Chapter 5

Phylogenetic analysis of extended MHC paralogous gene families

5.1 Introduction

The genome-wide survey presented in chapter 4 identified over 700 MHC paralogues with varying levels of confidence. Analysis of the distribution of the 82 L2- and L3-paralogues confirmed that there were paralogous regions on chromosomes 1q21.2-q25.3, 9q32-q34.3 and 19p13.3-p13.11. One of the most interesting and novel findings was that there are also paralogues scattered throughout the genome. However, the origin of these paralogues is not known. By definition paralogues have arisen by duplication of an ancestral gene, which can involve a chromosomal segment containing one or more genes (block duplication), an entire chromosome or the whole genome. One of the most useful approaches to study the history of paralogues is to reconstruct the evolutionary relationships using orthologous sequences.

These relationships are commonly represented by means of a phylogenetic tree using sequence data from a range of evolutionary distant organisms. A phylogenetic tree is simply a branching diagram in which each terminal element (e.g. a protein sequence) is linked only once to one or more other protein sequences, thus specifying a hierarchy. Trees can be rooted using a distantly related sequence, such as the *Drosophila* or amphioxus orthologue, and corresponds to a point at the base of a tree indicating the evolutionary direction. Internal branch points, or ‘nodes’, represent putative ancestors and are connected by ‘branches’. Two sequences that are very much alike will be located as neighbouring outside branches and will be joined to a common branch beneath them. Evolutionary trees can be constructed such that the length of a branch connecting two proteins is proportional to the number of residue differences in the sequences. Thus, the object of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.

The paralogues located in the paralogous regions on chromosomes 1, 9 and 19 are believed to be remnants of two rounds of large-scale duplication events involving the whole genome early in vertebrate history. This phenomenon is referred to as the 2R hypothesis (Sidow, 1996). The first whole-genome duplication event occurred after an 'amphioxus stage' prior to the divergence of Agnatha (jawless vertebrates, represented by lamprey and hagfish) and Gnathostomata (jawed vertebrates), while a second occurred after the divergence of Agnatha but before the divergence of cartilagenous fish (represented by sharks) (summarised in figure 5.1).

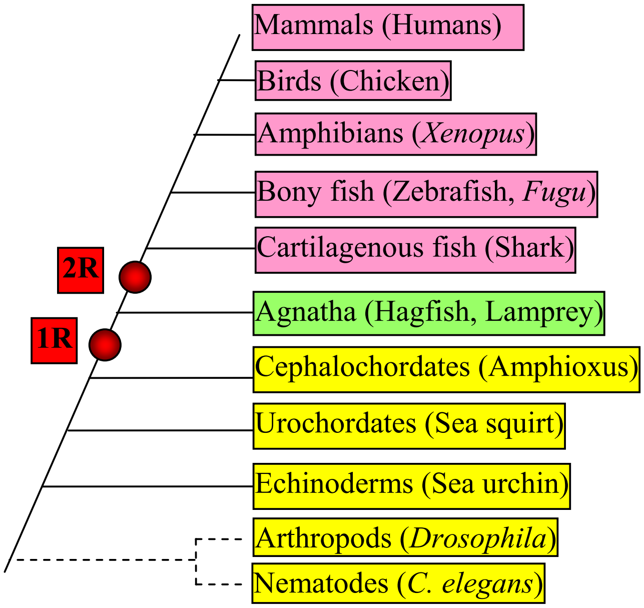
Mammals (Humans)
Birds (Chicken)
Amphibians (*Xenopus*)
Bony fish (Zebrafish, *Fugu*)
Cartilagenous fish (Shark)
2R
Agnatha (Hagfish, Lamprey)
1R
Cephalochordates (Amphioxus)
Urochordates (Sea squirt)
Echinoderms (Sea urchin)
Arthropods (*Drosophila*)
Nematodes (*C. elegans*)

Figure 5.1 Summary of the 2R hypothesis. The first round of genome duplication (1R) occurred after the emergence of amphioxus (yellow), prior to agnatha divergence (green) and the second occurred after the divergence of agnatha but before the divergence of cartilagenous fish (pink). The two duplication events are represented by red circles.

If two or more genes have been duplicated simultaneously as the result of a block duplication event, this should be revealed by phylogenetic analysis. If the 2R hypothesis is correct (assuming no genes have been lost since duplication) four paralogous genes should be found in humans and other jawed vertebrates, such as mice and chickens. Jawless fish, such as hagfish, should only have two paralogous genes, which are considered orthologous to the four paralogues in jawed vertebrates and the cephalochordate will have only one; corresponding to the closest relative of the 'ancestral gene' (figure 5.2). The branching pattern of the phylogenetic tree should be representative of the duplication events showing the double-forked tree topology, or the so-called 2+2 or (A,B)(C,D) topology. The age of the split of AB and CD is the same thus showing the history of successive rounds of duplication (summarised in figure 5.2).

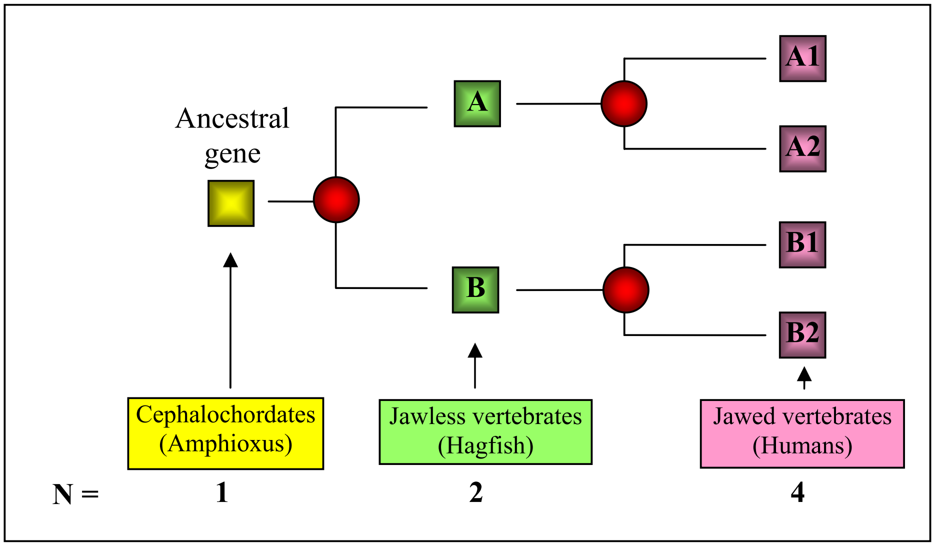
Ancestral gene
A
B
A1
A2
B1
B2
Cephalochordates (Amphioxus)
Jawless vertebrates (Hagfish)
Jawed vertebrates (Humans)
N =
1
2
4

Figure 5.2 Schematic representation of the effects of two rounds of gene, or genome, duplication on the topology of the phylogenetic tree (A1,A2)(B1,B2) and the resulting number of (N=) paralogues in 'key' species (1:2:4 ratio between amphioxus:hagfish:humans).

In this chapter, I present the phylogenetic analyses of ten paralogous gene families in order to determine the mechanism(s) by which they arose. Figure 5.3 summarises the topology of the phylogenetic tree expected if the paralogues arose from a common ancestor via two rounds of genome duplication (i.e. support the 2R hypothesis).

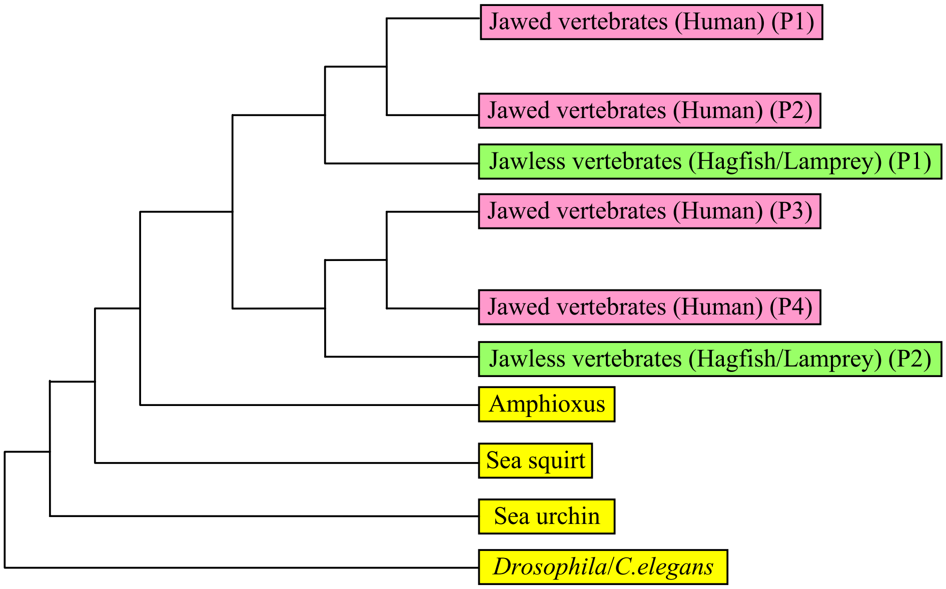
Jawed vertebrates (Human) (P1)
Jawed vertebrates (Human) (P2)
Jawless vertebrates (Hagfish/Lamprey) (P1)
Jawed vertebrates (Human) (P3)
Jawed vertebrates (Human) (P4)
Jawless vertebrates (Hagfish/Lamprey) (P2)
Amphioxus
Sea squirt
Sea urchin
Drosophila/*C.elegans*

Figure 5.3 Schematic representation of the 'ideal' phylogenetic tree in support of the 2R hypothesis. The species are colour coded according to the number of expected paralogues; species with one copy are highlighted in yellow, two copies (P1-2) in green and four paralogues (P1-4) in pink. It is important to note that in the phylogenetic trees presented in this chapter the species zebrafish, *Fugu* and *Xenopus* are highlighted pink, since they are jawed vertebrates, but they are expected to have more than four paralogues as an additional genome duplication event has occurred in their lineage.

5.2 MHC paralogous gene families used in phylogenetic analysis

In order to understand the evolutionary history of the MHC paralogues, 10 MHC genes and their paralogues (termed paralogous gene families) were selected for further analysis (summarised in figure 5.4).

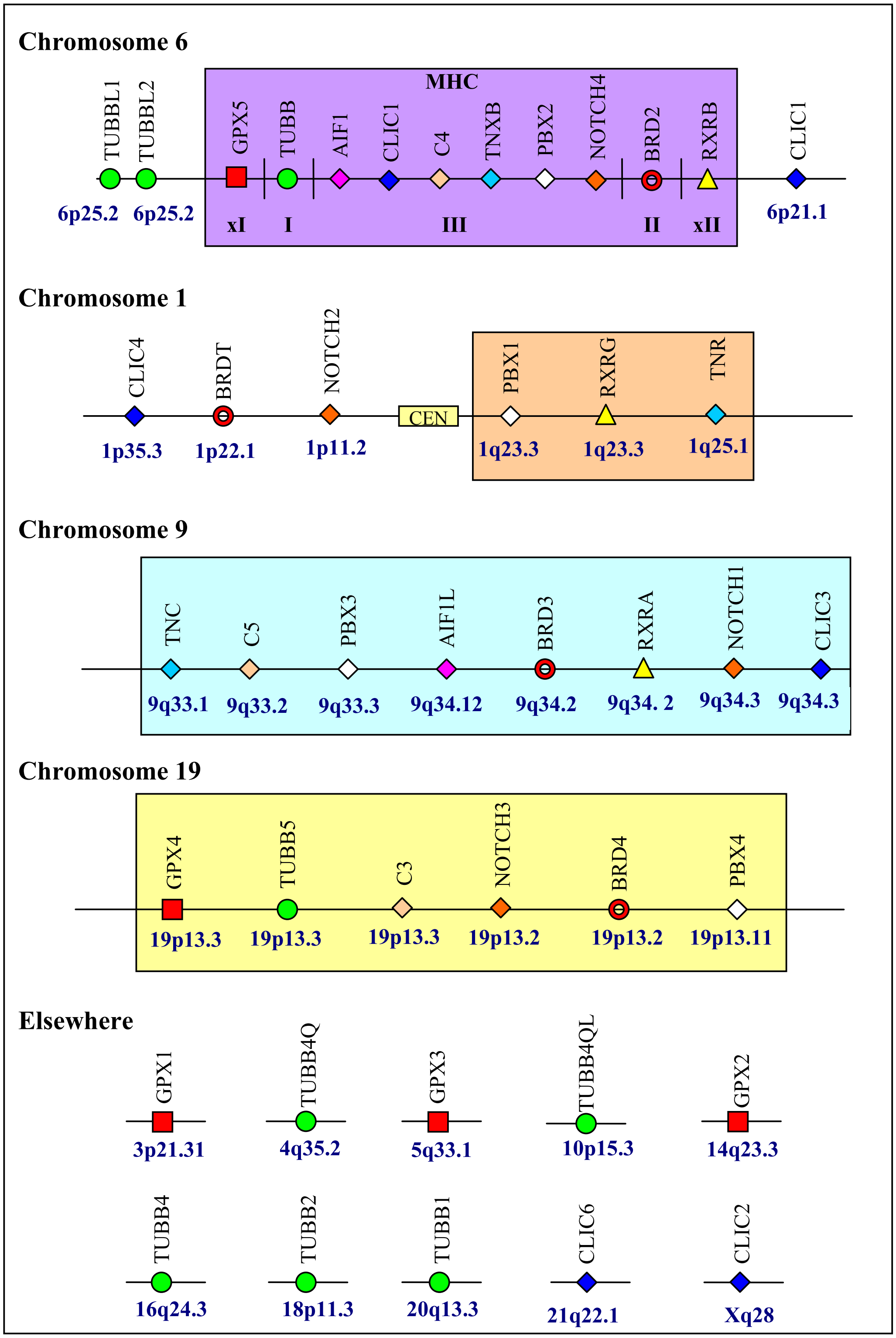

Figure 5.4 Summary of the MHC genes and paralogues selected for further investigation. The MHC genes and corresponding paralogues are represented by a shaded symbol. The cytogenetic locus for each gene on chromosomes 1, 9 and 19 is shown in blue text. 'CEN' corresponds to the centromere. The shaded areas correspond to the paralogous regions as defined in chapter 4.

The 10 paralogous gene families presented in this chapter were selected in order to satisfy a number of criteria. Firstly, the families were chosen to ensure that each of the five classes of the MHC region were represented by at least one paralogous gene family. Secondly, there were families with L2- and L3-paralogues located within the gene clusters on 1q21-q25, 9q32-q34.3 and 19p13.3-p31.3 only and, finally, there were also families with paralogues located elsewhere in the human genome.

5.3 Results

The protein sequences corresponding to the orthologues and paralogues of the 10 MHC genes were identified by searching the annotated protein databases and literature. The protein sequences were aligned with the ClustalW program using default parameters, and edited in Jalview. The sequence similarity between the MHC paralogues showed varying levels of divergence and, it was found that, the sequence alignments were often only reliable for conserved regions of the proteins. Therefore, in most circumstances, only these conserved regions were used to generate the trees. However, in cases, such as the TUBB family, where the sequence identity is very high (between 72.9 and 99.6%), the full length protein sequences were used. The number of sequences and protein regions used to produce the trees is summarised in table 5.1.

Table 5.1 Summary of the MHC paralogous gene families used to generate phylogenetic trees. The first three columns show the MHC gene locus, the location within the MHC region and the location of their paralogues, respectively. The remaining four columns, from left to right, show the number of sequences, the gamma-distribution alpha-parameter (α), number of amino acid (aa) residues and a description of the protein region used to generate the trees. The alpha-parameter is a measure of the rate of heterogeneity or change between amino acid sites (as described in 2.19.2). PR stands for paralogous region.

MHC Locus	*MHC class*	*Location of paralogues*	*No. of sequences used*	*α*	*aa residues used/length*	*Description of protein region used*
GPX5	xI	Outside PRs	22	1.09	221/221	Complete sequence
TUBB	I	Inside and outside PRs	27	0.29	444/444	Complete sequence
AIF1	III	In PR	11	1.21	92/147	Includes EF-hand domain
CLIC1	III	Inside and outside PRs	13	1.82	232/241	Most of sequence
C4	III	In PRs	27	1.42	1275/1744	Includes anaphylatoxin and macroglobulin domains
TNXB	III	In PRs	14	1.28	309/4289	Includes a fibronectin III domain and the fibrinogen c-terminal.
PBX2	III	In PRs	13	0.33	180/430	Includes homebox
NOTCH4	III	In PRs	17	0.97	385/2003	Includes 11 EGF-like domains
BRD2	II	In PRs	14	0.74	113/801	Includes a bromodomain
RXRB	xII	In PRs	20	0.23	313/533	Includes the DNA binding domain.

The phylogenetic trees presented in this chapter, unless otherwise stated, are a consensus of four trees generated using the three software packages: PHYLIP, MEGA2 and PUZZLE (as described in section 2.19). In each tree, the number on the branches of the tree correspond to the average percentage bootstrap or puzzling-step confidences from the three software packages. It should be noted that the protein names for all species, apart from human, are given in lower case.

5.3.1 Phylogenetic analysis of the BRD paralogous gene family

The BRD2, or the bromodomain containing 2, gene is located in the MHC class II region (Beck *et al*, 1992b). Denis and Green (1996) discovered that the RING3 product is a mitogen-activated nuclear kinase involved in signal transduction and that it is upregulated in certain types of leukeamia. In total, three paralogues of the BRD2 gene have been identified in the human genome with the highest level of confidence; these are BRDT on 1p22.3, BRD3 on 9q34.2 and BRD4 on 19p13.12. They all belong to the BET subgroup of bromodomain proteins and contain two bromodomains and an ET (or extraterminal) motif, which is a protein-protein interactive surface. The precise function of the bromodomains is unclear but it may be involved in protein-protein interactions and may play a role in assembly or activity of multi-component complexes involved in transcriptional activation (Tamkun, 1995).

The topology of the phylogenetic tree of the BRD paralogous gene family is (BRD2,BRD3)(BRD4,BRDT), thus supporting the 2R hypothesis (figure 5.5). Phylogenetic analysis shows that the timings of the two duplication events occurred after the divergence of cephalochordates and prior to the emergence of jawless fish. This is indicated by the positions of the amphioxus and hagfish orthologues on the tree. Overall, the phylogenetic analysis of the BRD2 paralogues and othologues shows that the BRD paralogous gene family arose by two rounds of duplication, but it should be noted that some branches show low levels of support.

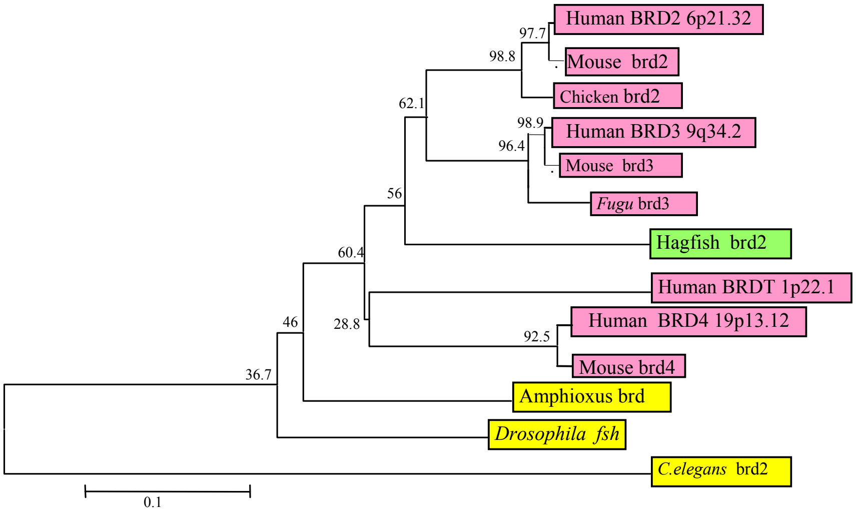

Figure 5.5 Phylogenetic tree of the BRD paralogous and orthologous family. The accession numbers are: P25440 (human BRD2), Q15059 (human BRD3), O14789 (human BRDT), O60885 (human BRD4), O54795 (mouse brd2), Q8K2F0 (mouse brd3), Q9ESU6 (mouse brd4), Q90971 (chicken brd2), Q8QFT7 (*Fugu* brd3), Q8T775 (amphioxus brd), P13709 (*Drosophila fsh*) and Q20948 (*C.elegans* brd2).

5.3.2 Phylogenetic analysis of the PBX paralogous gene family

The PBX2 (pre-B-cell leukaemia 2) gene encodes a homeodomain-containing protein. It was first identified on the basis of the extensive homology to the PBX1 gene involved in t(1;19)(q23;p13.3) translocation in acute pre-B-cell leukaemias (Monica *et al*, 1991). The genome survey identified three paralogues located within the paralogous regions on 1q23.3, 9q33.3 and 19p13.11, named PBX1, PBX3 and PBX4, respectively. Phylogenetic analysis shows that the PBX paralogous gene family arose by two rounds of duplication (figure 5.6). The topology of the tree is (PBX2, PBX4)(PBX1, PBX3), which supports the 2R hypothesis. The timings of the two duplication events can be determined as occurring after the divergence of cephalochordates and prior to the emergence of jawed fish, indicated by the positions of the amphioxus and zebrafish orthologues on the tree.

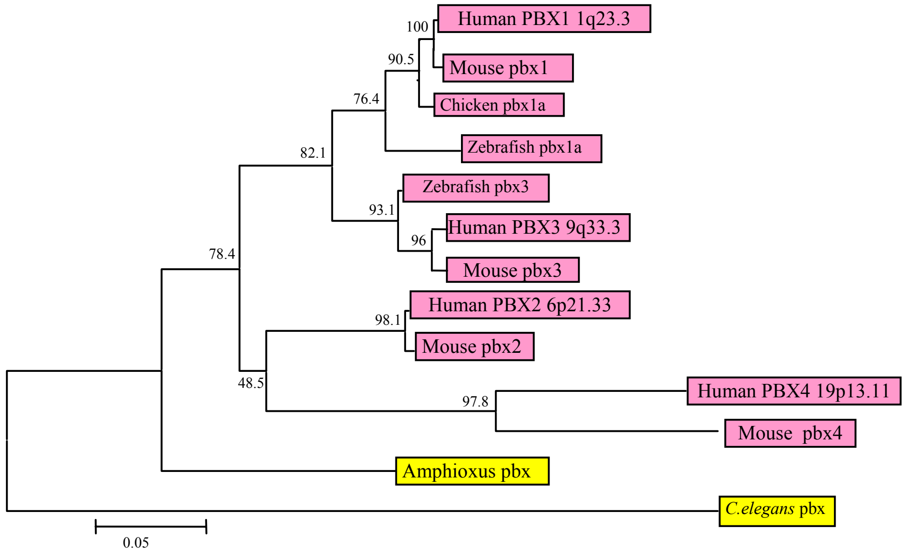

Figure 5.6 Phylogenetic analysis of the PBX paralogous gene family. The accession numbers of the vertebrate protein sequences used are: P40425 (human PBX2), P40424 (human PBX1), P40426 (human PBX3), Q9BYU1 (human PBX4), O35984 (mouse pbx2), P41778 (mouse pbx1), O35317 (mouse pbx3), Q99NE9 (mouse pbx4), Q9IB15 (chicken pbx1a), Q9I9B7 (zebrafish pbx1a), Q9I9B5 (zebrafish pbx3), AF39192_1 (amphioxus pbx) and P41779 (*C.elegans* pbx).

5.3.3 Phylogenetic analysis of the NOTCH paralogous gene family

The Notch gene was first identified in *Drosophila* as a regulator of cell fate determination and has been implicated in a large number of developmental processes in *Drosophila* and vertebrate systems (reviewed by Bray, 1998; Lewis, 1998). The phylogenetic tree using 14 vertebrate protein sequences and three invertebrate protein sequences is presented in figure 5.7.

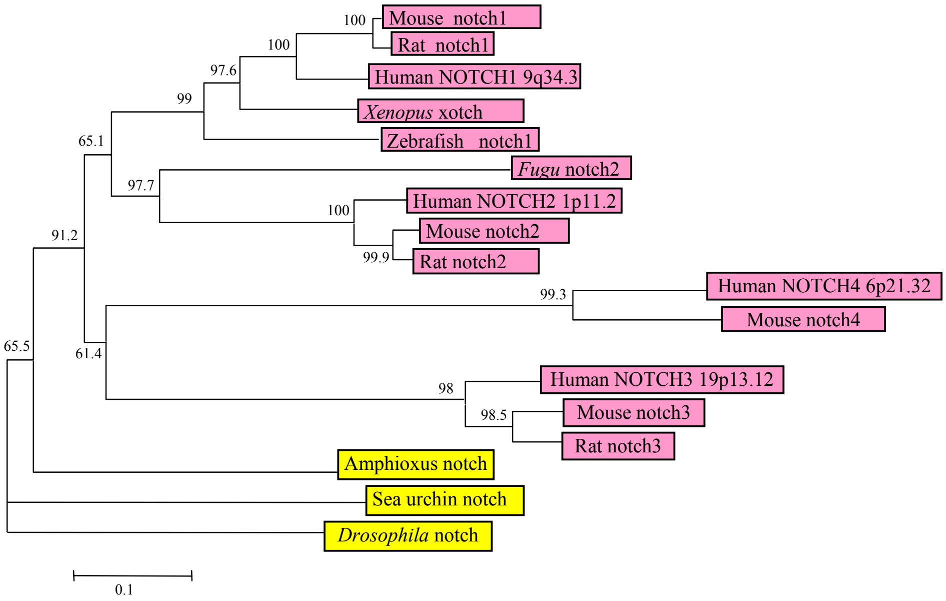

Figure 5.7 Phylogenetic analysis of the NOTCH paralogous gene family. The protein sequences, with the accession numbers given in parentheses, are: human NOTCH4 (O00306), human NOTCH1 (P46531), human NOTCH2 (Q04721), human NOTCH3 (Q9UM47), mouse notch4 (P31695), mouse notch1 (Q01705), mouse notch2 (O35516), mouse notch3 (Q61982), rat notch1 (O07008), rat notch2 (Q9QW30), rat notch3 (Q9R172), X*enopus* xotch (P21783), zebrafish notch1 (P46530), *Fugu* notch2 (O13149), amphioxus notch (Q9GPA5), sea urchin notch (O16004) and *Drosophila* notch (P07207).

In *Drosophila* and lower deuterostomes (such as sea urchins) there is a single Notch gene, while in vertebrates there are multiple Notch genes (four in humans and mouse). Phylogenetic analysis of the NOTCH4 paralogues and orthologues supports the hypothesis that NOTCH1-4 arose from a common ancestor via two duplication events. The single amphioxus notch protein branches at the base of the four vertebrate Notch proteins. Together with the presence of single Notch gene in the sea urchin it suggests that Notch duplicated within the vertebrate lineage. Thus, both duplications occurred after the divergence of amphioxus and prior to the divergence of bony fish and tetrapods.

5.3.4 Phylogenetic analysis of the complement paralogous gene family

The C4 gene is located in the MHC class III region and encodes the complement factor 4 protein. C4 plays a central role in the activation of the classical pathway of the complement system. The complement system is the principle effector mechanism of humoral immunity and consists of at least 24 serum proteins and 11 membrane-bound proteins. The interaction of these proteins leads to a complement cascade and results in a number of responses, including cell lysis, opsonisation of targets for phagocytosis by macrophages, regulation of B cell responses and the generation of potent anaphylatoxins (for review see Reid and Porter, 1981). Two paralogues of the C4 gene have been identified in the human genome; these are C5 located on 9q33.2 and C3 located on 19p13.3.

Phylogenetic analysis of the full length C3, C4 and C5 protein sequences (figure 5.8) supports the view that C5 diverged first with C3 and C4 subsequently diverging before the separation of jawed and jawless fishes (Hughes, 1994). The presence of C3 in jawless deuterostomes, such as sea urchin (Smith *et al*, 1999), hagfish (Ishiguro *et al*, 1992) and lamprey (Nonaka and Takashii, 1992) enables the divergence times of the complement genes to be determined and establishes the ancient origin of the complement system. The clustering of the lamprey and hagfish C3 with the other vertebrate C3 proteins clearly indicates that the duplication of the C3 and C4 genes from the ancestral gene occurred after the divergence of jawless vertebrates and prior to the divergence of jawed vertebrates. One would expect the orthologue of C4 to be revealed upon full sequencing of the hagfish and lamprey genomes. The divergence of C5 occurred after the cephalocordate split prior to the divergence of jawless fish. Phylogenetic analysis shows a ratio of 1:2:2:3 Amphioxus:Hagfish:Lamprey:Human.

Thus, supporting the 2R hypothesis, accompanied with the loss of one of the C4 duplicate giving the topology (C5(C4, C3)) rather than the predicted (A,B)(C,D).

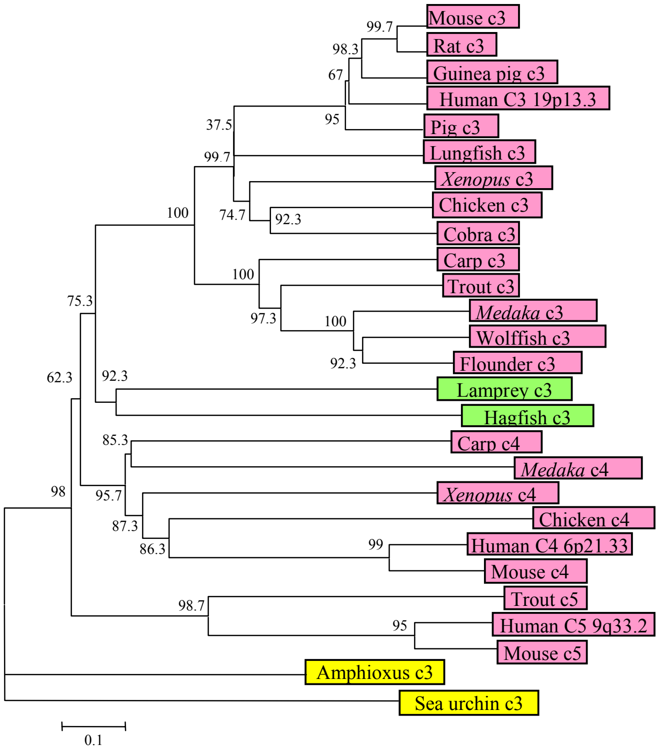

Figure 5.8 Phylogenetic analyses showing the relationship of the C4 paralogues and orthologues. The accession numbers of the proteins used to generate the trees are: human C4 (P01028), mouse c4 (P01029), chicken c4 (O73905), *Xenopus* c4 (Q91741), *Medaka* c4 (Q9IBG9), carp c4 (Q9I933), human C5 (P01031), mouse c5 (P06684), trout c5 (Q90XS7), human C3 (P01024), mouse c3 (P01027), chicken c3 (Q90633), *Xenopus* c3 (Q91588), rat c3 (P01026), guinea pig c3 (P12387), pig c3 (Q9GKP1), cobra c3 (Q01833), lungfish c3 (Q9W6G1), carp c3 (Q9YIB0), trout c3 (P98093), *Medaka* c3 (Q9IBH1), wolffish c3 (Q98TS6), flounder c3 (Q9PTY1), lamprey c3 (Q00685), hagfish c3 (P98094), amphioxus c3 (Q969A4) and sea urchin c3 (O44344).

5.3.5 Phylogenetic analysis of the RXR paralogous gene family

The retinoid X receptor beta, or RXRB, protein is a retinoid receptor and belongs to the steroid/thyroid hormone receptor superfamily of transcriptional regulators (Mangelsdorf *et al*, 1992). Retinoid receptors are soluble nuclear proteins that fall into two classes: retinoic acid receptors (RAR) and retinoid X receptors (RXR). The RXR subfamily consists of three polypeptide chains, namely alpha, beta and gamma, encoded by separate loci. The three loci encoding the alpha (RXRA), beta (RXRB) and gamma (RXRG) proteins are located on chromosomes 9, 6 and 1, respectively.

The RXR phylogenetic tree was rooted with the *Drosophila* orthologue, usp (figure 5.9).

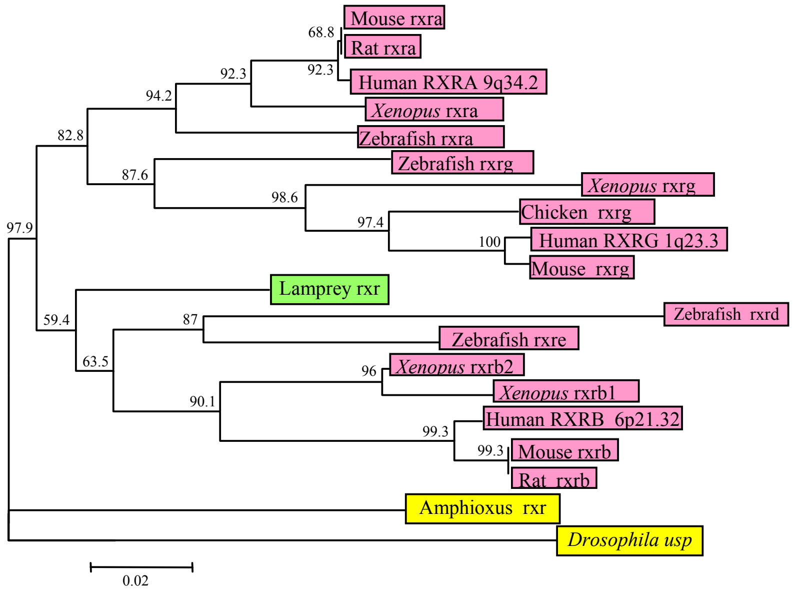

Figure 5.9 Phylogenetic tree showing the evolutionary relationship between the RXRB paralogues and orthologues. The accession numbers are: X52773 (human RXRA), X63522 (human RXRB), U38480 (human RXRG), M84817 (mouse rxra), M84818 (mouse rxrb), M84819 (mouse rxrg), L06482 (rat rxra), M81766 (rat rxrb), X58997 (chicken rxrb), L11446 (*Xenopus* rxra), X87366 (*Xenopus* rxrb2), S73269 (*Xenopus* rxrb1), L11443 (*Xenopus* rxrg), U29940 (zebrafish rxra), U29894 (zebrafish rxrg), U29941 (zebrafish rxrd), U29942 (zebrafish rxre), AF316878 (lamprey rxr), AF391296/5 (amphioxus rxr) and P20153 (*Drosophila* usp).

The results show that the human paralogues cluster, as expected, with equivalent orthologues. The RXR orthologues of the invertebrate species, *Drosophila* and amphioxus, both fall outside all of the vertebrate genes. However, the RXR orthologue of the invertebrate lamprey clusters with vertebrate RXRB. This indicates that RXRB diverged first, after the cephalochordate split prior to the divergence of jawless fish. This was followed by a duplication event between RXRA and RXRG after the divergence of jawless fish. The zebrafish RXRD and RXRE genes resulted from a duplication occurring around the time of teleost/mammalian divergence. The topology of the tree, (RXRB(RXRA, RXRG)), clearly supports at least one round of large-scale duplication but it is possible that the RXR family arose by two-rounds of large-scale duplication events and one paralogue has been lost over time. Thus, the present day topology is (RXRB (RXRA, RXRG)).

5.3.6 Phylogenetic analysis of the tenascin paralogous gene family

The tenascin proteins are a family of extracellular matrix proteins (ECM) (for a review see Erickson, 1993). The Tenascin X (TNX) gene is located within the MHC class III region overlapping the CYP21A2 and C4 genes. Two paralogues, tenascin C (TNC, cytoactin, hexabrachion) and tenascin R (TNR, restrictin) have been identified in the paralogous regions on chromosomes 9q33.1 and 1q24.1, respectively. Tenascin orthologues have been identified in a range of vertebrates but only one invertebrate (summarised in the legend of figure 5.10). The Drosophila protein, ten^{m}, contains the EGF-like and FN-III domains and is believed to be the closest relative of the vertebrate tenascins (Baumgartner *et al*, 1994). This has been used as the outgroup to root the phylogenetic tree presented in figure 5.10.

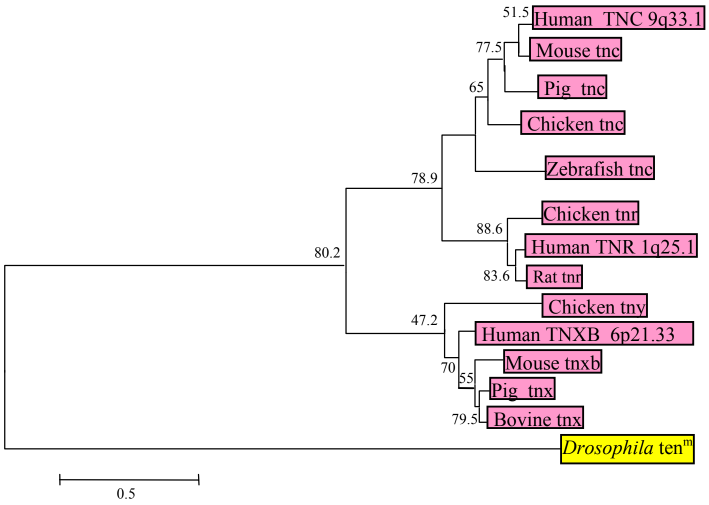

Figure 5.10 Phylogenetic analyses of the TNXB paralogues and orthologues. The accession numbers of the protein sequences used to generate this tree are as follows; P22105 (human TNXB), P24821 (human TNC), Q15568 (human TNR), O35452 (mouse tnxb), Q64706 (mouse tnc), Q05546 (rat tnr), P10039 (chicken tnc), Q00546 (chicken tnr), Q91008 (chicken tny), Q29038 (pig tnxb), Q29116 (pig tnc), O18977 (bovine tnxb) and Q24551 (*Drosophila* ten^{m}).

The topology of the tree strongly supports that TNXB diverged prior to the divergence of TNR and TNC as suggested by Katsanis and co-workers (1996) and Hughes (1998). Phylogenetic analysis shows that the TNC and TNR paralogues are most closely related and have arisen from a common ancestor. The clustering of the zebrafish TNC orthologue with the other TNC orthologous sequences indicates that the duplication which gave rise to the TNC and TNR paralogues occurred prior to the divergence of bony fish and tetrapods, approximately 450 million years ago. However, without the orthologous protein sequences of the key species (amphioxus, hagfish and lamprey) it cannot be determine whether the tenascin X gene supports the 2R hypothesis. Compelling evidence from the five other MHC paralogous gene families presented in sections 5.3.1-5.3.6 implies that these genes may have arisen via the same mechanism.

5.3.7 Phylogenetic analysis of the AIF paralogous gene family

The Allograft inflammatory factor 1 (AIF-1) gene was first isolated from activated macrophages in rat atherosclerotic allogenic heart grafts undergoing chronic transplant rejection (Utans *et al*, 1995). In humans, the full-length clone has been isolated and characterised (Autieri, 1996). Only one AIF-1 paralogue (AIF1-L) has been identified in the human genome (discussed in chapters 3 and 4) and is located in the chromosome 9 paralogous region.

The AIF1 encoded protein is evolutionarily well conserved within vertebrate species (Utans *et al*, 1996). To-date, it has been identified in seven vertebrates: humans, pig, rat, macaque, mouse, bovine, red sea bream and carp. It has only been identified in two invertebrates, the sea sponge and amphioxus. Phylogenetic analysis of the AIF1 paralogues and orthologues was carried out using the distantly related amino acid sequence from sea sponge as the outgroup (figure 5.11).

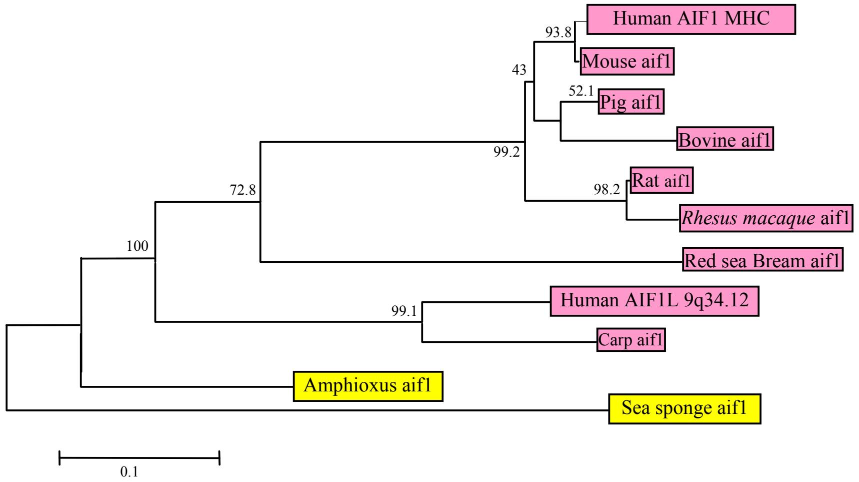

Figure 5.11 Phylogenetic tree of the AIF1 paralogues and orthologues. The species, and corresponding accession numbers given in parentheses, used to generate the tree are as follows: human AIF1 (P55008), human AIF1-L (Q9BQI0), pig aif1 (P81076), rat (P55009), *Rhesus macaque* aif1 (Q9GMH2), mouse aif1 (070200), bovine aif1 (Q9BDK2), red sea bream aif1 (Q9YI94), carp aif1 (O93246), sea sponge aif1 (Q966Y8) and aif1 amphioxus (translated from EMBL entry AU234552).

Evidence provided by the phylogenetic tree indicates that the duplication event involving the ancestral gene of the two AIF1 paralogues occurred prior to divergence of bony fish and post-dates the divergence of Amphioxus. Thus, this analysis supports at least one round of duplication prior to vertebrate emergence, or the 1R hypothesis. Interestingly, the AIF1 protein in carp clusters with AIF1-L in the phylogenetic tree suggesting that the carp AIF1 protein may actually be the orthologue of the human AIF1-L gene on chromosome 9. A sequence similarity search using the AIF1-L protein did not identify any orthologous sequences previously not identified for AIF1. To-date, a second AIF1 orthologue has not been identified in the carp genome to confirm that this is the true AIF1L orthologue. In summary, the AIF1 paralogous gene family have occurred via a large-scale duplication after the divergence of the cephalochordate lineage prior to the emergence of bony fish. Thus, supporting one round of large-scale duplication, or the 1R hypothesis.

5.3.8 Phylogenetic analysis of the **β-tubulin paralogous gene family**

The β-tubulins form the basic building blocks of the microtubulins when they form heterodimers with α-tubulins (reviewed by McKean *et al*, 2001). Microtubulins constitute a major component of the cytoskeleton in eukaryotic cells and are involved in essential processes, including cell division and intracellular transport. The survey of the human genome revealed seven paralogues of the TUBB gene scattered throughout the genome. The paralogues share very high sequence similarity, ranging from 72.9% to 99.6% at the protein level. The high level of similarity has resulted in the mis-annotation of these genes, i.e. the same SWISSPROT or SPTREMBL accession number has been given as the encoded protein sequence for multiple genes. In order to prevent confusion, the corresponding ENSEMBL accession numbers is given in table 5.2 that was identified in the genome survey.

Table 5.2 Summary of the TUBB paralogues in the human genome

Gene	***Locus***	***Genomic clone accession number***	***ENSEMBL gene ID***	***No. of amino acids***
TUBB	6p21.3	AB023051	ENSG00000137379	444
TUBBL1	6p25.2	AL031963	ENSG00000137267	445
TUBBL2	6p25.2	AL445309	ENSG00000137285	445
TUBB4QL	10p15.3	AL713922	ENSG00000173876	444
TUBB4	16q24.3	AC0092143	ENSG00000141037	442
TUBBL	18p11.3	AP001005	ENSG00000173213	433
TUBB5	19p13.3	AC010503	ENSG00000104833	444
TUBB1	20q13.3	AC109840	ENSG00000101162	451

The β-tubulin genes are extensively conserved evolutionarily, but the number of encoding genes varies dramatically among species (Lewis and Cowan, 1990). A search of the protein databases, SWISSPROT and SPTREMBL, revealed several vertebrate β-tubulin proteins; one chimpanzee, one squirrel monkey, one rhesus macaque, one baboon, two mouse, one rat, two chicken and three *Xenopus* β-tubulin proteins. In addition, seven invertebrate β-tubulin proteins were extracted from the database; two sea squirt, one sea urchin, two *Drosophila*, one *C.elegans* and one *C.briggsae*.

Phylogenetic analysis using the protein sequences encoded by the TUBB paralogues reveals evidence in support of a number of duplication events (figures 5.12 and 5.13). The phylogenetic tree presented in figure 5.12 reveals two main gene clusters; one including only new world monkey and human sequences and the other, also containing human sequences, including a number of more 'ancient' species, namely the sea squirt and sea urchin clustered with vertebrates.

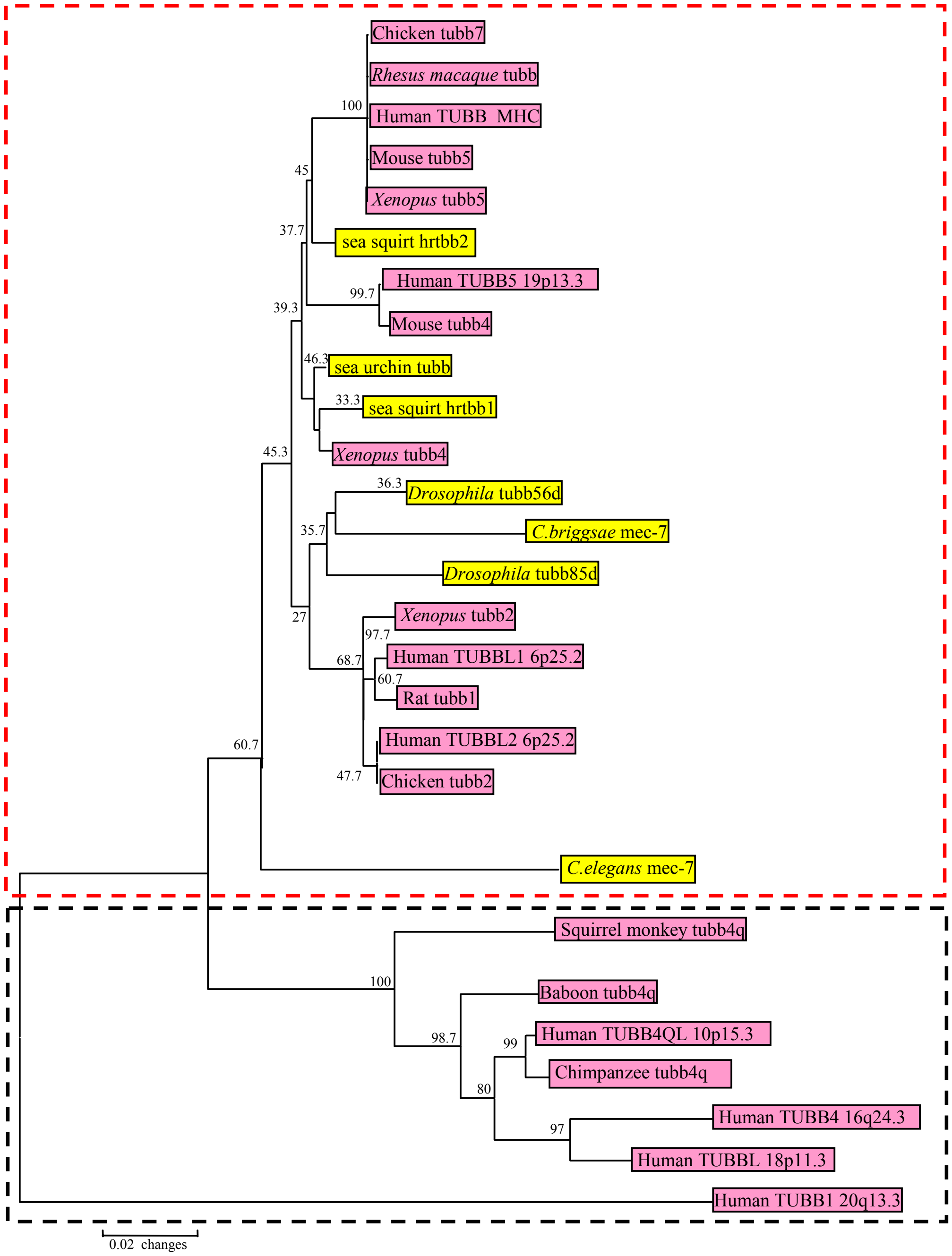

Figure 5.12 Phylogenetic analysis of the β-tubulin paralogues and orthologues. The two major species clusters corresponding to ancient duplication events (surrounded by a dashed red line) and more recent duplications (black dashed line). The accession numbers of the protein sequences used to generate this tree are: Q8WP14 (tubb4q, chimpanzee), Q8WP12 (tubb4q, squirrel monkey), AAD33992 (tubb4q, *Rhesus macaque*), Q8WP13 (tubb4q, baboon), Q9D6F9 (tubb4, mouse), P05218 (tubb5, mouse), P04691 (tubb1, rat), P32882 (tubb2, chicken) P09244 (tubb7, chicken), P13602 (tubb2, *Xenopus*), P30883 (tubb4, *Xenopus*), Q91575 (tubb5, *Xenopus*), O18343 (hrtbb2, sea squirt), O18342 (hrtbb1, sea squirt), P11833 (tubb, sea urchin), Q24560 (tubb56d, *Drosophila*), P08840 (tubb85d, *Drosophila*), P12456 (mec-7, *C.elegans*) and Q17299 (mec-7, *C.briggsae*).

The cluster containing the new world monkey β-tubulin proteins indicates that some of the tubulin paralogues in the human genome are the result of recent duplication events. This is supported by the analysis of the TUBB4Q pseudogene on 4q35.2 and related paralogues and orthologues by van Geel and co-workers (2002). Analysis of the human chromosomal segment, 4q35, containing the TUBB4Q pseudogene has indicated a substantial amount of duplication throughout the genome (Grewal *et al*, 1999; van Geel *et al*, 1999). Van Geel and colleagues (2002) revealed that this segment has undergone a number of duplications at different time points within the last 25 million years of catarrhine (New World Monkeys and humans) evolution.

The phylogenetic tree presented in figure 5.13 reveals evidence of ancient duplication events which occurred earlier than those proposed by the 2R hypothesis. The timings of the two rounds of duplication are proposed as follows; the first round of duplication occurred prior to the divergence of sea squirt and sea urchin and the second, after their divergence, prior to vertebrate emergence. The β-tubulin paralogues have been involved in a much earlier round of duplication followed by further duplications; this is supported by the clustering of the sea squirt and sea urchin orthologues with the mammalian counterparts.

There is evidence of a more recent duplication event telomeric to the MHC region on chromosome 6. The two newly identified paralogues, termed TUBBL1 and TUBBL2, on 6p25.2 share identical exon fingerprints and have 99.6% protein sequence similarity. The two paralogues are located within 70 kb of each other and span approximately 3.4 kb. They appear to have arisen by a tandem duplication event after amphibian divergence and prior to the emergence of rodents (the orthologue of TUBBL2 in rat may not be functional or the dataset used to generate the tree may not be complete).

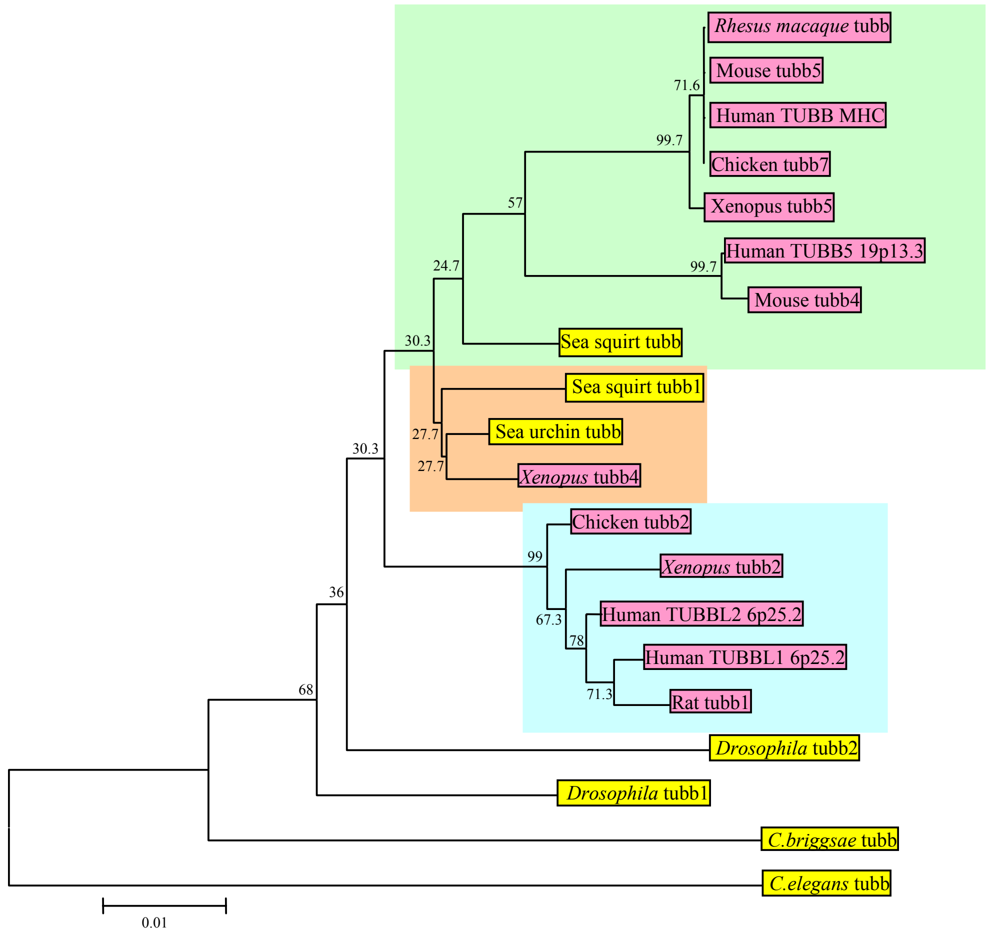

Figure 5.13 Phylogenetic tree showing the ancient duplication events that have shaped the present day β-tubulin paralogues and orthologues. The three main groups are highlighted in different colours. The protein accession numbers are as described in figure 5.12.

Analysis of the distribution of β-tubulin paralogues in the human genome (Chapter 4) reveals a strong positional bias towards the pericentromeric and subtelomeric regions of the genome. It is known that frequent exchange of sequences occurs between these dynamic chromosome regions (Eichler *et al*, 1996; Pryde *et al*, 1997; Eichler, 1998; IHGSC, 2001), which can result in the acquisition of new genes as well as genetic diversity. There is evidence to suggest that the TUBB4Q pseudogene on 4q35.2 was once a functional gene and, because of its proclivity to duplicate to subtelomeric locations, a novel tubulin member was transposed to 10p15.3 (TUBB4QL) approximately 7.3 MYA (van Geel *et al*, 2002). It has also been suggested that GC-rich repeat elements play a direct role in the pericentromeric localisation of intra- and interchromosomal duplication events (Eichler *et al*, 1999). It would be interesting to investigate whether there are GC-rich repeat elements bordering the duplicated segments containing the TUBB paralogues but this is beyond the scope of this thesis.

5.3.9 Phylogenetic analysis of the GPX paralogous gene family

The glutathione peroxidase proteins (GPX) are enzymes involved in the protection of the cell against oxidative damage. Using glutathione as the reducing agent they metabolise hydroperoxides generated by normal oxidative metabolism which otherwise would have deleterious effects, mainly on cell-wall integrity (Dufaure *et al*, 1996) phylogenetic relationship of the GPX5 paralogues and orthologues is shown in figure 5.14.

The human GPX protein sequences were aligned with the protein sequences obtained from a range of vertebrates and a single invertebrate species, *Suberites domuncula* also known as the sea sponge (see legend of figure 5.14). In total, one true paralogue of GPX5, named GPX3, was identified in the genome survey on 5q33.1. Two putative paralogues were also identified on chromosomes 3p21.31, GPX1, and GPX3, 5q33.1 but have a very different exon structure compared with GPX5. In addition, another member of the glutathione peroxidase family, GPX4, has been identified in the paralogous region on 19p13.3 but was not identified as a paralogue in this analysis.

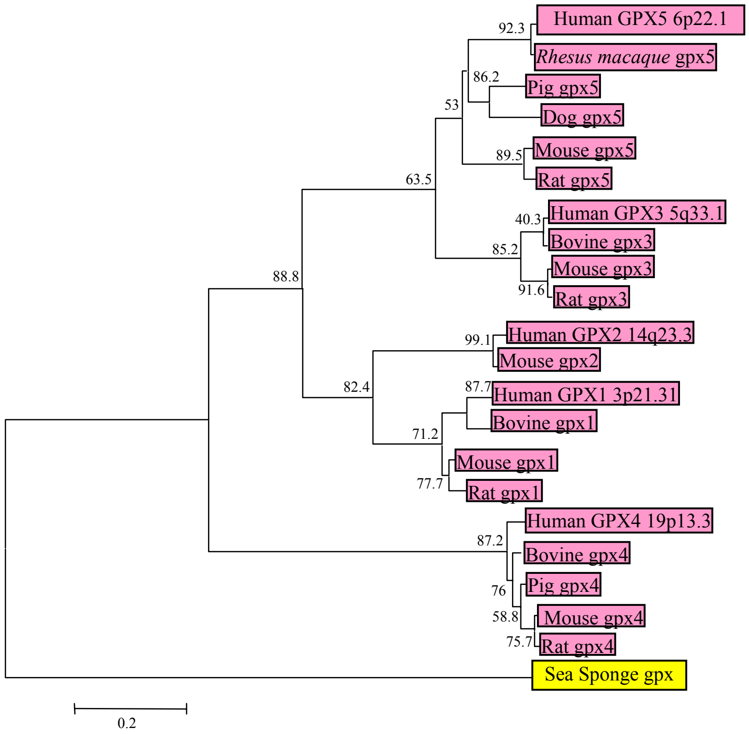

Figure 5.14 Phylogenetic analysis of the GPX family. The protein sequences used, with the corresponding accession numbers given in parentheses, were as follows; human GPX5 (O75715), human GPX3 (P22352), human GPX1 (P18283), human GPX2 (P18283), human GPX4 (P36969), mouse gpx5 (P21765), mouse gpx3 (P46412), mouse gpx1 (P11352), mouse gpx2 (Q9JHC0), mouse gpx4 (O70325), pig gpx5 (O18994), pig gpx4 (P36968), rat gpx5 (P30710), rat gpx3 (P23704), rat gpx1 (P04041), rat gpx4 (P36970), bovine gpx3 (P37141), bovine gpx1 (P00435), bovine gpx4 (Q9N2N2), dog gpx5 (O46607), *Rhesus macaque* gpx5 (P28714) and sea sponge gpx (Q966Y9).

Phylogenetic analysis shows that GPX4 is the most distantly related member. This is to be expected as it has a very different exon structure than the other GPX family members and shares less than 30% sequence identity. The GPX genes with identical exon fingerprints and highest protein sequence identity, GPX5-GPX3 and GPX1-GPX2, cluster together indicating that the genes have descended from a common ancestor. The duplications occurred after sea sponge divergence but prior to rodent divergence thus could have resulted from two rounds (or more) of whole-genome duplication. More data is needed to determine the precise times of the duplication events.

5.3.10 Phylogenetic analysis of the CLIC paralogous gene family

The chloride intracellular channel (CLIC) paralogous gene family encode for chloride channels, which are involved in chloride ion transport within various subcellular compartments (reviewed by Jentsch *et al*, 2002). Phylogenetic analysis of the CLIC family suggests a series of successive duplication events including at least 2 rounds of whole-genome duplications (figure 5.15).

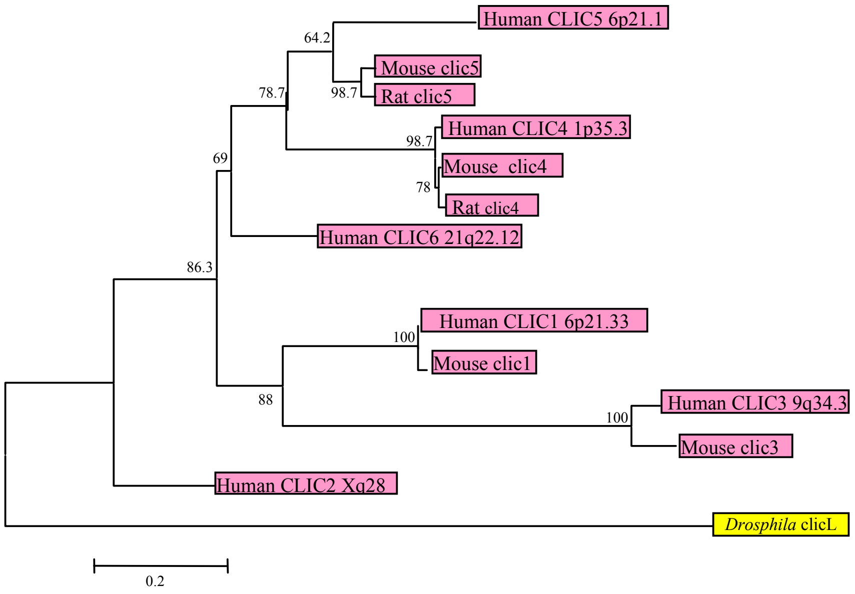

Figure 5.15 Phylogenetic analysis of the CLIC family. The protein sequences used, with the corresponding accession numbers given in parentheses, were as follows; human CLIC1 (O00299), human CLIC2 (O15247), human CLIC3 (O95833), human CLIC4 (Q9Y696), human CLIC5 (Q9NZA1), human CLIC6 (Q9NY7), mouse clic1 (Q9Z1Q5), mouse clic3 (Q9D7P7), mouse clic4 (Q9QYB1), mouse clic5 (Q9CYD1), rat clic4 (Q9Z0W7), rat clic5 (Q9EPT8), *Drosophila* clicL (NM_132700).

The topology of the tree indicates that the Xq28 CLIC2 gene diverged first. Following the divergence of CLIC2 there were two rounds of whole-genome duplication. The first duplication event resulted in the CLIC (MHC,9q34.3) and the CLIC (21q22.12(1p35.3,6p21.1)) gene precursors. This was followed by a second round of whole-genome duplication resulting in CLIC1 (MHC), CLIC3 (9q34.3), CLIC6 (21q22.12) and the CLIC (1p35.3, 6p21.1) precursor (possibly located on 19p13). A further segmental duplication event occurred resulting in the present day location of the CLIC4 and CLIC5 genes on 1p35.3 and 6p21.1, respectively. The latter segmental duplication is supported by the observation that there has been a large-scale triplication involving the chromosomal regions 1p35, 6p21.1 and 21q22.12 (Strippoli *et al*, 2002). Strippoli and colleagues (2002) identified a large (approximately 500 kb) segment on human chromosome 21q22 that is triplicated on chromosomes 1p35 and 6p12-p21. The region on chromosome 21 contains the CLIC6 gene, along with two other genes, DSCR1 and AML1, which have functional copies in the other regions. The gene order within these regions, termed the ACD clusters, is identical and it was suggested that the triplication occurred by segmental duplication as part of the genome-duplication events before the divergence of tetrapods and teleosts. However, more sequence data is needed to confirm this prediction and to fully understand the complex history of this paralogous gene family.

5.4 Discussion

The evolutionary histories of 10 MHC paralogous gene families have been reconstructed using phylogenetic trees. Analysis of the topologies of the trees and the arrangement of the paralogues and orthologues has revealed that the evolution of the MHC paralogues is complex. What is evident is that gene duplication has played a major role in the evolution of these gene families. In particular, there is evidence in support of the 2R hypothesis. The 2R hypothesis proposes that the genome evolved via two rounds of whole-genome duplication events early in the vertebrate lineage; one occurring after amphioxus divergence, prior to the emergence of hagfish and lamprey and the second just after. In order to support the 2R hypothesis, phylogenetic analyses of gene families should meet the following criteria: (a) the vertebrate members of the gene family can be shown to have duplicated within the vertebrate lineage and (b) the gene family phylogenies show the (A,B)(C,D) topology.

The 2R hypothesis would give rise to four copies of an ancestral gene therefore this is best exemplified by the paralogous gene families with four members in the human genome. The BRD, NOTCH and PBX paralogous gene families all have four paralogues, including the MHC locus, in the human genome. The topology of the three phylogenetic trees support the 2R hypothesis, showing the (A,B)(C,D) topology. Furthermore, some of the orthologous genes have been identified in the three key organisms, amphioxus, lamprey and hagfish, and the positions of these organisms in the phylogenetic trees are in support of the timings of the duplication events proposed by the 2R hypothesis. Thus, if the sequences are available, a single amphioxus orthologue is positioned at the base of each tree and at least one hagfish or lamprey orthologue clusters with the mammalian counterparts.

The paralogous gene families with three members also support the 2R hypothesis, albeit accompanied by gene loss. This appears particularly likely as extensive gene loss has been shown to take place after gene duplication events (Gu and Huang, 2002). Furthermore, the paralogous gene family with only two members is also in support of at least one round of genome duplication in the vertebrate history (the 1R hypothesis). Alternatively, it also supports the 2R hypothesis accompanied with the loss of two genes. The timings of the duplication events as suggested by the 2R hypothesis are also supported by the clustering of the 'key' organisms in these phylogenetic trees.

Ideally, if the paralogues emerged simultaneously by block or whole-genome duplication, the genes from the same chromosomal regions should cluster together on the tree. For example, previously published phylogenetic analyses of three paralogous gene families indicated that the paralogous regions on 1q21-q25 and 9q33-q34 were most related (Katsanis *et al*, 1996; Kasahara, 1997; Hughes, 1998). It would therefore be expected that the paralogues on chromosomes 1 and 9 will cluster and 6 and 19 will also cluster. However, this is not the case. The NOTCH and PBX paralogous gene families support this clustering however the BRD paralogues do not; with the BRD paralogues on chromosomes 6 and 9, and, 1 and 19 clustering. Since the construction of phylogenetic trees utilises sequence information the different rates by which the sequences of the paralogues have evolved since duplication, dictated by the evolutionary pressures acting upon them, may explain why different sets of paralogues cluster.

Phylogenetic analysis of the MHC paralogous gene families with five or more members revealed that the evolution of the MHC paralogues involved more than just the two rounds of large-scale duplication events proposed by the 2R hypothesis. This is exemplified by the β-tubulin family, which shows evidence of both ancient duplication events, dated prior to the divergence of sea squirt and sea urchin prior to the emergence of amphioxus, as well as much more recent duplications. This is in concordance with previously published phylogenetic studies of the MHC paralogous gene families (Endo *et al*, 1997; Hughes; 1998). These studies revealed that duplication events of multigene families, such as the proteasome component (PSMB) genes, occurred much earlier than those proposed by the 2R hypothesis.

In conclusion, there is strong evidence that some MHC paralogues evolved via the mechanism proposed by the 2R hypothesis but that others have emerged by independent means. Therefore, there could still be a selective advantage, potentially related to function, for these genes to have been brought together and remained clustered.