002).
23. E. B. Brittebo, *Pharmacol. Toxicol.* **76**, 76–79 (1995).
24. A. Nakashima *et al.*, *Cell* **154**, 1314–1325 (2013).
25. J. L. Lefebvre, D. Kostandinov, W. V. Chen, T. Maniatis, J. R. Sanes, *Nature* **488**, 517–521 (2012).
26. S. L. Zipursky, W. B. Grueber, *Annu. Rev. Neurosci.* **36**, 547–568 (2013).

Acknowledgments: We thank R. Axel, E. Morrow, and members of the Barnea laboratory for critical reading of the manuscript. We thank R. Y. Korsak and M. Talay for artwork for the figures. This work was supported by NIH grants

T32GM007601 (L.T.) and 5R01MH086920 (G.B.), as well as by funds from the Pew Scholar in the Biomedical Sciences program.

Supplementary Materials

www.sciencemag.org/content/344/6180/197/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S4
Reference (27)

21 November 2013; accepted 28 February 2014
10.1126/science.1248806

Acquisition of Germ Plasm Accelerates Vertebrate Evolution

Teri Evans,¹ Christopher M. Wade,¹ Frank A. Chapman,² Andrew D. Johnson,^{1*} Matthew Loose^{1*}

Primordial germ cell (PGC) specification occurs either by induction from pluripotent cells (epigenesis) or by a cell-autonomous mechanism mediated by germ plasm (preformation). Among vertebrates, epigenesis is basal, whereas germ plasm has evolved convergently across lineages and is associated with greater speciation. We compared protein-coding sequences of vertebrate species that employ preformation with their sister taxa that use epigenesis and demonstrate that genes evolve more rapidly in species containing germ plasm. Furthermore, differences in rates of evolution appear to cause phylogenetic incongruence in protein-coding sequence comparisons between vertebrate taxa. Our results support the hypothesis that germ plasm liberates constraints on somatic development and that enhanced evolvability drives the evolution of germ plasm.

The germ line of metazoans is established early in development with the specification of primordial germ cells (PGCs). Among vertebrates, the conserved mechanism for PGC specification involves their induction from pluripotent cells by extracellular signals, a process referred to as epigenesis (1, 2). However, in several lineages of vertebrates, an alternative mechanism evolved, termed preformation. Here, PGCs are determined by inheritance of germ plasm. Preformation evolved by convergence, which suggests that it may confer a selective advantage. Accordingly, the evolution of germ plasm is associated with morphological innovations and enhanced numbers of species within individual clades (1, 3, 4). Why this derived mode of PGC specification evolved repeatedly in vertebrates is unknown.

The best-studied contrast of epigenesis and preformation is within amphibians. The PGCs of urodele amphibians (salamanders) are specified by epigenesis, whereas in its sister lineage, anurans (frogs), PGCs contain germ plasm (5). Using the axolotl (*Ambystoma mexicanum*) as a model urodele, the ancestral gene regulatory networks (GRNs) for pluripotency and mesoderm specification in vertebrates were identified (6, 7). These

GRNs were conserved through the evolution of mammals (6, 7), which also employ epigenesis (8). In contrast, in frogs the master regulators of pluripotency as employed in mammals have been deleted (6, 9, 10), and the GRN for mesoderm underwent expansions of key regulatory mole-

cules (7, 11). Similar genetic innovations evolved in the GRNs for zebrafish development (12), which also uses preformation (13). The correlation of germ plasm with genetic change has been proposed to result from the relaxation of constraints on somatic development imposed by maintaining the PGC induction pathway (1, 3, 4). To investigate this possibility, we compiled available expressed sequence tag, mRNA and cDNA sequences from vertebrates (fig. S1A and table S1) identifying ortholog pairs shared between sister taxa with different modes of PGC specification and an appropriate mammal and outgroup sequence (14) (fig. S1B). To increase sequence numbers from organisms using epigenesis, we generated transcriptomes from the axolotl and an Acipenseriforme, *Acipenser ruthenus* (the sterlet) (14), identifying 82,954 sequence clusters across all vertebrates. All analyses were performed with protein coding DNA sequence, excluding the saturated third position (14) (figs. S2 and S3).

Of the 56 published gene trees involving an anuran and a urodele, 29 do not recapitulate the known species phylogeny (table S2). The majority of the incongruent gene trees group urodele

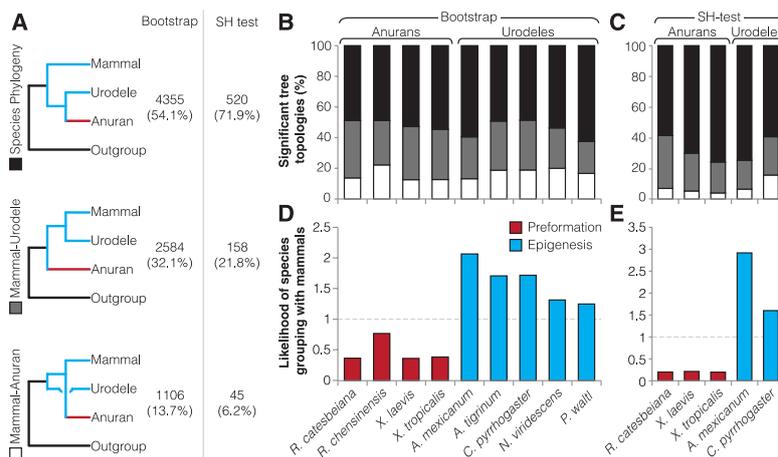


Fig. 1. Amphibian four-taxon tree topologies. (A) Number of significant trees by bootstrapping (>70%) and SH test ($P < 0.05$) for each topology rooted with a Teleostei sequence. (B and C) The proportions of species phylogeny (black), mammal-urodele (gray), and mammal-anuran (white) topologies per species. (D and E) The likelihood of each species grouping with mammals when the tree is incongruent; species using preformation are shown in red, those using epigenesis in blue. Dashed lines indicate equal probability of species grouping with mammal or outgroup. [(B) to (E)] Only species with >20 significant trees are shown. The results excluding the transcriptome are shown in fig. S4.

¹School of Life Sciences, University of Nottingham, Nottingham, NG7 2UH, UK. ²Program of Fisheries and Aquatic Sciences, University of Florida, Gainesville, FL 32653–3071, USA.

*Corresponding author. E-mail: matt.loose@nottingham.ac.uk (M.L.); andrew.d.johnson@nottingham.ac.uk (A.D.J.)

and mammal species together, both of which undergo epigenesis, to the exclusion of anurans. We generated unrooted four-taxon trees to investigate the extent of this incongruity (14), presenting trees rooted on the known outgroup (a Teleostei sequence) (Fig. 1A). Within these trees, 54.1% (4355 of 8045) of amphibian sequences show the expected species phylogeny (>70% bootstrap), grouping anuran with urodele. The majority of the remainder [32.1% (2584 of 8045)] incongruently group urodeles with mammals (Fig. 1A). The Shimodaira-Hasegawa (SH) test reduces the number of significant trees overall ($P < 0.05$), increasing the proportion of trees reflecting the species phylogeny (14). Orthology groups that do not reflect the species phylogeny (28%) are three times as likely to place the urodele sequence with the mammal (Fig. 1A, fig. S4A, and table S3). We next considered each amphibian species in turn, grouping them by mode of PGC specification (Fig. 1, B and C). We show that when a tree is incongruent, any given anuran sequence is less likely than its orthologous urodele sequence to group with mammals (Fig. 1, D and E). These results do not depend on the inclusion of the urodele transcriptome data (fig. S4).

Within Actinopterygii (ray-finned fishes), Teleostei (teleosts) use preformation, whereas Acipenseriformes (sturgeons and paddlefish), which maintain primitive embryological and adult traits, most likely have retained epigenesis (1, 4, 13). We identified 19,394 trees with >70% bootstrap support, of which 68.2% (13,233) reflect the species phylogeny. The majority of the remainder [24.5% (4757)] incongruently group Acipenseriformes with mammals (Fig. 2A). The SH test reduces the total, but still Acipenseriforme sequences are 5 times as likely to group with mammals when the species phylogeny is not obtained (Fig. 2A). Subdividing the data by species reveals a clear distinc-

tion between Teleostei and Acipenseriformes; in incongruent trees, Acipenseriformes are more likely to group with a mammal (Fig. 2, B to E). This is true for all 59 Teleostei analyzed with bootstrap-supported trees and 22 of 23 Teleostei supported by the SH test (fig. S5, A and B, and table S4). These results remain true even if the transcriptome data are excluded (fig. S5, C and D).

We next investigated the sauropsids (reptiles and birds), determining four-taxon tree topologies. In sauropsids, preformation evolved independently in lepidosaurs (lizards and snakes) and in archosaurs (crocodiles and birds) (15–18). The lepidosaurs, which experienced a change in the rate of evolution (19), and archosaurs, separated ~280 million years ago (20). The turtle lineage (testudines), using epigenesis, is closer to the archosaurs than lepidosaurs (21). Thus, we analyzed the sauropsids in two subdivisions—the archosaurs and testudines, and the lepidosaurs. Within these groups, we compared birds (preformation) with crocodiles and testudines (epigenesis) and similarly snakes (preformation) with Gekkota and Iguanidae lizards (epigenesis) (15–18). Within the sauropsids, almost all the four-taxon trees support the expected species phylogeny and do not subdivide by the mode of germ cell specification in this analysis (fig. S6 and table S5), although the total number of sequence comparisons was low (fig. S1A).

Nonetheless, among the amphibians and actinopterygians, but not the sauropsids, when in an incongruent tree, species using epigenesis are more likely to group with mammals (Figs. 1,2). Such incongruent phylogenies may be driven by differences in the rate of sequence evolution (19, 22), and organisms that have acquired preformation are typically more speciose than those using epigenesis (3, 4). We therefore used three-taxon multiple alignments to determine how the relative rate of sequence evolution differs be-

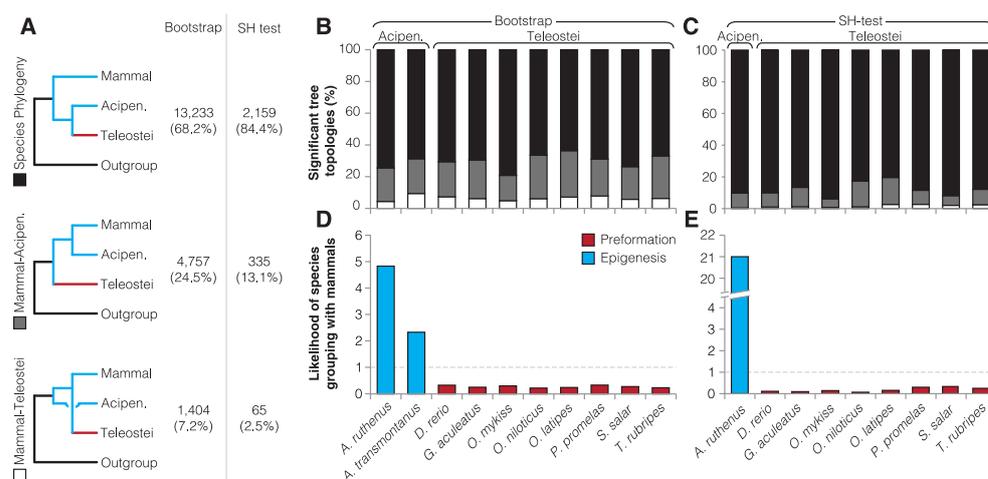
tween sister taxa. Among amphibian species, 32.3% of sequences are evolving at significantly different rates ($P < 0.05$), of which 87% show urodele sequences evolving slower than anurans (Fig. 3A and fig. S7B). Within the actinopterygians, ~50% of sequences evolve at significantly different rates ($P < 0.05$), with almost all showing that Acipenseriforme sequences are slower than Teleostei (Fig. 3B and fig. S7, A and B). Furthermore, in the sauropsids, ~20 to 25% of sequences are evolving at significantly different rates, with the majority of slow-evolving sequences in organisms using epigenesis ($P < 0.05$) (Fig. 3, C and D). Thus, sauropsid sequences exhibit differences in the rate of sequence evolution that correlate with the mode of PGC specification.

Combining these data across classes, 56% (69,165 of 121,373) of all analyses show no significant difference in the rate of sequence evolution (Fig. 3E and table S6). Only 2319 of 121,373 relative-rate tests (<2%) showed a sequence derived from an organism with epigenesis evolving faster than its ortholog. The remaining 41.1% of comparisons (49,898 of 121,373) suggest that sequences from organisms using epigenesis are evolving more slowly. Ranking each species by the proportion of slower-evolving sequences separates organisms using epigenesis from those using preformation, regardless of taxonomic class (fig. S7F).

To investigate functional properties of sequences showing accelerated rates of evolution and incongruent phylogenies, we mapped our results to the mouse and zebrafish genomes (see supplementary text). The proportion of sequences showing evidence of accelerated evolution is significantly higher among genes expressed early in development (chi-square test, $P < 0.05$) and decreases in genes expressed at later stages (fig. S8). Previous reports demonstrate that early genes are under the highest levels of developmental

Fig. 2. Actinopterygian four-taxon tree topologies.

(A) Number of significant trees by bootstrapping (>70%) and SH test ($P < 0.05$) for each topology, rooted with an amphioxus sequence. (B and C) The proportion of species phylogeny (black), Mammal-Acipenseriforme (gray) and Mammal-Teleostei (white) topologies per species. (D and E) The likelihood of each species grouping with mammals when the tree is incongruent; species using preformation are shown in red, those using epigenesis in blue. Dashed lines indicate equal probability of species grouping with mammal or outgroup. [(B) to (E)] Only Acipenseriformes with >20 significant trees and eight Teleostei species are shown; the results for all species are in fig. S5, A and B. The results excluding the transcriptome are shown in fig. S5, C and D.



constraint (23), and our data suggest that early genes are the most likely to evolve at faster rates in species employing preformation. Together, this supports the hypothesis that the evolution of germ plasm liberates constraints on early development (1, 4).

We next considered the correlation of results where individual sequences had been tested for both rate and incongruent tree topologies (table S11). All four-taxon trees were rooted on species using preformation, yet our preceding analyses suggest that these outgroups also have accelerated rates of evolution compared with their sister taxa (Fig. 4A). Because tree reconstruction may fail as a consequence of two long branches clustering (long-branch attraction)(Fig. 4B) (24), we asked whether the incongruent trees were driven by the

differences in rates observed in the outgroup sister taxa.

For the majority of amphibian four-taxon trees where the outgroup sister taxa differ in rate, the Teleostei sequence is evolving significantly faster than its Acipenseriforme ortholog [$P < 0.05$, (fig. S9A)]. We therefore rebuilt trees using Acipenseriformes as the outgroup. This increased the proportion of trees congruent with the species phylogeny at the expense of trees grouping urodele sequences with mammals (fig. S9B and table S12). Grouping sequences by both relative rate and tree topology revealed that the highest proportion of incongruent trees occur when the relative rate differs within amphibians (Fig. 4C). If rate differences drive incongruence, changing out-

group should only affect those trees where the outgroup rates differ. Where the actinopterygian sequences do not significantly differ in rate, proportions of incongruent trees remain similar as the outgroup species changes (Fig. 4, D and E, and table S13). Where actinopterygian sequences significantly differ in rate, the proportion of incongruent trees is reduced using an Acipenseriforme rather than a Teleostei outgroup ($P < 0.05$). The most dramatic change occurs when both amphibian and actinopterygian sequences significantly differ in rate ($P < 0.05$), suggesting that the faster rate of evolution in both anuran and Teleostei sequences drives the observed incongruence.

The natural history of vertebrates is punctuated with the repeated evolution of germ plasm associated with embryological innovations, gross morphological changes in adults, and enhanced speciation (1–4). Germ plasm functions to segregate PGCs from somatic cells at the inception of development, and we propose that this relaxes genetic constraints on the mechanisms that govern early somatic development (4). Our results identify a consistent bias in changes in the rate of sequence evolution in species using preformation compared with their sister taxa that use epigenesis. No other biological property correlates as well with the observed changes in rate (fig. S10). Sequences expressed during early development are under high levels of developmental constraint (23), and we show that these sequences exhibit a release of constraint in species using preformation. Taken together, these data suggest that the acquisition of germ plasm liberates developmental constraints, leading to increased rates of sequence evolution and enhanced speciation. They support the hypothesis that enhanced evolvability is responsible for the repeated evolution of germ plasm (1–4).

Fig. 3. Relative-rate test results. (A to D) Proportion of sequences evolving at significantly slower (filled) or faster (clear) rates in each species ($P < 0.05$; preformation shown in red, epigenesis in blue). (A) Amphibians. (B) Actinopterygians, including eight Teleostei species (all Teleostei shown in fig. S7G). (C) Archosaurs and Testudines. (D) Lepidosaurs. Only species with >20 significant sequences are shown. (E) Summary of relative-rate data across all vertebrates grouping species by epigenesis or preformation. [(A) (B), and (E)] Excluding transcriptomes are in figs. S7, C to E, respectively.

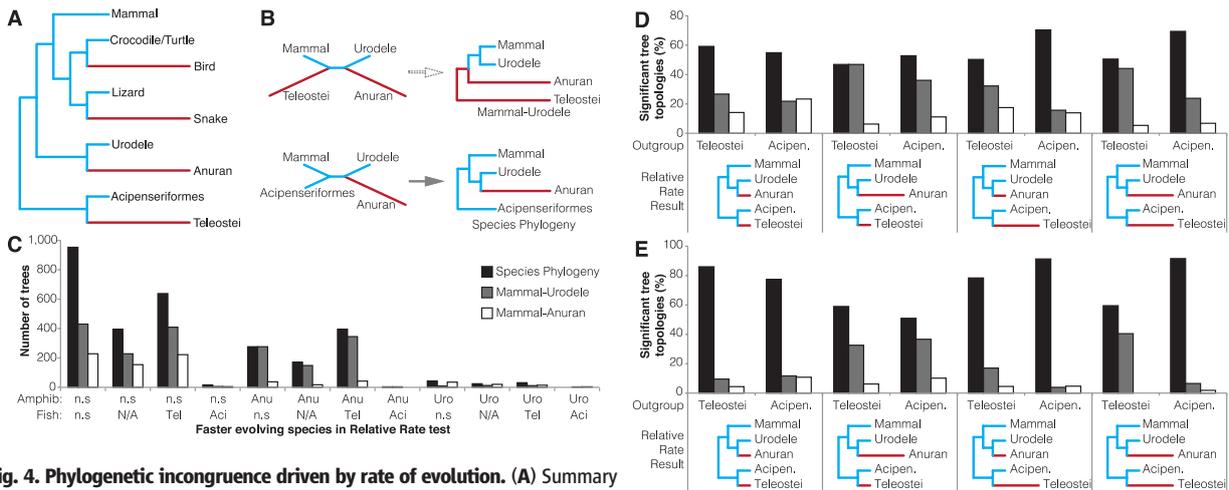
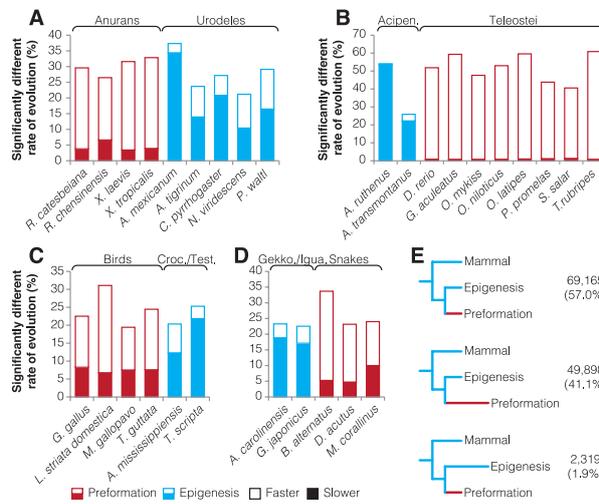


Fig. 4. Phylogenetic incongruence driven by rate of evolution. (A) Summary of relative-rate test results. (B) Tree diagrams illustrating long-branch attraction driven by outgroup choice in four-taxon trees. (C) Number of amphibian four-taxon tree topologies grouped by relative-rate differences between anurans and urodeles, and Teleostei and Acipenseriformes. n.s., not significant; N/A, not available; Anu, Anuran; Uro, Urodele; Tel, Teleostei; Aci, Acipenseriforme. (D and E) For the four common relative-rate test results between Amphibians and Actinopterygii, the proportions of Amphibian four-taxon tree topologies are shown when using Teleostei or Acipenseriforme outgroups. (D) Bootstrap trees. (E) SH test trees.

References and Notes

- A. D. Johnson *et al.*, *Evol. Dev.* **5**, 414–431 (2003).
- A. D. Johnson *et al.*, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **358**, 1371–1379 (2003).
- B. I. Crother, M. E. White, A. D. Johnson, *J. Theor. Biol.* **248**, 322–330 (2007).
- A. D. Johnson, E. Richardson, R. F. Bachvarova, B. I. Crother, *Reproduction* **141**, 291–300 (2011).
- P. D. Nieuwkoop, L. A. Sutasurya, *Primordial Germ Cells in the Chordates, Embryogenesis and Phylogenesis* (Cambridge Univ. Press, Cambridge, 1979).
- J. E. Dixon *et al.*, *Development* **137**, 2973–2980 (2010).
- G. Swiers, Y. H. Chen, A. D. Johnson, M. Loose, *Dev. Biol.* **343**, 138–152 (2010).
- H. G. Leitch, W. W. Tang, M. A. Surani, *Curr. Top. Dev. Biol.* **104**, 149–187 (2013).
- S. Frankenberg, M. B. Renfree, *BMC Biol.* **11**, 56 (2013).
- U. Hellsten *et al.*, *Science* **328**, 633–636 (2010).
- M. Loose, R. Patient, *Dev. Biol.* **271**, 467–478 (2004).
- X. Fan, S. T. Dougan, *Dev. Genes Evol.* **217**, 807–813 (2007).
- E. Raz, *Nat. Rev. Genet.* **4**, 690–700 (2003).
- Materials and methods are available as supplementary materials on Science Online.
- R. F. Bachvarova, B. I. Crother, A. D. Johnson, *Evol. Dev.* **11**, 603–609 (2009).
- R. F. Bachvarova *et al.*, *Evol. Dev.* **11**, 525–534 (2009).
- J. Hubert, in *Biology of the Reptilia: Development A.*, C. Gans, F. Billet, P. F. A. Maderson, Eds. (John Wiley & Sons, New York, 1985), pp. 41–74.
- N. Tsunekawa, M. Naito, Y. Sakai, T. Nishida, T. Noce, *Development* **127**, 2741–2750 (2000).
- S. Hughes, D. Mouchiroud, *J. Mol. Evol.* **53**, 70–76 (2001).
- S. B. Hedges, J. Dudley, S. Kumar, *Bioinformatics* **22**, 2971–2972 (2006).
- Z. Wang *et al.*, *Nat. Genet.* **45**, 701–706 (2013).
- L. M. Dávalos, A. L. Cirranello, J. H. Geisler, N. B. Simmons, *Biol. Rev. Camb. Philos. Soc.* **87**, 991–1024 (2012).
- J. Roux, M. Robinson-Rechavi, *PLOS Genet.* **4**, e1000311 (2008).
- M. J. Telford, R. R. Copley, *Curr. Biol.* **15**, R296–R299 (2005).

Acknowledgments: All data are available from the authors' Web site at www.nottingham.ac.uk/~plzloose/phyloinc and are deposited at <http://datadryad.org> (DOI:10.5061/dryad.rd70f). The authors thank B. Crother, G. Morgan, and M. Blythe for helpful discussion. Sequencing was carried out at the Genome Centre, Queen Mary, University of London by L. Bhaw-Rosun and M. Strueggig. This work was supported by the Biotechnology and Biological Sciences Research Council, Medical Research Council, and University of Nottingham.

Supplementary Materials

www.sciencemag.org/content/344/6180/200/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S10
Tables S1 to S13
References (25–78)

4 December 2013; accepted 18 March 2014
10.1126/science.1249325

PINK1 Loss-of-Function Mutations Affect Mitochondrial Complex I Activity via NdufA10 Ubiquinone Uncoupling

Vanessa A. Morais,^{1,2*} Dominik Haddad,^{1,2} Katleen Craessaerts,^{1,2} Pieter-Jan De Bock,^{3,4} Jef Swerts,^{1,2} Sven Vilain,^{1,2} Liesbeth Aerts,^{1,2} Lut Overbergh,⁵ Anne Grünewald,⁶ Philip Seibler,⁶ Christine Klein,^{6,7} Kris Gevaert,^{3,4} Patrik Verstreken,^{1,2} Bart De Strooper^{1,2,8*}

Under resting conditions, *Pink1* knockout cells and cells derived from patients with *PINK1* mutations display a loss of mitochondrial complex I reductive activity, causing a decrease in the mitochondrial membrane potential. Analyzing the phosphoproteome of complex I in liver and brain from *Pink1*^{-/-} mice, we found specific loss of phosphorylation of serine-250 in complex I subunit NdufA10. Phosphorylation of serine-250 was needed for ubiquinone reduction by complex I. Phosphomimetic NdufA10 reversed *Pink1* deficits in mouse knockout cells and rescued mitochondrial depolarization and synaptic transmission defects in *pink1*^{B9}-null mutant *Drosophila*. Complex I deficits and adenosine triphosphate synthesis were also rescued in cells derived from *PINK1* patients. Thus, this evolutionary conserved pathway may contribute to the pathogenic cascade that eventually leads to Parkinson's disease in patients with *PINK1* mutations.

Mutations in *PINK1*, a mitochondrial targeted Ser/Thr kinase, cause a monogenic form of Parkinson's disease (PD) (1, 2). Loss of *PINK1* function mutations interfere with Parkin-mediated carbonyl cyanide m-chlorophenyl hydrazone (CCCP)-induced mitophagy (3–5) and mitochondrial fusion and fission de-

fects (4). However, an early and invariant phenotype of *PINK1* loss of function in different species is an enzymatic defect in mitochondrial complex I and a decrease in mitochondrial membrane potential ($\Delta\Psi_m$) (6–8). In contrast to effects of *PINK1* on toxin-induced mitophagy like CCCP (4, 5, 9), these complex I defects are observed in cell culture and *Drosophila* neurons under resting conditions with normal-appearing mitochondria (6, 8). Because *Pink1*^{-/-} mice display only subtle, and somewhat controversial, phenotypes of altered mitochondrial morphology (6, 8, 10, 11), it remains unresolved to what extent decreased mitophagy, or, alternatively, primary complex I deficiency, or both, are involved in those defects (6). In *pink1* and *parkin* *Drosophila* models (12–14), phenotypes are more pronounced. Thorax muscle degeneration and flight deficits can be rescued by expression of the fission-promoting gene *drp1* or by ablating the fusion-promoting gene *opa1* (15, 16), linking these molecules to the role of *PINK1* and Parkin in fusion and fission defects.

Intriguingly, other *pink1*-related phenotypes, such as defective neurotransmitter release, adenosine triphosphate (ATP) depletion, and loss of $\Delta\Psi_m$, cannot be rescued efficiently in *Drosophila* neurons by fission gene *Drp1* (17, 18) but can be rescued by genes restoring the proton motive force (19) or by NDi1, a yeast rotenone-insensitive reduced form of nicotinamide adenine dinucleotide (NADH)-quinone oxidoreductase (17). This suggests that two parallel molecular pathways are affected by *pink1* deficiency in flies, and both could be relevant to our understanding of the role of *PINK1* in PD.

First, we confirmed the pathological relevance of the previously reported $\Delta\Psi_m$ defects in *Pink1*^{-/-} mice and *Drosophila* using human fibroblasts and two induced pluripotent stem (iPS) cell lines derived from PD patients with *PINK1* mutations. Fibroblasts contained homozygous p.Q456X nonsense (L2122) or p.V170G missense (L1703) mutations (20), and iPS cells were derived from two PD patients with c.1366C>T; p.Q456X nonsense (L2122 and L2124) mutations. Integrity of the mitochondrial-targeted red fluorescent protein-labeled mitochondrial network was qualitatively and quantitatively (fragmented versus elongated) not different between control (L2134 and L2132) and patient (L1703 and L2122) fibroblasts (fig. S1, A and B). $\Delta\Psi_m$ was significantly decreased in the patient fibroblasts as assessed by the electrochemical potentiometric dye tetramethyl rhodamine ethyl ester (TMRE) (fig. S1, C and D). Overall ATP content in these *PINK1* mutant fibroblasts was also decreased when compared with age-matched controls (fig. S1E) (20). In neuronal differentiated iPS cells (L2124 and L2122) (21), $\Delta\Psi_m$ and ATP content (fig. S1, F to H) were lowered compared with controls (L2134 and L2135), confirming that clinical mutations in the context of human cells and human neurons display similar deficits as cell lines derived from *Pink1*-null mice and flies.

Cells display a specific deficit in the enzymatic activity of complex I (6). Therefore, we immunocaptured complex I from isolated mouse mitochondria (Fig. 1, A and B) and obtained independent phosphoproteomes from three brain and three liver

¹VIB Center for the Biology of Disease, 3000 Leuven, Belgium.

²Center of Human Genetics, University Hospitals Leuven and Department of Human Genetics, KU Leuven, and Leuven Research Institute for Neuroscience and Disease (LIND), 3000 Leuven, Belgium. ³Department of Medical Protein Research, VIB, 9000 Ghent, Belgium. ⁴Department of Biochemistry, Ghent University, 9000 Ghent, Belgium. ⁵Laboratory of Clinical and Experimental Endocrinology, KU Leuven, 3000 Leuven, Belgium. ⁶Section of Clinical and Molecular Neurogenetics at the Department of Neurology, University of Lübeck, 23538 Lübeck, Germany. ⁷Wellcome Trust Centre for Mitochondrial Research, Institute of Ageing and Health, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. ⁸University College London, Institute of Neurology, Queen Square, London, UK.

*Corresponding author. E-mail: bart.destrooper@cme.vib-kuleuven.be (B.D.S.); vanessa.morais@cme.vib-kuleuven.be (V.A.M.)