Molecular evolution of cytochrome P450 in vertebrates: rapid turnover of the detoxification-type and conservation of the biosynthesis-type

Ayaka Kawashima

DOCTOR OF PHILOTHOPHY

Department of Evolutionary Studies of Biosystems

School of Advanced Sciences

The Graduate University for Advanced Studies

2014

Acknowledgements

I am grateful for the help and support of many people. First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Yoko Satta, for giving me the opportunity to study fascinating relationships and interactions between organisms and chemicals. She gave me invaluable advice, academic guidance, and many opportunities not only to study our field but also to learn how to do science and how to think scientifically in general. She carefully guided me through my PhD life with help and support, and nurtured my desire to learn and study at all times.

I would also like to thank to Prof. Naoyuki Takahata for helpful suggestions and comments -- they were truly educational and allowed me to diversify my scientific thinking. I thank Associate Prof. Hideyuki Tanabe for teaching me 3D-FISH allowing me to study chromosomes. I appreciate the constructive questions from Associate Prof. Tatsuya Ota during my presentation, this has allowed me to rethink some of my work and has helped improve my thesis. I am indebted to Assistant Prof. Jun Gojobori for teaching me “Perl”, and developing my understanding of human evolution in general. I

thank Dr. Yoshiki Yasukochi for valuable advice on the study of CYP.

I would like to express my thanks to the members of the ESB department for their invaluable assistance and comments, which have made an enormous contribution to my work. Thank you all very much. I would also like to express my gratitude to my family for their continuing support, encouragement and understanding. My father picked me up from the station with his car and drove me to my house every time I came back late from work, my mother prepared lunch boxes for me every day, and my brother also always made sure I come home safely. Last, I thank my husband, Dr. Hirotaka Moriguchi, for cheering me up everyday. He supported me in all respects, especially psychologically.

Abstract

In this thesis, I aim to elucidate the birth and death processes of vertebrate *Cytochrome P450* (*CYP*) genes to understand the evolution of human CYPs. Members of the CYP family are important metabolic enzymes that are present in all metazoans. Genes encoding CYP form a multi-gene family, and the number of genes varies widely among species. The enzymes are classified as either biosynthesis- (B-type) or detoxification-type (D-type), depending on their substrates, but their origin and evolution have not been fully understood. In order to elucidate the birth and death process of CYP genes, I performed a phylogenetic analysis of 710 sequences from 14 vertebrate genomes and 543 sequences from 6 invertebrate genomes. The results showed that vertebrate D-type genes have independently emerged three times from B-type genes and that invertebrate D-type genes differ from vertebrates in their origins. B-type genes exhibit more conserved evolutionary processes than do D-type genes, with regard to the rate of gene duplication, pseudogenization, and amino acid substitutions. The differences in the evolutionary mode between B- and D-type genes may reflect differences in their respective substrates. The phylogenetic tree also revealed 11 clans comprising an upper category to families in the CYP nomenclature. Here, I report novel clan-specific amino acids that may be used for the qualitative definition of clans.

The novelties of this thesis are these three points and it is shown in Chapter 3:

1) The difference in the evolutionary mode between B- and D-types was shown quantitatively. Especially, I estimated the time and rate of gene duplications and pseudogenizations or losses by comparing genome sequences.
2) The origin of B-type genes has been believed to be ancient. On the other hand, the origin of D- type is only known to be a duplication of B-type. I showed vertebrate D-type genes have emerged independently from three different B-type genes, and invertebrate D-type genes appear to have an independent origin from vertebrae D-type genes.
3) The category of clan has been defined as an upper category of families in metazoan and plant *CYP* genes. I showed a clan-specific amino acids and this information is useful for qualitative classification of *CYP* clans.

In particular, I compare and contrast the origin and evolution of B- and D-types, and present an evolutionary model of vertebrate *CYP* genes. There is no report about evolutionary mode on *CYP* genes in vertebrates until now. In addition, CYP is thought that it evolved with adapting to the environment habitat. Besides, the presence of 58 CYP pseudogenes in humans and the presence of four human specific CYP pseudogenes have been reported, but there is no study about rest of 54 pseudogenes. Pseudogene is important tool for explore the evolutionary process of the gene. Hence, in this study I also made clear the cause of pseudogenization or time of pseudogenization of all human CYP pseudogenes.

In Chapter 4, I discussed the evolutionary mode of CYP genes in vertebrates, especially, focusing on the evolution of *CYP* genes that is driven by substrate specificity.

The metabolism of chemicals is the response system to the environment. Organisms had constructed many systems for chemical metabolism with adapting the intake of chemical materials. CYP is one of the important systems of these mechanisms. *CYP* genes are indispensable enzymes not only in the human but also in the almost all organisms. Modern humans use many medicines for the treatment of disease. Mice or macaques are used as model animals in the most study on CYP metabolism for medicines or medical sciences. It is necessary to understand the metabolic system in humans and to apply the result of the study based on model animals. But, it is difficult to confirm and reexamine the findings of model animals in humans directly. Therefore I tried to elucidate the characteristics of metabolic systems for chemicals in humans by using the evolutionary point of view. Relating to the results in vertebrate species that was used in this analysis, I showed the perspectives for evolution of chemical metabolic systems for other species in vertebrates. Cetacean and Carnivora are placental mammals, and they must have almost similar variety of CYP genes. But their foods or habitats are quite different from other mammals. Most Cetacean stop feeding plants and shifted their foods from plants to aquatic organisms. Canivora also changed their foods from plants to meets. Therefore their evolutionary processes on *CYP* genes are great interest by the viewpoint of the adapt to the environment.

I would like to show the relationships between CYP evolution in vertebrates and their foods or habitats. These evolutionary findings may become useful for the application of medical studies.

TABLE OF CONTENTS

Chapter 1. Introduction

1.1 Back ground

1.1.1 Organisms and chemicals

Chemicals had existed on the earth before the first organism had arisen. Ancestors have been treating a variety of chemical materials for a long time. In other words they have been evolving with adapting each environment for many years. Environment of organisms has changed frequently during the course of evolution. Some organisms have stayed in an ocean and the others move to a land. According to the great changes of environments such as from the ocean to the land, foods or habitats of organisms had changed dramatically.

Nowadays, there are so many different types of chemicals in this world. *Homo sapience* as well as other organisms became to be able to react to a variety of new chemical materials, which did not exist in the paleoenvironment. How we come to be able to respond for unseen chemicals? Animals take foods from outside of their bodies, including chemicals of alkaloids. It is not fully understood how and when organisms acquire systems for metabolizing unexpected chemical materials?

1.1.2 Detoxification systems in animals

It is thought that CYPs are more flexible than the other detoxification enzymes. The reason is that the CYP directly metabolize unexpected xenobiotics that came from outside of living organisms in the cell. On the other hand, other enzymes

metabolize the products oxidized by CYPs. Basic detoxification system in mammals is shown in Figure 1-1-1.

CYP genes have mainly two types of function in metabolizms: One is responsible for the detoxification of alkaloids or metabolisms of medicines and the other is for the biosysnthesis of physiologically active substances. CYP may have something specific evolutionary mode compare to other detoxification genes.

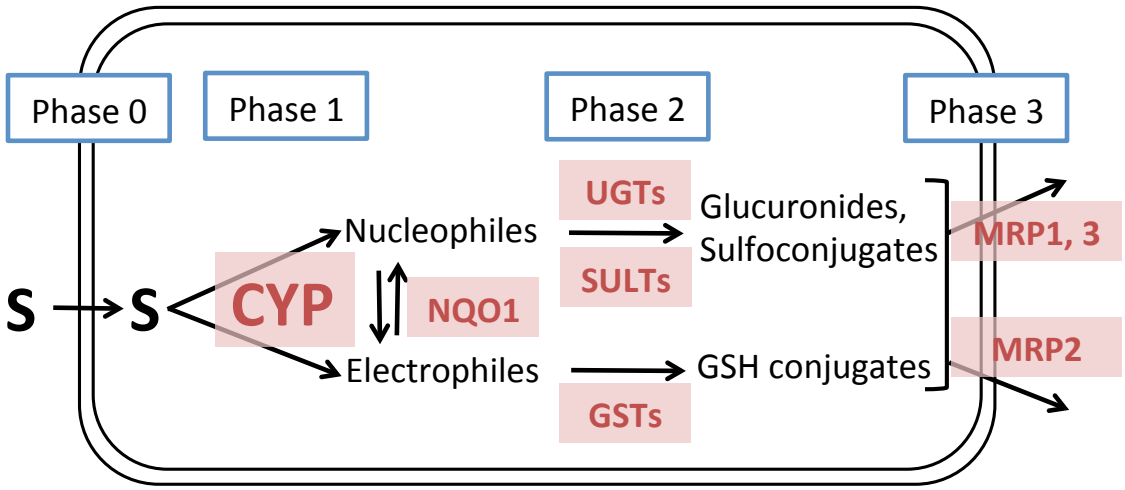


Figure 1-1-1. The detoxification systems in a mammalian cell.

There are 4 phases in the detoxification systems (Bock, 2003). S is a substrate. A round cornered square shows a cell membrane. Each red colored text are detoxification genes. *NQO1*: *NAD(P)H dehydrogenase, quinone 1*, *UGT*: *UDP glucuronosyltransferase*, *SULT*: *sulfotransferase*, *GST*: *glutathione S-transferase*, *MRP*: *Multidrug-Resistance like protein*.

1.2 Previous studies on CYPs

1.2.1 The classification of CYPs and chemicals

CYP is one of the heme-binding proteins. First it is found as “CO-binding pigment” that has specific absorbance at 450 nm from microsome in a rat liver (Klingenberg M, 1958), but their function or detail was not known. In 1962, Omura and Sato proofed that it is the heme-binding protein.

In 1971, Estabrook et al proposed “cyclic reaction mechanisms” (Estabrook RW *et al.*, 1971) for CYP function. The mechanisms are shown in Figure 1-2-1. A substrate binds to an oxygenated CYP and produce a complex. An electron is added to the complex, then heme-iron in the complex is reduced to the heme. And oxygen molecule added to the complex. Moreover, another electron is introduced to the complex, and the oxygen molecule is activated. Finally, the substrate will be oxygenated and released from the CYP.

In 1962, Hashimoto Y et al also found some proteins. They elucidated the material object represents Electron Paramagnetic Resonance (EPR) signals and estimated that the protein is including iron. After that, it has become clear that the material is same as CYP. Since then, CYPs have been found in yeasts, bacteria and plants in addition to animals. More than 10,000 *CYP* genes have been found until now and about 6,000 genes of them are named (Cytochrome P450 Homepage: http://drnelson.uthsc.edu/cytochromep450.html; Nelson DR, 2009). The nomenclature for classification is as shown in Figure 1-2-2. All CYPs are considered as being originated from a single ancestor, based on the conservation of amino acid sequences

(Gotoh O, 1993). Monooxygenases activity of CYP is characterized by the incorporation of one of two atoms of molecular oxygen into substrate, which results in hydroxylation in most cases.

CYP genes form a multi-gene family and encode proteins with amino-acid sequence identities higher than 40%. Each family comprises subfamilies with amino-acid sequence identities higher than 55%. In the classification of CYPs, a clan is defined as a higher-order category of *CYP* families (Nelson DR, 1998). Although clans can be useful for defining the relationships among *CYP* genes in different phyla within each kingdom (Gotoh O, 2012), the definition of "clan" is rather arbitrary compared with the definitions of "family" and "subfamily."

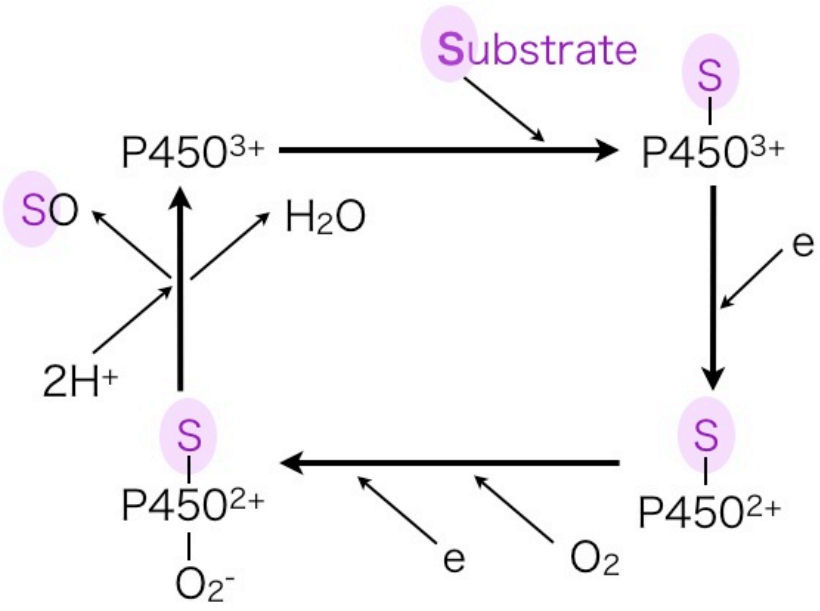


Figure 1-2-1. The cyclic reaction mechanisms of CYP.

The symbol on the right shoulder of CYP represents the electric charge of heme-iron. S, SO, e

shows substrate, a product of the reaction and an electron which is introduced to CYP, respectively.

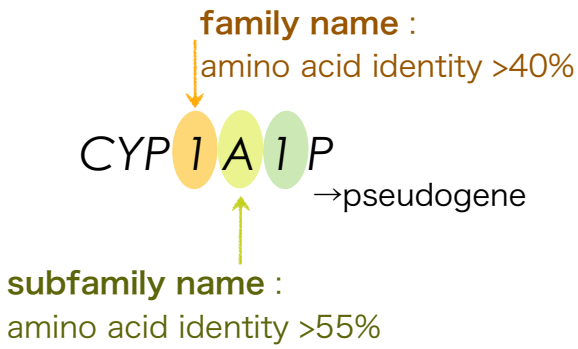


Figure 1-2-2. The name of CYP and their classifications

The number after "CYP" shows a family name. If the amino acid identity of two genes is over 40%, the genes are classified to a same family. The character after family name is a subfamily name. If the identity is over 55%, those genes are considered as of same subfamily. The last number of name is added in the order of discovery. In addition, if the gene had lost its function by flame shift or nonsense mutation, "P" is added at the end to indicate that it is pseudogene.

1.2.2 CYP in organisms

CYP genes are present in vertebrates, invertebrates, plants, fungi, and even some prokaryotes (Munro AW and Lindsay G, 1996). The number of known *CYP* genes in metazoan, plant, and fungus genomes is moderately large. For example, there are 115 *CYP* genes in the human genome, 97 in the sea squirt (*Ciona intestinalis*) (Cytochrome P450 Homepage), 120 in the sea urchin (*Strongylocentrotus purpuratus*) (Sea urchin genomesequencing consortium, 2006), 457 in the rice (*Oryza sativa*) (Nelson DR, 2009), 272 in *Arabidopsis thalian* (Mao G *et al*., 2013), and 159 in *Aspergillus oryzae* (Nelson

DR *et al*., 2013). In contrast, there are relatively a small number of *CYP* genes in eubacteria or Archaea, ranging from none in *Escherichia coli* to 33 in *Streptomyces avermitilis* (Nelson DR *et al*., 2013). Among metazoan *CYP* genes, *CYP51* is particularly conserved and participates in the synthesis of cholesterol, which is an essential component of eukaryotic cell membrane. A possible prokaryotic homolog (*CYP51B1*) to the metazoan *CYP51* is reported in the genome of *Mycobacterium tuberculosis* (Quaderer R *et al*., 2006). It is therefore thought that *CYP51* is the most ancient *CYP* gene. Although the functional role of CYPs in prokaryotes is not well defined (Aoyama Y *et al*., 1998, Yoshida Y *et al*., 2009, Debeljak N *et al.,* 2003), the presence of *CYP* genes in prokaryotes indicates that the emergence of *CYP*s preceded the origin of eukaryotes (Qi X *et al*., 2006). However, it has also been suggested that bacterial *CYP51* arose through lateral transfer from plants (Nelson DR, 1999). The absence of *CYP* genes in some bacteria, such as *E. coli,* suggests that they are not essential at least in some prokaryotes.

Plants have many *CYP* genes. *Oryza sativa* have 458 genes, for example, and the number is much greater than that in humans (57 genes). It is thought that the reason for the having various *CYP* genes in plants is to maintain lifecycle of plants by synthesizing many secondary metabolic products (Nelson D and Werck-Reichhart D, 2011). Many CYPs are responsible for this process (Ohmura T *et al*., 2009).

These days, it is reported that insects have 50-150 *CYP* genes; 85 in *Drosophila melanogaster*, 106 in *Anophelinae*, 46 in *Apis mellifera* and 87 in *Bombyx mori*. Insects are most prosperous organisms and they have many their specific

mechanisms for growing up; e.g. exuviation, metamorphosis or quiescence. CYPs assume metabolize the biologically active agent for these steps (Sutherland TD *et al*., 1998, Rewitz KF *et al*., 2007, Sandstrom P *et al*., 2006). In addition, insects eat various plants. Therefore insect specific *CYP* may play an important role in metabolizing and detoxifying a large variety of defensive chemicals of plants. Moreover, some are used for the resistance to the insecticides (Feyereisen R 1999, Scott JG 1999, Scott JG & Wen Z, 2001).

Yeast has the smallest number of *CYP* genes among the organisms. *Saccharomyces cerevisiae* have 3 *CYP* genes and *Schizosaccharomyces pombe* have only 2 *CYP*s. It is thought that *S. pombe* have the minimum number of *CYP* genes among eukaryotes. CYP51 was purified from S. cerevisiae by Yoshida Y *et al*., in 1977. This gene is important for biosynthesis of ergosterol which is one component of cell membrane and their function is almost same in animals (Mallory JC *et al*., 2005). Most Eukaryote has *CYP51,* which is the most conserved gene among *CYP*s.

1.2.3 CYP in animals and the function of CYPs in human

In animals, most of detoxification genes are expressed in liver for metabolizing toxins or drugs. On the other hand, *CYP* genes are expressed in almost all organs except erythrocyte and sperm. *CYP* genes are divided into two groups. One is detoxification type (D-type) and the other is biosynthesis type (B-type). The former is mainly located in liver microsomal fractions and the latter is in liver, adrenal and gonad microsomal fractions or mitochondrial fractions.

The function of CYP is listed in Table 1-2-1. The *CYP1*, *CYP2*, *CYP3* and *CYP4* family are responsible for detoxification of drugs or toxins. Especially, *CYP3A* is related to detoxification of more than 50% of drugs and medicines. CYP3A7 is specially expressed in fetus and the main enzyme for detoxification in their liver (50% in all expressed CYPs). However, after the birth the expressed rate of them decreases to 5% (Kato R *et al*., 2010). B-type genes are listed in Table 1-2-1 in blue color. The main function for B-type is biosynthesis of physiological bioactive substances (e.g. vitamin D, steroids etc.). Disease is sometimes happened, if these genes have some mutations (see chapter 4).

Table 1-2-1. The function and tissue distribution of *CYP* genes in human

Gene	Function	Tissue distribution
CYP1A1	polycyclic hydrocarbon metabolism	ubiquitous (after induction)
CYP1A2	aryl amine, drug metabolism	liver, GI tract brain
CYP1B1	polycyclic hydrocarbon metaabolism	adrenal ovary, testis, breast
CYP2A6	coumarin, nicotine metabolism	liver, nasal mucosa
CYP2A7	Unknown	liver breast
CYP2A13	drug metabolism	Respiratory tract
CYP2B6	drug metabolism	liver
CYP2C8	drug, steroid arachidonic acid metabolism	liver
CYP2C9	drug, steroid arachidonic acid metabolism	liver
CYP2C18	drug, steroid metabolism	liver
CYP2C19	drug metabolism	liver
CYP2D6	carcinogen, drug metabolism	liver
CYP2E1	carcinogen, drug metabolism	liver, WBCs, brain, kidney, placenta
CYP2F1	drug metabolism	liver

CYP2J2	drug steroid aracidonic. acid metabolism	brain, heart, pancreas, uterus, colon, kidney
CYP2R1	Unknown	pancreas, tonsil, kidney, lung, aorta, uterus, prostate
CYP2S1	Unknown	ubiquitous
CYP2U1	Unknown	brain, uterus, kidney, tonsil, lung
CYP2W1	Unknown	hepatoblastoma
CYP3A4	drug metabolism	liver
CYP3A5	drug metabolism	liver
CYP3A7	drug, steroid metabolism	low in adult liver, GI tract
CYP3A43	Unknown	testis
CYP4A11	fatty acid, arachidonic acid metablism	kidney
CYP4A20	unknown	breast uterus
CYP4A22	similar to CYP4A11	liver, kidney, fetal liver and spleen
CYP4B1	fatty acid, drug arachidonic acid metablism	lung, placenta, colon
CYP4F2	leukotriene, arachidonic acid, fatty acid, 12-HETE, drug metabolism	liver
CYP4F3	leukotriene, B4, fatty acid, HETE, drug metabolism	WBCs
CYP4F8	prostaglandin metabolism	seminal vesicles
CYP4F11	arachidonic Acid, fatty acid metabolism	breast, ovary, liver, kidney, lung, colon
CYP4F12	arachidonic Acid, fatty acid metabolism	colon
CYP4F22	arachidonic Acid, fatty acid metabolism	No ESTs
CYP4V2	unknown	lung, kidney, colon, bone, brain, uterus
CYP4X1	unknown	uterus, kidney, brain, aorta, lung, breast, prostate
CYP5A1	Thromboxane A2 synthase	Endothelial tissues
CYP7A1	bile acid biosynthesis	hepatocellular carcinoma, normal liver
CYP7B1	brain-specific	brain
CYP8A1	prostacyclin synthase	lung, aorta, endothelial cells
CYP8B1	sterol 12 α-HOase	kidney, fetal liver/spleen
CYP11A1	steroid biosynthesis	placenta, uterus, breast
CYP11B1	steroid biosynthesis	adrenal
CYP11B2	aldosterone synthase	breast, adrenal
CYP17A1	steroid 17 a-HOase, 17/20-lyase	adrenal

CYP19A1	aromatase, estrogen formation	ovary, placenta, testis, adipose, breast, brain
CYP20A1	unknown	brain, lung, testis, kidney, stomach, fetal liver, spleen heart
CYP21A2	steroid 21-HOase	adrenal
CYP24A1	vitamin D3 degradation, 24-Hoase	kidney
CYP26A1	retinoic acid HOase	breast, liver, keratinocytes
CYP26B1	retinoic acid HOase	eye, brain kidney
CYP26C1	retinoic acid HOase	rare transcript
CYP27A1	bile acid biosynthesis, vitamin D3 25- and 27-HOase	liver
CYP27B1	vitamin D3 1α-HOase	kidney
CYP27C1	unknown	testis
CYP39A1	24-HO-cholesterol 7α-HOase	stomach, testis, kidney, parathyroid, fetal liver/spleen
CYP46A1	Cholesterol 24-HOase	brain
CYP51A1	Lanosterol 14α-demethlase	testis, ovary, adrenal, liver,prostate, lung, ubiquitous

B-type and D-type gene is shown in blue and red, respectively.

(Each data of functions and tissue distribution were retrieved from NCBI databases.)

1.2.4 The structure of CYP protein and gene structure of *CYP* genes

The first crystal structure of CYP reported was derived from *Pseudomonas putida* (Poulos TL et al, 1985). CYP crystal structure revealed is triangle-prism shape. It looks like a rice ball, “Onigiri” in Japanese, and the place for a pickled plum, “Umeboshi”, is corresponding to an iron ion. The ion is located in the center as an active site of the enzyme reaction. Crystal structures of about 40 CYP had been reported until 2008 (Ohmura et al, 2009). CYP in prokaryote is hydrophilic while CYP in eukaryote is membrane-bound. It is difficult to crystalize eukaryotic CYP proteins and the number of crystalized eukaryotic CYPs are less than that in prokaryotes. *CYP* genes

have 7-9 exons and 6-8 introns. The number of exons and introns in the human *CYP* genes was shown in Table 1-2-2. In particular *CYP8B1* is intron-less.

Table 1-2-2 The number of exon and intron in human *CYP* genes

Gene	exon	intron						
CYP1A1	7	6	*CYP3A4*	13	12	*CYP8A1*	10	9
CYP1A2	7	6	*CYP3A5*	15	14	*CYP8B1*	1	0
CYP1B1	3	2	*CYP3A7*	13	12	*CYP11A1*	10	9
CYP2A6	9	8	*CYP3A43*	13	12	*CYP11B1*	9	8
CYP2A7	9	8	*CYP4A11*	12	11	*CYP11B2*	9	8
CYP2A13	9	8	*CYP4A22*	12	11	*CYP17A1*	7	6
CYP2B6	9	8	*CYP4B1*	12	11	*CYP19A1*	11	10
CYP2C8	10	9	*CYP4F2*	13	12	*CYP20A1*	13	12
CYP2C9	9	8	*CYP4F3*	14	13	*CYP21A2*	10	9
CYP2C18	9	8	*CYP4F8*	13	12	*CYP24A1*	12	11
CYP2C19	9	8	*CYP4F11*	12	11	*CYP26A1*	8	7
CYP2D6	9	8	*CYP4F12*	13	12	*CYP26B1*	6	5
CYP2E1	9	8	*CYP4F22*	14	13	*CYP26C1*	6	5
CYP2F1	10	9	*CYP4V2*	11	10	*CYP27A1*	9	8
CYP2J2	9	8	*CYP4X1*	12	11	*CYP27B1*	9	8
CYP2R1	5	4	*CYP4Z1*	12	11	*CYP27C1*	8	7

CYP2S1	9	8	*CYP5A1*	18	17	*CYP39A1*	12	11
CYP2U1	5	4	*CYP7A1*	6	5	*CYP46A1*	15	14
CYP2W1	9	8	*CYP7B1*	6	5	*CYP51A1*	11	10

B-type and D-type gene is represented as blue and red color.

1.3 The significance of this study

The metabolism of chemicals is the response system to the environment. Organisms had constructed many systems for chemical metabolism with adapting the intake of chemical materials. CYP is one of such important systems in these mechanisms. CYP genes are indispensable enzymes not only in the human but also in the almost all organisms. Modern humans use many medicines for the treatment of disease. Mice or macaques are used as model animals in the most study on CYP metabolism for medicine or medical science until now. It is necessary to understand the metabolic system in human to apply the result of the study obtained from model animals. But, it is difficult to confirm and reexamine the findings in humans directly. Therefore I tried to elucidate metabolic systems of chemicals in humans by using evolutionary point of view.

The presence of 58 CYP pseudogenes in human and the presence of human specific pseudogenes have been reported, Pseudogenes are important tool to explore the evolutionary process of a multi-gene family. Hence, in this study I made clear causes of

pseudogenization or time of pseudogenization of all human CYP pseudogenes.

In this paper, I aim to elucidate the birth and death processes of vertebrate *CYP* genes to understand human CYPs. In particular, I compare and contrast the origin and evolution of B- and D-types, and present an evolutionary model of vertebrate *CYP* genes. There is no report about evolutionary mode on CYP genes in vertebrates until now. Additionally CYP is thought to have been evolved with adapting to the environment habitat. Therefore I would like to show the relationships between CYP evolution in vertebrates and their foods or habitats. These evolutionary findings may become useful for the application of medical studies.

Chapter 2 Materials and Methods

2.1 Data collection

2.1.1 Sequence datasets and identification of B- and D-type genes in vertebrates

The nucleotide sequences of 115 *CYP* genes in the human genome were obtained from the Cytochrome P450 Homepage. Using these sequences as queries, I performed a basic local alignment search tool (BLAST) search by using BLASTn and downloaded coding sequences (CDS) of homologous nucleotide sequences from 14 vertebrate species (*Pan troglodytes*: CHIMP2.1.4, *Macaca mulatta*: MMUL_1, *Callithrix jacchus*: C_jacchus3.2.1, *Bos taurus*: UMD 3.1, *Canis lupus familiaris*: CanFam3.1, *Mus musculus*: GRCm38.p2, *Rattus norvegicus*: Rnor_5.0, *Monodelphis domestica*: BROADO5, *Gallus gallus*: Galgal4, *Taeniopygia guttata*: 3.2.4, *Anolis carolinensis*: AnoCar2.0, *Xenopus tropicalis*: JGI 4.2, *Oryzias latipes*: MEDAKA1.70, and *Danio rerio*: Zv9) from NCBI (http://www.ncbi.nlm.nih.gov/) or ENSEMBL databases (http://www.ensembl.org/index.html). In the BLAST search, the top two hits and the top five hits were retrieved when B- and D-type genes were used as queries, respectively. The nucleotide sequences of ref-seq from NCBI were obtained, and sequences from ENSEMBL were filtered by length (> 1000bp) and their identity with human genes. The extent of sequence identity was dependent on the divergence time between each vertebrate species and humans. For example, in fish, I filtered out sequences with identity > 60%. Orthology was confirmed by the presence of a syntenic region and the presence of adjacent loci, if any.

2.1.2 Sequence datasets in invertebrates

The following invertebrate species were included in the analysis: amphioxus (*Branchiostoma floridae*), sea squirt (*C. intestinalis*), sea urchin (*S. purpuratus*), sea anemone (*Nematostella vectensis*), water flea (*Daphnia pulex*), and fruit fly (*Drosophila melanogaster*). Protein sequences obtained from the Cytochrome P450 Homepage were used for the analysis of invertebrate CYPs. Only protein sequences > 350 amino acids in length were included in the phylogenetic analysis. Because of the too extensive sequence divergence between vertebrate and invertebrate *CYP* genes, BLAST searches of the NCBI and ENSEMBL databases were not performed. Therefore I just retrieved protein sequences of invertebrate CYPs from Homepages.

2.2 Methods

2.2.1 Molecular evolutionary analysis

Vertebrate nucleotide sequences and invertebrate amino acid sequences in *CYP* coding regions were aligned separately using ClustalW (Larkin MA *et al.*, 2007) implemented in MEGA5 (Tamura K *et al.*, 2011), and each alignment was further edited by hand. In the alignment of the vertebrate nucleotide sequences, I first translated them into the amino acid sequences and after checked by eye, reconverted them to the nucleotide sequences. I excluded sites at which > 20% of the operating taxonomic units (OTUs) showed gaps. As a result, 28.7% of the aligned sites showed > 60% identity,

48.5% showed > 50% identity, and 71.9% showed > 30% identity (data not shown). I then constructed Neighbor-joining (NJ) trees (Saitou N, Nei M, 1987) using either nucleotide differences per site (p-distance) (Nei M and Kumar S, 2000) or amino acid distances (JTT distance) (Jones D *et al.*, 1992). I performed missing data treatment under both the pairwise deletion and complete deletion options. The maximum likelihood (ML) (Felsenstein J, 1981) method was used to test the tree topology. All methods for tree construction were implemented in MEGA5 (Tamura K *et al.*, 2011).

2.2.2 Collection and classification of pseudogenes

The name and position in build 33 of 54 *CYP* pseudogenes (except human specific *CYP* pseudogenes: *CYP2G2P*, *2G3P*, *2T2P* and *2T3P*) was obtained from the Cytochrome P450 Homepage. Then, each pseudogene was searched by using their position and chromosome number from Archives of UCSC Apr 2003 (build33). Their DNA sequences were downloaded and checked whether the sequences are same as the latest version of the human genome by blastn. If their sequences are different from new one, I used current sequences for analyses. In this step, I also checked their location in the human genome and their functional paralogous genes. If the name of pseudogene is including "-se", which means that the gene contains a solo exon, I checked whether the sequence is corresponding to the exon number of pseudogene name or not.

2.2.3 Detection of pseudogenization or deletion of genes

The nucleotide sequences of the human *CYP* pseudogenes in the human genome were obtained from the Cytochrome P450 Homepage. I selected genes containing > 1000 bp out of the 1500 bp CDS. I retrieved orthologous genes from other vertebrate genomes by performing BLAST searches, using the human sequences as queries. The orthologous sequences were aligned with their human counterparts by ClustalW. Based on this alignment, I searched other vertebrates for nonsense or frame-shift mutations and examined if the positions are the same as in human. To estimate the time of pseudogenization, I calculated the ratio of non-synonymous substitutions per site to synonymous substitutions per site, for pairs of a pseudogene and an orthologous functional gene. Using this ratio, I estimated the pseudogenization time for all *CYP* pseudogenes based on the formula in Sawai *et al*. (2008). I used the TimeTree (http://www.timetree.org/index.php, Hedges SB *et al*., 2006) as a reference for species divergence time. When an orthologous gene was not detected in any non-human vertebrate, I searched for the syntenic region in the genome in order to confirm its deletion.

2.2.4 Estimation of functional constraint

In order to compare the functional constraint of each *CYP* gene in primates, I normalized the non-synonymous nucleotide substitution rate with the synonymous substitution rate. To be complete, I assumed that the gene tree is the same as the species tree for four primates (humans, chimpanzees, rhesus macaques, and marmosets) and

placed the numbers of synonymous and non-synonymous substitutions on each branch by the least squares method (Rzhetsky A and Nei M, 1992). The degree of functional constraint 1 - *f* is obtained from the ratio (*f*) of the sum of non-synonymous substitutions to that of synonymous substitutions of all branches in each tree. Finally, I compared the degree of functional constraint or directly the *f* value between B-type and D-type genes by using the Mann–Whitney *U* test (Mann HB and Whitney DR, 1947).

2.2.5 Detection of genome structure

Genome structure in the *CYP* cluster regions was compaired by Genome matcher 1.331 (Ohtsubo Y *et al*., 2008). Genome Matcher was used to obtained detailed information on nucleotide sequences similarity between duplicate units. A diagram drawn by this program depicts the extent of similarity between sequences using color codes, with red representing similarity greater than 95%, orange representing 90-95%, green representing 85-90%, and blue representing lower than 85%. Firstly, nucleotide positions of the *CYP* clusters on the human chromosome were obtained from Map Viewer in NCBI. Two anchor genes for each cluster were identified (*CYP2* cluster; *EGLN2* and *AXL*, *CYP4F* cluster; *RASAL3* and *OR10H4*, *CYP2C* cluster; *HELLS* and *PDLM*, *CYP4* cluster; *ATDAF1* and *TAL1*, *ZNF655* and *AZGP1*). Those anchor genes were selected since each pair of genes bound the respective CYP cluster in the most species analyzed. In other vertebrate, the sequences of syntenic region to human were obtained from a database of Synteny in ENSEMBL. Each sequence was compared with itself by Genome Matcher and the condition of blast search implemented in the

application was FF-W10-e0.01.

2.2.6 Identification of *Alu*

All *Alu*s in human *CYP* cluster were found out by Repeat Masker 3.2.9 (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org). Their positions and names were retrieved from the results. The number of *Alu*s in each cluster or gene was counted and recorded.

2.2.7 Causes of pseudogenization

To know causes of loss of function in *CYP* genes or seven human specific pseudogenes, I searched their functional paralogous genes in the human genome. I aligned those sequences with pseudogenes and looked for frame shift mutations or nonsense mutations. In addition, those mutations were sought in other vertebrates. If I detect deteriorated mutations, I examined the position and the codon with the mutation in all species. When a premature stop codon (TGA, TAG and TAA) was detected, I also checked whether their ancestral state was CGA or not. CGA codon tend to become TGA codon due to frequent methylation at CpG sites, and likely to produce nonsense mutation compared with other type of codons.

Chapter 3 Results

3.1 Origins of D-type *CYP* genes: Vertebrate D-type genes emerged independently three times from B-type genes

Among the 57 functional *CYP* genes in the human genome, 35 are D-type genes and 22 are B-type genes. This classification is based on the description of the enzyme substrate (Nelson DR, 1999), if any, and subfamily or family classification (Nelson DR, 2009). D-type genes constitute four *CYP* families: *CYP1* (3 genes), *CYP2* (16 genes), *CYP3* (4 genes), and *CYP4* (12 genes). B-type genes are grouped into 14 families: *CYP*5 (1 gene), *CYP7* (2 genes), *CYP8* (2 genes), *CYP11* (3 genes), *CYP17* (1 gene), *CYP19* (1 gene), *CYP20* (1 gene), *CYP21* (1 gene), *CYP24* (1 gene), *CYP26* (3 genes), *CYP27* (3 genes), *CYP39* (1 gene), *CYP46* (1 gene), and *CYP51* (1 gene) (Table 3-1-1). Using the definition proposed by Nelson (Nelson DR *et al*., 1996), the 57 CYPs can be classified into 10 clans: clans 2, 3, 4, mito, 7, 19, 20, 26, 46, and 51 (Table 3-1-2). Clan "mito" contains genes encoding enzymes that operate in mitochondria. Of the 10 clans, 6 (2, 3, 4, mito, 7, and 26) contain more than two families, whereas 4 (19, 20, 46, and 51) contain only one single family. The amino acid alignment of the 57 functional *CYP* genes showed that four amino aid sites are conserved. Two of these (310F and 316C) are located within the heme-binding region (Figure 3-1-1). The latter site (316C) is known to be structurally close to the iron ion in the heme-binding region and to operate as an active center of the enzyme (Meunier B *et al*., 2004). This

conserved cysteine is said as the proximal Cys (Meunier B *et al*., 2004). The other two sites (242E and 245R) are located about 80 amino acids upstream from the proximal Cys. Although it is unknown whether these amino acids are involved in any specific function, their conservation suggests some evolutionary or functional importance. Furthermore, several clan-specific amino acids were found in the 57 functional human *CYP*s (Figure 3-1-2). Some of them were conserved not only in vertebrates but also in metazoans, although the number of conserved sites correlates with the number of genes in each clan.

To characterize the phylogenetic relationships among the 57 functional human *CYP* genes, an NJ tree was constructed based on the total nucleotide differences (*p*-distances) between the CDSs (Figure 3-1-3). In the resulting tree, members of each family formed monophyletic groups with respect to other families, and each monophyletic group was supported by a relatively high bootstrap value. The phylogeny showed that D-type genes emerged independently from B-type genes at least three times: first, an ancestral gene of *CYP17A1* and *CYP21A1* was duplicated, generating the ancestor of the *CYP1* and *CYP2* families (node *a* in the tree, Figure 3-1-3). Second, the *CYP3A* subfamily arose from the common ancestor of *CYP3* and *CYP5* (node *b* in the tree). Third, an ancestor of *CYP46A1* was duplicated, generating the ancestor of the *CYP4* family (node *c* in the tree). All nodes (*a*, *b*, and *c*) were supported by high bootstrap values (94~100% in Figure 3-1-3). In addition to these bootstrap values, amino acids that could distinguish B- from D-type genes were also identified (Figure 3-1-4). For example, an amino acid site in the middle of the sequence supported node *a*.

In the D-type genes, F was shared by all members of the *CYP1* family whereas V was shared by most members of the *CYP2* family (except T in CYP2U1) at 274. In contrast, the B-type *CYP*s, *CYP17A1*, and *CYP21A1*, shared T at that site. Similarly, several other amino acid changes that support nodes *a*, *b*, and *c* were observed (Figure 3-1-4; red column in Clan2, green in Clan3 and blue in Clan4 and 46, respectively).

To investigate the duplicaion times of three major D-types from their ancestral B-types, orthologs and paralogs of human B-type and D-type *CYP* genes were retrieved from 14 vertebrate genomes. This resulted in a total of 710 *CYP* nucleotide sequences so that I examined twice as many vertebrate sequences as in the previous study (388) (Nelson DR *et al*., 2013). The presence or absence of vertebrate orthologs to the 57 functional human genes is summarized in Table 3-1-3, showing that almost all 14 genomes contain orthologs of B-type genes. I used the pairwise deletion option and constructed a phylogenetic tree by using nucleotide sequences (Figure 3-1-5); its topology readily confirmed the orthologous relationship between human and other vertebrate B-type genes. However, it was difficult to identify orthologous relationships between D-type genes from humans and other vertebrates, especially in the *2A*, *2C*, *3A*, and *4F* subfamilies, owing to frequent species-specific duplications. Nevertheless, monophyletic relationships within each D-type family (*CYP1–4*) were observed with relatively high bootstrap values (> 80%), so that vertebrate genes in each monophyletic group are classified as the D-type. The phylogenetic analysis revealed that human D- and B-type genes had already emerged when vertebrates diverged, and that three duplication events occurred in the B-type genes from which the D-type genes were

originated. Assuming a molecular clock and that zebrafish and humans diverged 400 million years ago (mya) (TimeTree; http://www.timetree.org/), I calculated the total branch lengths leading to both B- and D-type genes (branch b_A, b_B, and b_C to B-type and b'_A, b'_B, and b'_C to D-type in Figure 3-1-6A, B, C) to estimate the timing of the emergence of the *CYP1–4* families (nodes *a*, *b*, and *c* in Figure 3-1-3). Since b_B, b_C, b'_B, and b'_C correspond to 400 million years (myr), each ratio of (b_A+ b_B) to b_B, (b_A+ b_C) to b_C, (b'_A+ b'_B) to b'_B, and (b'_A+ b'_C) to b'_C yielded an estimate of the duplication time. The estimates varied from 623–1316 mya for *a*, 601–664 mya for *b*, and 681–926 mya for *c*. To be conservative, I used the youngest estimate for each node: 623 ± 35 mya for *a*, 601 ± 34 mya for *b*, and 681 ± 37 mya for *c*. As anticipated, these estimates preceded the emergence of vertebrates (608 mya, TimeTree) but occurred after the divergence of vertebrates and chordates (774 mya, TimeTree). This finding suggests that invertebrates do not possess orthologs to vertebrate D-type genes, despite the presence of D-type *CYP*s in insects, which function in insecticide resistance and detoxification of plant alkaloids (Urabe K *et al*., 1990).

Table 3-1-1. The number of *CYP* gene in Human

	Functional gene	Pseudogene
Detoxification	35[a]	14[b]
Biosynthesis	22[c]	3[d]

After exclusion of truncated pseudogeens each category includes genes as below

a: *CYP1A1*, *CYP1A2*, *CYP1B1*, *CYP2A6*, *CYP2A7*, *CYP2A13*, *CYP2B6*, *CYP2C8*, *CYP2C9*, *CYP2C18*, *CYP2C19*, *CYP2D6*, *CYP2E1*, *CYP2F1*, *CYP2J2*, *CYP2R1*, *CYP2S1*, *CYP2U1*, *CYP2W1*, *CYP3A4*, *CYP3A5*, *CYP3A7*, *CYP3A43*, *CYP4A11*, *CYP4A20*, *CYP4A22*, *CYP4B1*, *CYP4F2*, *CYP4F3*, *CYP4F8*, *CYP4F11*, *CYP4F12*, *CYP4F22*, *CYP4V2*, *CYP4X1*,

b: *CYP1D1P*, *CYP2A7P1*, *CYP2B7P1*, *CYP2D7P1*, *CYP2D8P1*, *CYP2F1P*, *CYP2G1P*, *CYP2G2P*, *CYP2T2P*, *CYP2T3P*, *CYP4F9P*, *CYP4F23P*, *CYP4F24P*, *CYP4Z2P*,

c: *CYP5A1*, *CYP7A1*, *CYP7B1*, *CYP8A1*, *CYP8B1*, *CYP11A1*, *CYP11B1*, *CYP11B2*, *CYP17A1*, *CYP19A1*, *CYP20A1*, *CYP21A2*, *CYP24A1*, *CYP26A1*, *CYP26B1*, *CYP26C1*, *CYP27A1*, *CYP27B1*, *CYP27C1*, *CYP39A1*, *CYP46A1*, *CYP51A1*,

d: *CYP21A1P*, *CYP51P1*, *CYP51P2*

Table 3-1-2. The classification of *CYP* genes (using a partial list of Clans)

Class	Group	Vertebrates	Arthropoda	Nematode	Fungi	Plant	Protist	Prokaryote
B					55			101
E	I	2	2	Cel	64, 501		508	
	II	3, 4, 46	3, 4	3, 4	52, 56, 505, 534, 72, 86, 97, 711		525	102, 110, 132
	III	mito	mito	mito				
	IV	7, 16, 26, 51			51, 524, 550	51, 85	51	51, 120
	V	19						
	VII	20				727	526	117
	X	74				74		

Class B: specific to prokaryotes (except for *CYP55* genes), E: eubacteria, archaea, viruses and eukaryotes

Figure 3-1-1. Conserved amino acids at a position.

The *x*-axis indicates amino-acid positions in an alignment of 710 vertebrate *CYP* genes (after excluding gaps), and the y-axis indicates the proportion of most conserved amino acids at each position. Red bars indicate highly (> 95%) conserved positions. The chart below the bar graph shows the approximate position of six substrate recognition sites (SRS): SRS1–6. The bracket represents the heme-binding region (~10 amino acids) of a *CYP* gene.

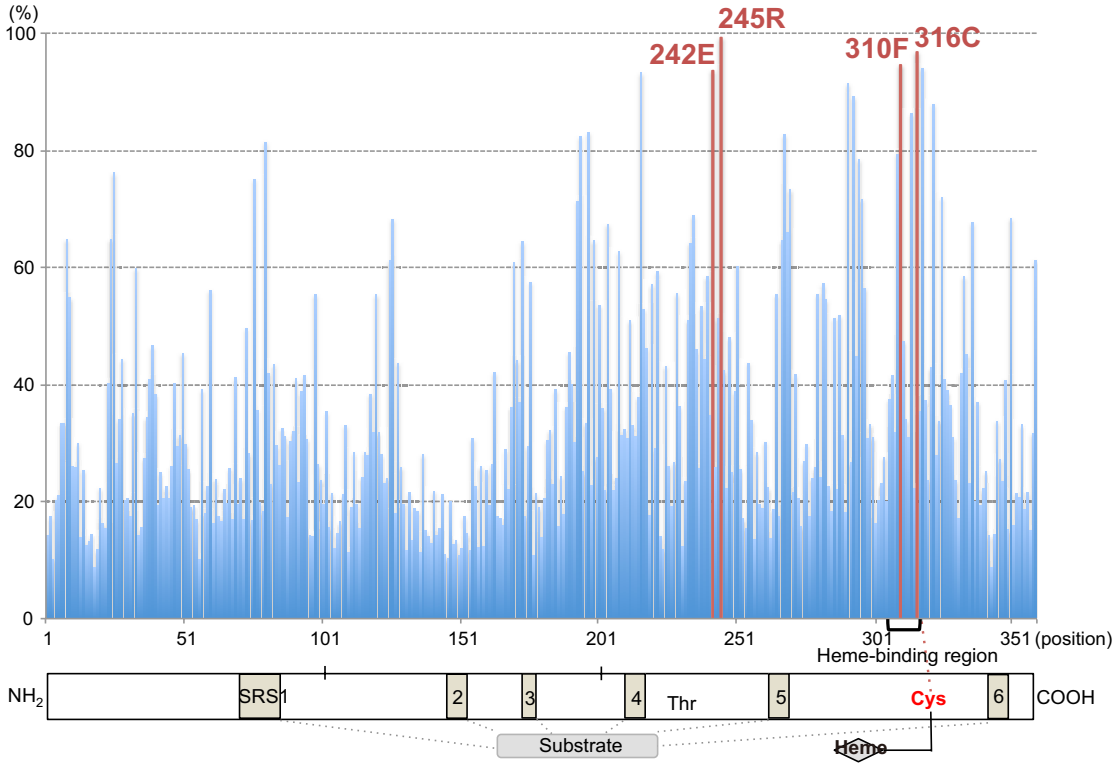


Figure 3-1-3. Conserved amino acids within each clan of vertebrate *CYP* genes.

Conserved amino acids within each clan are shown in colored columns: the red column is specific for clan 2; aqua, mito; orange, 4; blue, 3; yellow, 26, and brown, 7. The four sites in purple (E, R, G, and C) are highly conserved (> 95%) sites among all vertebrate species.

Clan	gene name	(positions) 275	276	277	278	279	280	281	282	283	284	285	350	351	352	353	354	355	356	357	358	359	360
	Hosa CYP1A1	I	L	E	T	F	R	H	S	S	F	V	F	G	M	G	K	R	K	C	I	G	E
	Hosa CYP1A2	.	.	.	.	.	.	.	.	.	.	L	.	.	.	.	.	.	R	.	.	.	.
	Hosa CYP1B1	L	Y	.	A	M	.	F	.	.	.	.	.	S	V	.	.	.	R	.	.	.	.
	Hosa CYP2A6	.	H	.	I	Q	.	F	G	D	V	I	.	S	I	.	.	.	N	.	F	.	.
	Hosa CYP2A7	.	H	.	I	Q	.	F	G	D	V	I	.	S	I	.	.	.	N	.	F	.	.
	Hosa CYP2A13	.	H	.	I	Q	.	F	G	D	M	L	.	S	I	.	.	.	Y	.	F	.	.
	Hosa CYP2B6	.	Y	.	I	Q	.	F	.	D	L	L	.	S	L	.	.	.	I	.	L	.	.
	Hosa CYP2C8	V	H	.	I	Q	.	Y	.	D	L	.	.	S	A	.	.	.	I	.	A	.	.
	Hosa CYP2C9	V	H	.	V	Q	.	Y	I	D	L	L	.	S	A	.	.	.	I	.	V	.	.
	Hosa CYP2C18	V	H	.	I	Q	.	Y	I	D	L	L	.	S	A	.	.	.	M	.	M	.	.
2	*Hosa CYP2C19*	V	H	.	V	Q	.	Y	I	D	L	I	.	S	A	.	.	.	I	.	V	.	.
	Hosa CYP2D6	.	H	.	V	Q	.	F	G	D	I	.	.	S	A	.	R	.	A	.	L	.	.
	Hosa CYP2E1	V	H	.	I	Q	.	F	I	T	L	.	.	S	T	.	.	.	V	.	A	.	.
	Hosa CYP2F1	.	H	.	V	Q	.	F	A	D	I	I	.	S	A	.	R	.	L	.	L	.	.
	Hosa CYP2J2	.	H	.	V	Q	.	M	G	N	I	I	.	S	I	.	.	.	A	.	L	.	.
	Hosa CYP2R1	L	H	.	V	L	.	F	C	N	I	.	.	S	L	.	R	.	H	.	L	.	.
	Hosa CYP2S1	L	H	.	A	Q	.	L	L	A	L	.	.	S	L	.	.	.	V	.	L	.	.
	Hosa CYP2U1	.	M	.	V	Q	.	L	T	V	V	.	.	.	I	.	.	.	V	.	M	.	.
	Hosa CYP2W1	L	H	.	V	Q	.	F	I	T	L	L	.	S	A	.	R	.	V	.	V	.	.
	Hosa CYP17A1	.	R	.	V	L	.	L	R	P	V	A	.	.	A	.	P	.	S	.	.	.	.
	Hosa CYP21A2	.	A	.	V	L	.	L	R	P	V	.	.	.	C	.	A	.	V	.	L	.	.
19	*Hosa CYP19A1*	.	Y	.	S	M	.	Y	Q	P	V	.	.	.	F	.	P	.	G	.	A	.	K
	Hosa CYP11A1	.	K	.	.	L	.	L	H	P	I	S	.	.	W	.	V	.	Q	.	L	.	R
	Hosa CYP11B1	L	K	.	.	L	.	L	Y	P	V	G	-	-	-	-	-	-	-	-	-	-	-
	Hosa CYP11B2	L	K	.	.	L	.	L	Y	P	V	G	.	.	F	.	M	.	Q	.	L	.	R
mito	*Hosa CYP24A1*	L	K	.	S	M	.	L	T	P	S	.	.	.	V	.	.	.	M	.	.	.	R
	Hosa CYP27A1	L	K	.	.	L	.	L	Y	P	V	.	.	.	Y	.	V	.	A	.	L	.	R
	Hosa CYP27B1	V	K	.	V	L	.	L	Y	P	V	.	.	.	F	.	.	.	S	.	M	.	R
	Hosa CYP27C1	L	K	.	.	L	.	L	F	P	V	L	.	.	H	.	V	.	S	.	.	.	R
	Hosa CYP4A11	.	K	.	A	L	.	L	Y	P	P	.	.	S	G	.	S	.	N	.	.	.	K
	Hosa CYP4A22	.	K	.	A	L	.	L	Y	P	P	.	.	S	G	.	S	.	N	.	.	.	K
	Hosa CYP4B1	.	K	.	S	.	.	L	Y	P	P	.	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4F2	M	K	.	S	L	.	L	H	P	P	.	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4F3	.	K	.	S	L	.	L	H	P	P	.	.	S	A	.	P	.	N	.	.	.	Q
4	*Hosa CYP4F8*	L	K	.	S	L	.	L	H	P	P	I	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4F11	.	K	.	S	L	.	L	H	P	P	.	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4F12	V	K	.	S	L	.	L	H	P	P	A	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4F22	.	K	.	S	L	.	Q	Y	P	P	.	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4V2	.	K	.	.	L	.	L	F	P	S	.	.	S	A	.	P	.	N	.	.	.	Q
	Hosa CYP4X1	.	K	.	.	C	.	L	I	P	A	.	.	S	A	.	S	.	N	.	.	.	Q
	Hosa CYP4Z1	.	K	.	C	L	.	L	Y	A	P	.	.	S	A	.	L	.	N	.	.	.	Q
46	*Hosa CYP46A1*	L	K	.	S	L	.	L	Y	P	P	A	.	S	L	.	H	.	S	.	.	.	Q
	Hosa CYP3A4	V	N	.	.	L	.	L	F	P	I	A	.	.	S	.	P	.	N	.	.	.	M
	Hosa CYP3A5	V	N	.	.	L	.	L	F	P	V	A	.	.	T	.	P	.	N	.	.	.	M
3	*Hosa CYP3A7*	V	N	.	.	L	.	L	F	P	V	A	.	.	S	.	P	.	N	.	.	.	M
	Hosa CYP3A43	V	N	.	.	L	.	L	F	P	V	.	.	.	A	.	P	.	N	.	.	.	M
	Hosa CYP5A1	.	A	.	.	L	.	M	Y	P	P	A	.	.	A	.	P	.	S	.	L	.	V
20	*Hosa CYP20A1*	L	C	.	.	V	.	T	A	K	L	T	.	S	-	.	T	Q	E	.	P	E	L
	Hosa CYP26A1	.	K	.	.	L	.	L	N	P	P	.	.	.	G	.	L	.	S	.	V	.	K
26	*Hosa CYP26B1*	.	K	.	V	M	.	L	F	T	P	I	.	.	G	.	V	.	T	.	L	.	K
	Hosa CYP26C1	V	K	.	V	L	.	L	L	P	P	.	.	.	G	.	A	.	S	.	L	.	Q
	Hosa CYP7A1	.	K	.	S	L	.	L	.	.	A	S	.	.	S	.	A	T	I	.	P	.	R
	Hosa CYP7B1	.	F	.	A	L	.	L	.	.	Y	S	.	.	T	.	T	S	.	.	P	.	R
7	*Hosa CYP8A1*	L	S	.	S	L	.	L	T	A	A	P	W	.	A	.	H	N	H	.	L	.	R
	Hosa CYP8B1	V	E	.	.	L	.	L	R	A	A	P	W	.	S	.	V	S	I	.	P	.	R
	Hosa CYP39A1	V	.	.	.	I	.	L	K	-	-	A	.	.	S	.	.	F	Q	.	P	A	R
51	*Hosa CYP51A1*	.	K	.	.	L	.	L	R	P	P	I	.	.	A	.	R	H	R	.	.	.	.

Figure 3-1-3. Phylogenetic tree of *cytochrome P450* genes in humans.

The tree includes all functional *CYP* genes in humans (*Hosa*) and all yeasts (*Sace, Saccharomyces cerevisiae*; *Scpo*, *Schizosaccharomyces pombe*). The tree was constructed using the NJ method for nucleotide differences between the CDS and rooted with yeast *CYP51* gene sequences (*Sace ERG11* and *Scpo erg11*). Red text indicates D-type *CYP* genes, and black and blue text indicate B-type *CYP* genes. The *CYP1–4* families are indicated by a red bracket on the right side of the tree. Three diamonds (*a*, *b*, and *c*) indicate gene duplications that arose both B- and D-type genes. The B-type genes that were the ancestors of D-type genes are indicated with a blue line and character. Black brackets and roman numerals (i–v) at the tips of the tree show five clusters (see Chapter3-5) of D-type genes: i, the *CYP2* family on chromosome 19q; ii, the *CYP2C* subfamily on chromosome 10q; iii, the *CYP3A* subfamily on chromosome 7q; iv, the *CYP4* family on chromosome 1p; v, the *CYP4F* subfamily on chromosome 19p. The number near each node indicates the bootstrap value (> 94%) supporting the node.

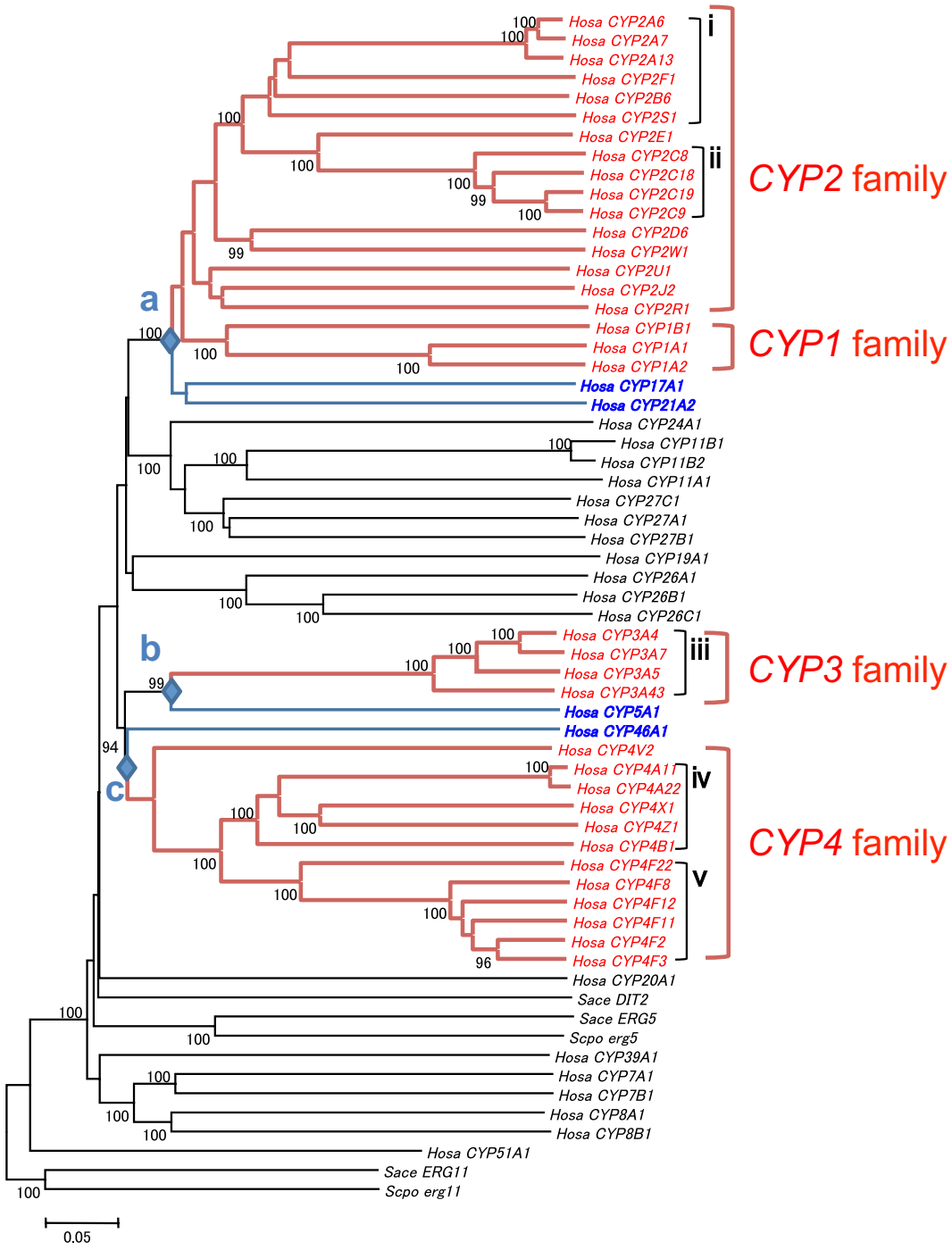
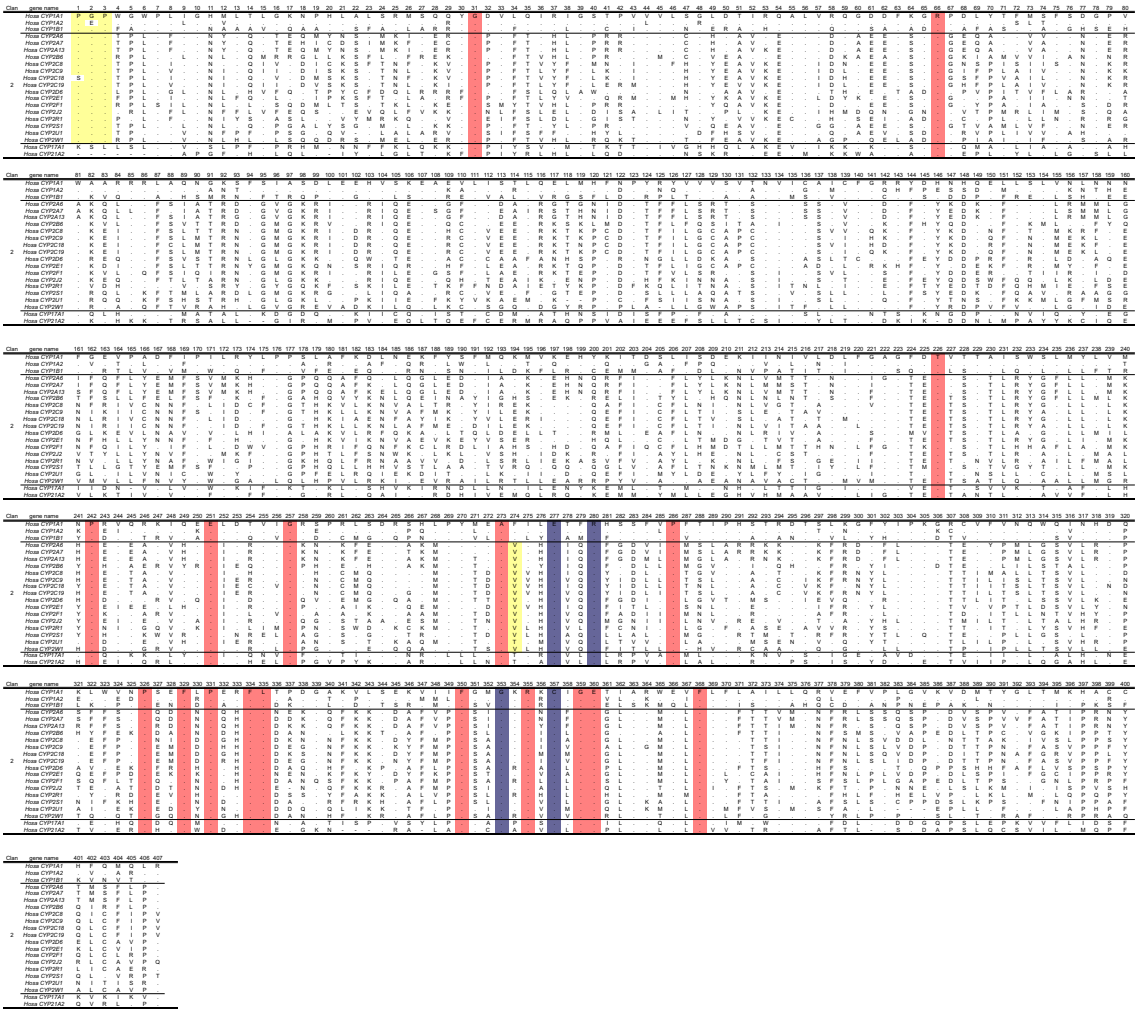

Hosa CYP2A6
Hosa CYP2A7
Hosa CYP2A13
Hosa CYP2F1
Hosa CYP2B6
Hosa CYP2S1
Hosa CYP2E1
Hosa CYP2C8
Hosa CYP2C18
Hosa CYP2C19
Hosa CYP2C9
Hosa CYP2D6
Hosa CYP2W1
Hosa CYP2U1
Hosa CYP2J2
Hosa CYP2R1
Hosa CYP1B1
Hosa CYP1A1
Hosa CYP1A2
Hosa CYP17A1
Hosa CYP21A2
Hosa CYP24A1
Hosa CYP11B1
Hosa CYP11B2
Hosa CYP11A1
Hosa CYP27C1
Hosa CYP27A1
Hosa CYP27B1
Hosa CYP19A1
Hosa CYP26A1
Hosa CYP26B1
Hosa CYP26C1
Hosa CYP3A4
Hosa CYP3A7
Hosa CYP3A5
Hosa CYP3A43
Hosa CYP5A1
Hosa CYP46A1
Hosa CYP4V2
Hosa CYP4A11
Hosa CYP4A22
Hosa CYP4X1
Hosa CYP4Z1
Hosa CYP4B1
Hosa CYP4F22
Hosa CYP4F8
Hosa CYP4F12
Hosa CYP4F11
Hosa CYP4F2
Hosa CYP4F3
Hosa CYP20A1
Sace DIT2
Sace ERG5
Scpo erg5
Hosa CYP39A1
Hosa CYP7A1
Hosa CYP7B1
Hosa CYP8A1
Hosa CYP8B1
Hosa CYP51A1
Sace ERG11
Scpo erg11
a
b
c
i
ii
iii
iv
v
CYP2 family
CYP1 family
CYP3 family
CYP4 family
0.05

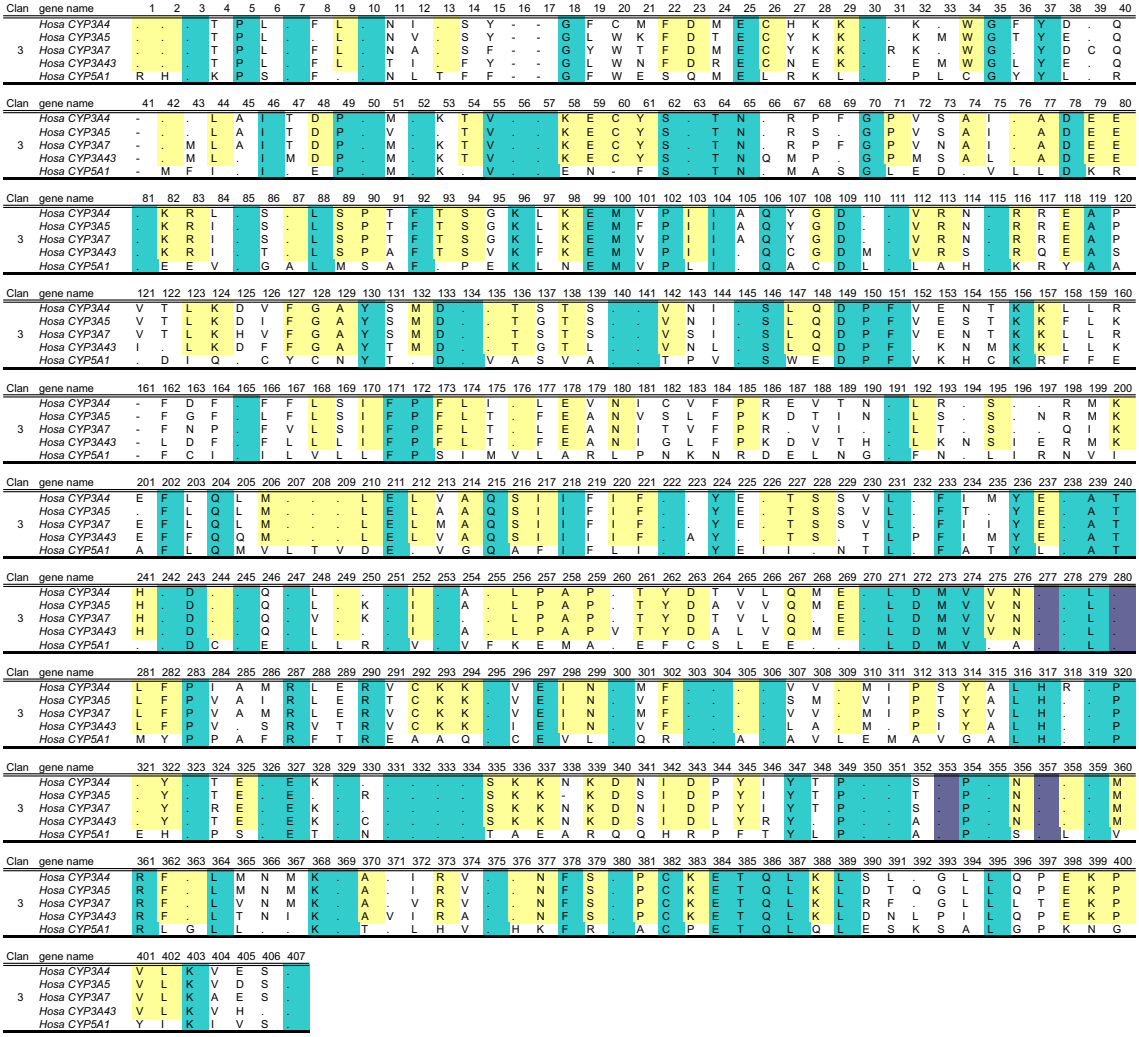
Figure 3-1-4. Amino acids distinguishing B-type from D-type genes within each clan.

The yellow column indicates the amino acids in B-type genes that differ from D-type genes. The other colors are the same as in Figure 3-1-2. A) Clan 2; B) Clan 3; C) Clan 4 and 46.

A) Clan2



B) Clan3



C) Clan4 and 46

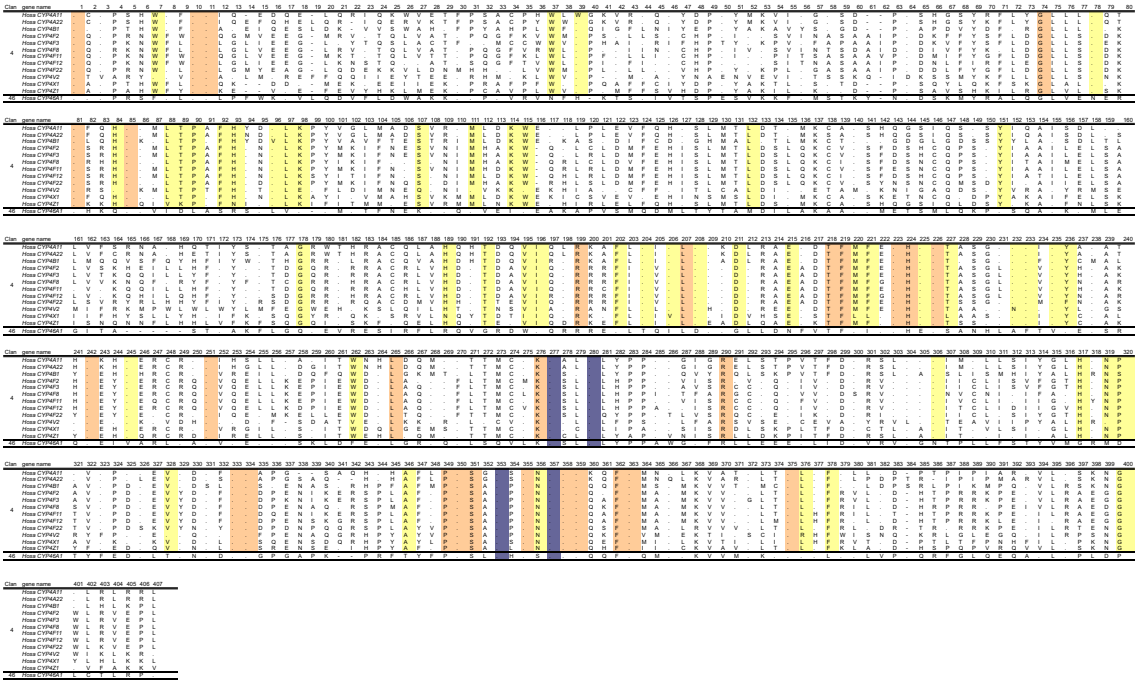


Table 3-1-3. Presence or absence of vertebrate orthologs to human *CYP* genes

CYP		*Species name*													
family	*genes*	*Patr*	*Mamu*	*Caja*	*Bota*	*Cafa*	*Mumu*	*Rano*	*Modo*	*Anca*	*Gaga*	*Tagu*	*Xetr*	*Orla*	*Dare*
1	*A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*							
	A2	*1*	*1*	*1*	*1*	*1*	*1*	*1*							
	A1 or A2								*2*	*3*	*1*	*1*	*1*	*1*	*1*
	B1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*
	Others	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*1*	*2*	*1*	*0*	*3*	*2*	*3*
2	*A6*	*0*	*1*	*0*	*0*	*0*									
	A7	*1*	*2*	*0*	*0*	*1*									
	A13	*1*	*2*	*1*	*1*	*1*									
	Other A*						*4*	*2*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
	B6	*1*	*1*	*1*	*1*	*1*	*4*	*4*	*2*	*0*	*0*	*0*	*0*	*0*	*0*
	C8	*1*	*1*	*1*											
	C9	*1*	*1*	*1*											
	C18	*1*	*1*	*1*											
	C19	*1*	*1*	*1*											
	Other C*				*7*	*2*	*9*	*6*	*8*	*0*	*0*	*0*	*0*	*0*	*0*
	D6	*1*	*1*	*1*	*2*	*0*	*5*	*5*	*1*	*1*	*1*	*1*	*5*	*0*	*0*
	E1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
	F1	*1*	*1*	*1*	*1*	*3*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*
	J2	*1*	*1*	*1*	*5*	*1*	*6*	*3*	*6*	*1*	*4*	*2*	*1*	*0*	*0*
	R1	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
	S1	*1*	*1*	*1*	*1*	*1*	*2*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
	U1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*1*
	W1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*2*	*1*	*1*	*0*	*0*	*0*
	Others	*1*	*1*	*1*	*1*	*2*	*1*	*2*	*1*	*15*	*7*	*7*	*18*	*10*	*20*
3	*A4*	*1*	*1*	*1*											
	A5	*1*	*1*	*1*											
	A7	*1*	*1*	*1*											
	A43	*1*	*1*	*1*											
	*Other A**				*3*	*4*	*8*	*4*	*4*	*3*	*2*	*1*	*5*	*1*	*1*
	Others	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*2*	*5*
4	*A11*	*1*	*1*	*1*	*2*	*1*									
	A22	*1*	*1*	*1*	*2*	*1*									
	A11 or						*5*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*

	22														
	*OtherA**	*0*	*0*	*0*	*0*	*3*									
	B1	*1*	*1*	*1*	*1*	*1*	*2*	*1*	*1*	*6*	*2*	*2*	*4*	*2*	*3*
	F2	*1*	*1*	*1*											
	F3	*1*	*1*	*1*											
	F8	*1*	*1*	*1*											
	F11	*1*	*1*	*1*											
	F12	*1*	*1*	*1*											
	F22	*1*	*1*	*1*											
	*Other F**				*6*	*3*	*9*	*8*	*6*	*1*	*0*	*1*	*3*	*1*	*1*
	V2	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*0*	*2*	*1*	*2*
	X1	*1*	*1*	*1*	*1*	*2*	*1*	*1*	*2*	*0*	*0*	*0*	*0*	*0*	*0*
	Z1	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
5	*5A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
7	*A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
	B1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*
8	*A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*1*	*1*	*1*
	B1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*2*	*1*	*3*
11	*A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*
	B1	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*1**
	B2	*1*	*1*	*1*	*0*	*1*	*1**	*1**	*0*	*0*	*0*	*0*	*0*	*0*	*0*
17	*17A1*	*1*	*1*	*1*	*2*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*2*
19	*19A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*2*	*2*
20	*20A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
21	*21A2*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*2*	*2*
24	*24A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*2*	*2*	*1*	*1*
26	*A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
	B1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*1*	*1*
	C1	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
27	*A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*2*	*1*	*2*
	B1	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*
	C1	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*
39	*39A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*
46	*46A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*1*	*1*	*1*	*2*	*2*	*2*
51	*51A1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*

Others: *CYP* genes are not orthologs to the human *CYP* genes listed here but are subfamily members belonging to each family.

*: Genes are included in the subfamily, but the subfamily number differs from that in humans

The abbreviations are as below. *Patr*: *Pan troglodytes*, *Mamu*: *Macaca mulatta*, *Caja*: *Callithrix jacchus*, *Bota*: *Bos taurus*, *Cafa*: *Canis lupus familiaris*, *Mumu*: *Mus musculus*, *Rano*: *Rattus norvegicus*, *Modo*: *Monodelphis domestica*, *Anca*: *Anolis carolinensis*, *Gaga*: *Gallus gallus*, *Tagu*: *Taeniopygia guttata*, *Xetr*: *Xenopus tropicalis*, *Orla*: *Oryzias latipes*, *Dare*:*Danio rerio*.

Figure 3-1-5. The phylogenetic tree of B- and D-type *CYP* genes in vertebrates.

An internal bracket at the tips of the tree indicates the *CYP* family in vertebrates, and an external bracket indicates clusters for a clan. D-type *CYP* genes in humans are shown in red, and B-type *CYP* genes are shown in blue. Red-shaded branches indicate the divergence of D-type from B-type. Text colors indicate the following: red for D- and blue for B-type in humans; dark brown for *Bos taurus*; light blue for *Canis lupus familiaris*; pink for *Mus musculus*; aqua for *Rattus norvegicus*; dark red for *Monodelphis domestica*; dark orange for *Gallus gallus*; purple for *Taeniopygia guttata*; brown for *Anolis carolinensis*; blue-purple for *Xenopus tropicalis*; red-purple for *Oryzias latipes*; orange for *Danio rerio*.

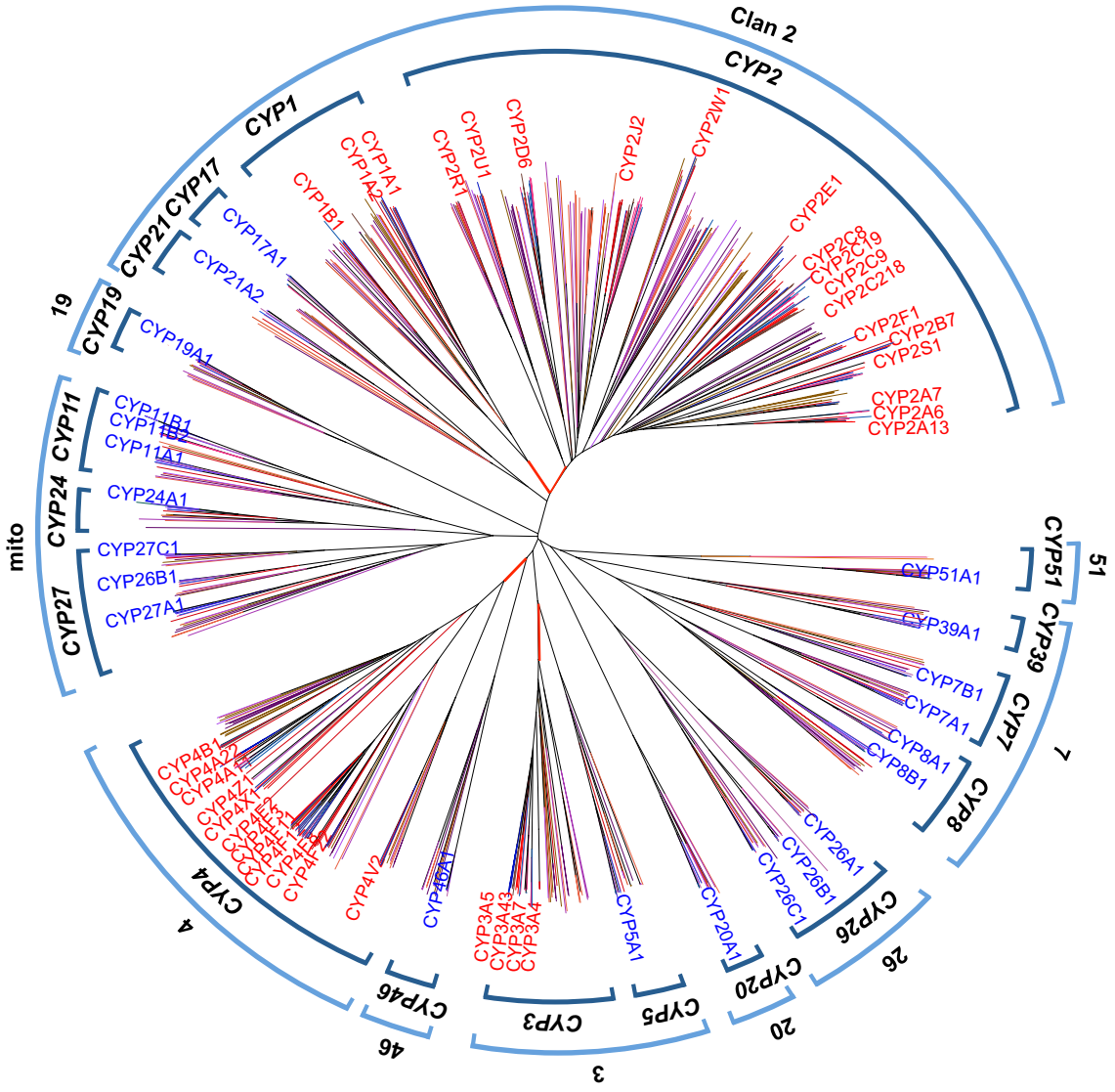
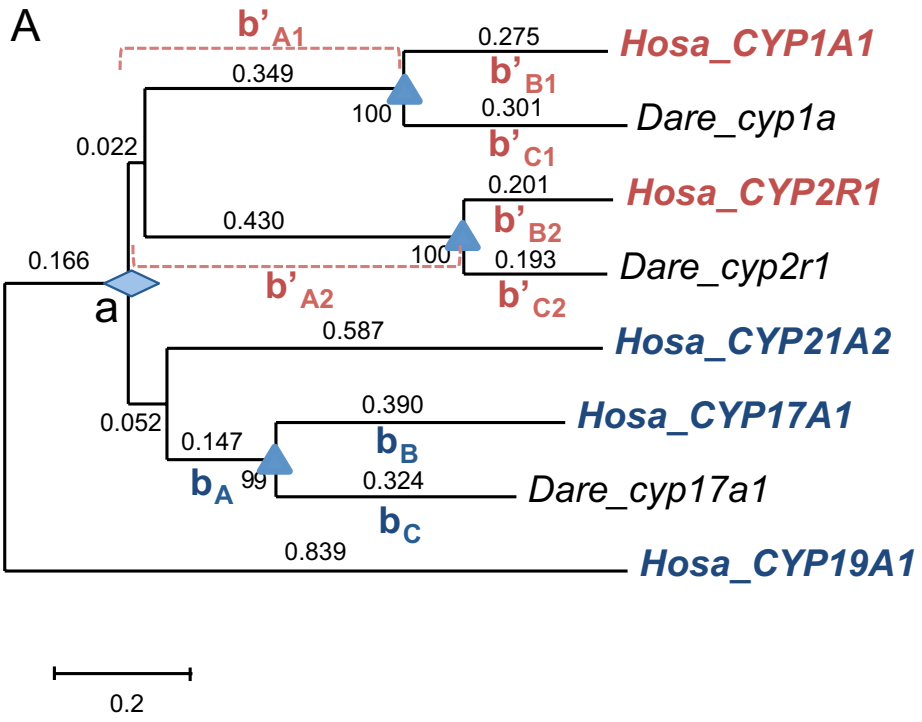
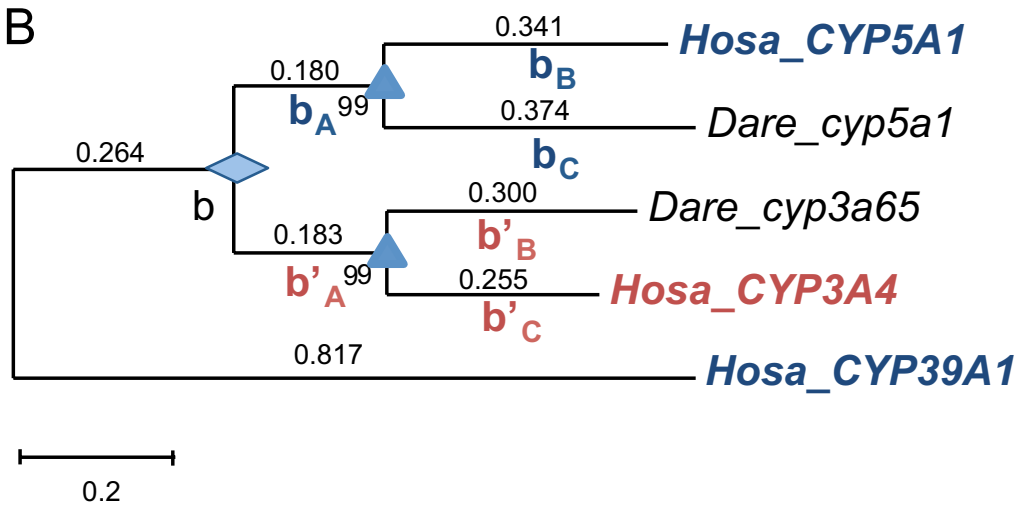
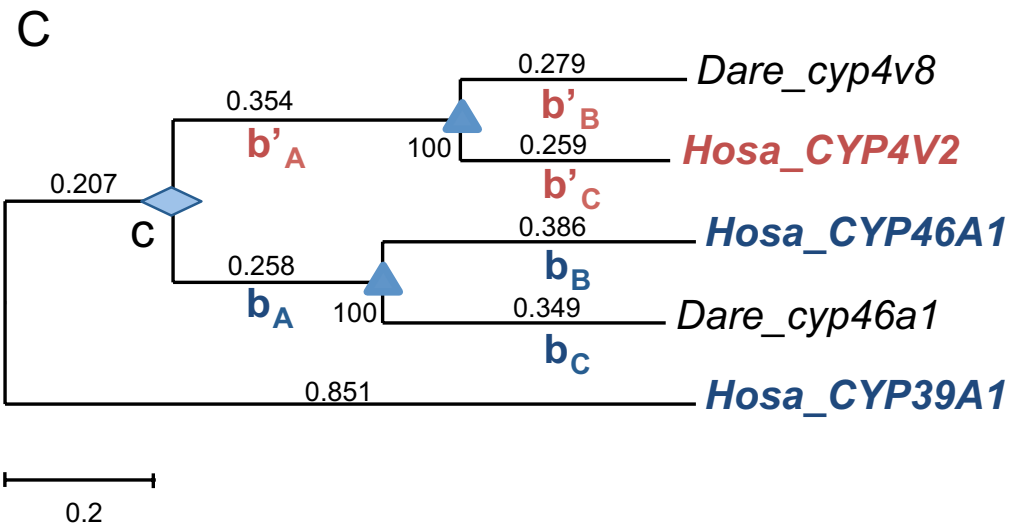

Clan 2
CYP2
CYP1
CYP17
CYP21
CYP19
19
CYP11
CYP24
CYP27
mito
CYP1A1
CYP1A2
CYP1B1
CYP2U1
CYP2R1
CYP2D6
CYP2J2
CYP2W1
CYP2E1
CYP2C8
CYP2C19
CYP2C9
CYP2C18
CYP2F1
CYP2B7
CYP2S1
CYP2A7
CYP2A6
CYP2A13
CYP17A1
CYP21A2
CYP19A1
CYP11B1
CYP11B2
CYP11A1
CYP24A1
CYP27C1
CYP26B1
CYP27A1
CYP51
51
CYP51A1
CYP39
CYP39A1
CYP7
CYP7B1
CYP7A1
CYP8
CYP8A1
CYP8B1
7
CYP26
CYP26A1
CYP26B1
CYP26C1
26
CYP20
CYP20A1
20
CYP5
CYP5A1
CYP3
CYP3A5
CYP3A43
CYP3A7
CYP3A4
3
CYP46
CYP46A1
46
CYP4
CYP4V2
CYP4B1
CYP4A22
CYP4A11
CYP4Z1
CYP4X1
CYP4F3
CYP4F11
CYP4F12
CYP4F2
4

Figure 3-1-6. Duplication time of B-type and D-type genes.

A) The divergence between *CYP1/2* and *17A1/21A2*. B) The divergence between *CYP3A* and *5A1*. C) The divergence between *CYP4* and *46A1*. The divergence between humans and zebrafish was used as a calibration time (= 400 mya), and is shown as a triangle in each tree. The duplication event is shown as a diamond. b_A, b_B, and b_C represent the branch length between the duplication event and species divergence. The branches b'_A, b'_B, and b'_C represent the length of D-type genes. The number near each branch shows the branch length.




B
0.341
Hosa_CYP5A1
0.180
b_B
b_A 99
0.374
Dare_cyp5a1
0.264
b_C
b
0.300
Dare_cyp3a65
0.183
b'_B
b'_A 99
0.255
Hosa_CYP3A4
b'_C
0.817
Hosa_CYP39A1
0.2


C
0.279
Dare_cyp4v8
0.354
b'_B
b'_A
100
0.259
Hosa_CYP4V2
0.207
b'_C
0.386
Hosa_CYP46A1
C
b_B
0.258
b_A
100
0.349
Dare_cyp46a1
b_C
0.851
Hosa_CYP39A1
0.2

3.2 Evolutionary relationship between invertebrate and vertebrate *CYP*s

Based on the calculation of divergence time, vertebrate D-type *CYP* gene had diverged from B-type around 601-1316 mya. To further examine the emergence time of an ancestor D-type in vertebrate and invertebrate, I searched for homologs of human D-type *CYP*s in six invertebrate species. In Cytochrome P450 Homepage, there are 33 invertebrate CYP nucleotide sequences. I used six invertebrate species (amphioxus: *B. floridae*, sea squirt: *C. intestinalis*, sea urchin: *S. purpuratus*, sea anemone: *N. vectensis*, water flea: *D. pulex*, and fruit fly: *D. melanogaster*) in this study. A total of 543 CYP amino acid sequences were retrieved from the Cytochrome P450 Homepage.

A preliminary search to determine the phylogenetic relationships between vertebrate and invertebrate *CYPs*, I compared nucleotide sequences of invertebrate genes with those of humans (Figure 3-2-1A, B, C, D, E and F). I examined whether the gene is belongs to B-type or D-type by comparing with human genes or *Drosophila* ones. It is already known that *Drosophila* has D-type genes (*CYP6U2*, *CYP6D2*). In hydra or daphnia, each species' *CYP* genes formed monophyletic cluster, although it was difficult to distinguish from vertebrate B- and D-types. Sea urchins seem to have a lot of D-like genes 33 genes) and a few B-types (8 genes). On the other hand, in Amphioxus, they have many D-type *CYP* genes and small number of B-types. Furthermore, I searched other invertebrate genomes, sponge and choanoflagellate, to determine the divergence time of D-type (Figure 3-2-1F and G). Sponge and choanoflagellate seem to have both B-type and D-like genes, although the genome information of them has not been fully completed. Therefore I couldn't determine the

detailed time of the birth of D-type. These observation mean that D-type *CYP* genes already have occurred before the divergence of vertebrate species and at least before the divergence of Deuterostomia, Figure 3-2-2.

In addition, I constructed the phylogenetic tree of vertebrates and invertebrates to determine the phylogenetic position of vertebrate *CYP*s in the tree. The tree included both vertebrate and invertebrate *CYP*s revealed that each vertebrate *CYP* family formed a monophyletic group. To simplify the phylogenetic analysis, amino acid sequences from these invertebrates were aligned only with sequences from humans, as a representative vertebrate, and the tree was constructed on the basis of amino acid distances (Figure 3-2-3).

The amino-acid distance tree shows that 10 clans (clans 2, 3, 4, mito, 7, 19, 20, 26, 46, and 51) are common to vertebrates; the tree also reveals one *Drosophila*-specific clan. A previous study of 1,572 *CYP* sequences also identified 11 clans in metazoans, but with inclusion of clan 74, which was present only in lancelets, sea anemones, and Trichoplax, but absent in vertebrates (Nelson DR *et al*., 2013). In the present analysis, despite the inclusion of both lancelet and sea anemone, clan 74 was not detected. However, a further phylogenetic analysis that included only yeasts, humans, lancelets, and sea anemones identified clan 74, although it was supported by a relatively low bootstrap value (55%). In addition, the genes that comprised the *Drosophila*-specific clan (*CYP6D2*, *6U1*, *28A5*, *28C1*, *28D1*, *308A1*, *309A1*, *350A1*, and *317A1*) were all included in clan 3 (Nelson DR *et al*. 2013). This holds true when I draw trees with different methods (maximum likelihood), although the bootstrap value for this clan is

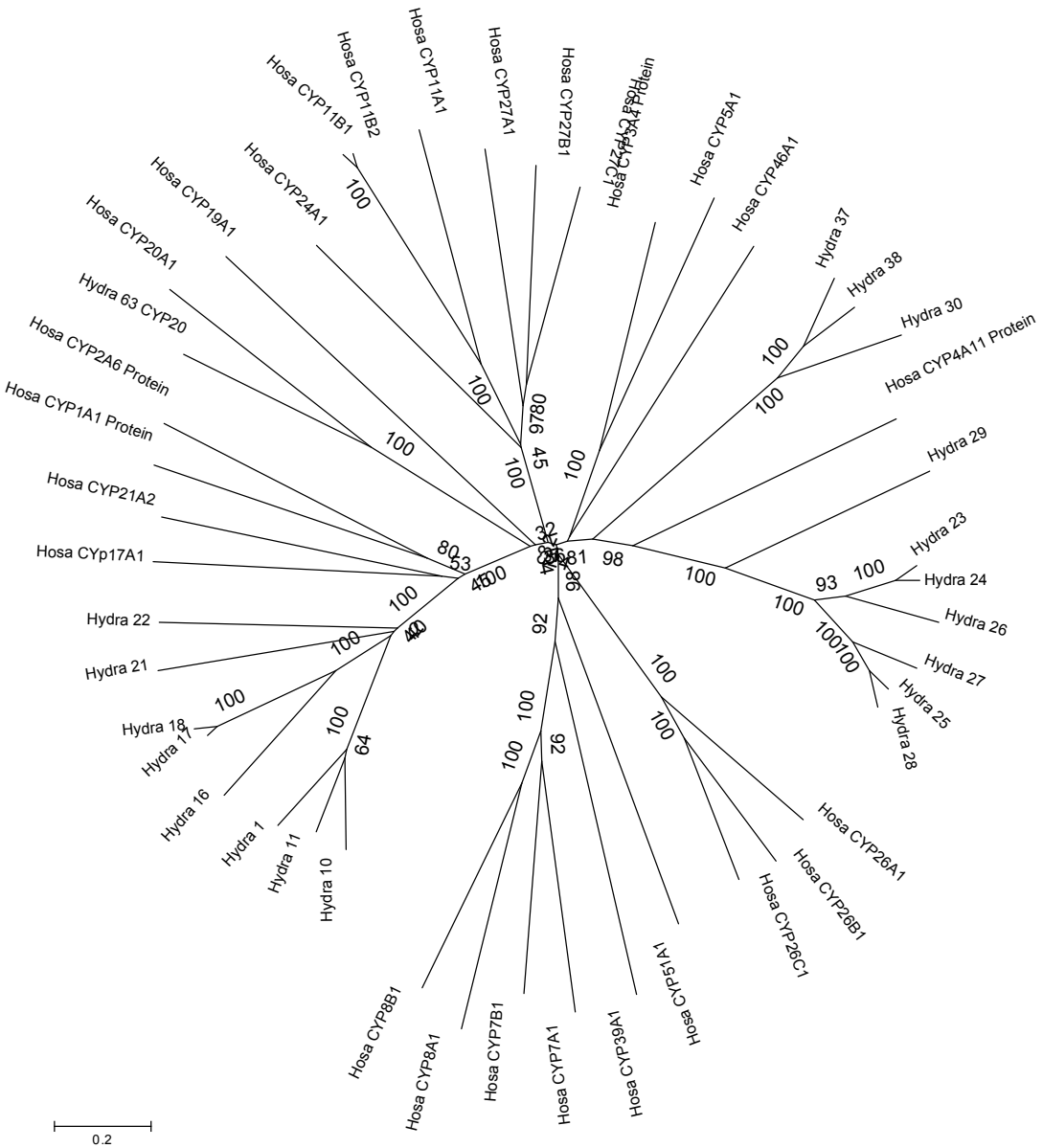
too low (<20%) to confirm this inclusion. I also observed some other differences from the previous study (Nelson DR *et al*., 2013): clan 51 did not include any sea-urchin gene, and clan 20 included neither sea urchin nor sea-anemone gene (Figure 3-2-2). The absence of a sea-urchin *CYP51* ortholog can be explained by the incompleteness of the database used here. In fact, a blast search of the NCBI database using human *CYP51* as a query identified a *CYP51* gene (Accession number: NM_001001906) in the recently published sea urchin genome. However, clan 20-like genes were absent from the sea urchin and the sea-anemone genomes in the database. In addition, clan 19 in the present tree appeared to include the *Drosophila* genes (*313A1*, *313B1*, *316A1*, and *318A1*) that were included in clan 4 in the previous study. In fact, the *Drosophila*-specific genes in clan 19 shared 16 of 433 amino acids with human *CYP19* (Figure 3-2-4), and these 16 amino acids were conserved among vertebrate CYPs. However, an ML tree supported the presence of the *Drosophila* sequences in clan 4, with very low bootstrap support (6%).

Clans including invertebrate *CYP* genes were supported by low bootstrap values, and clan definitions were dependent on the methods used for tree construction. Thus, the notion of clan becomes ambiguous and ill-defined for distantly related metazoan *CYP* genes.

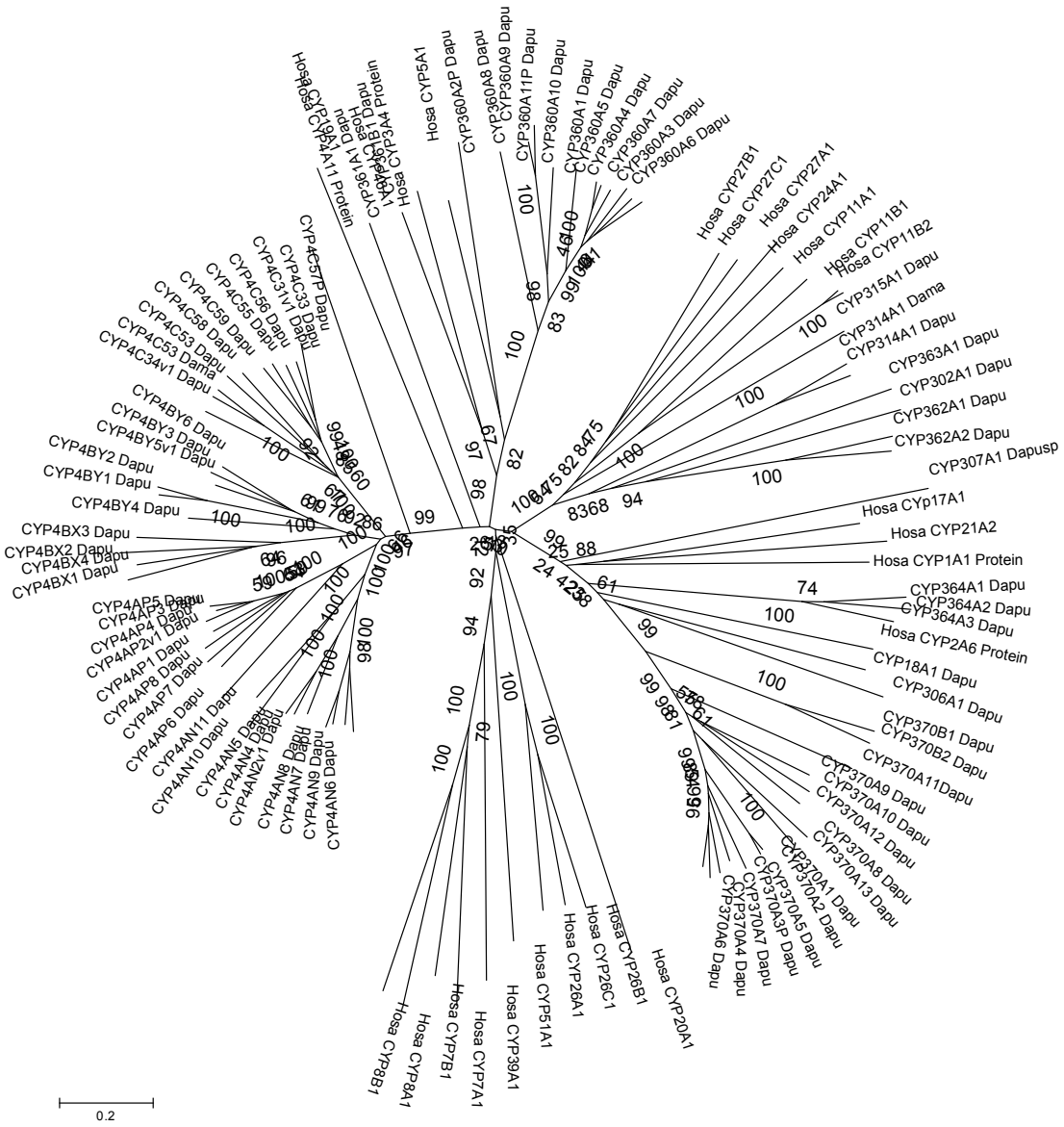
Figure 3-2-1. The NJ tree of human and each invertebrate species

Each tree is A) *Hydra*, B) *Daphnia pulex* (*Dapu*), C) *Drosophila melanogaster* (*Drme*), D) sea anemone (Seaan), E) sea urchin (GLEAN) and *Amphioxus* (*Amphi*) F) sponge, G) choanoflagellate (*Mobr*).

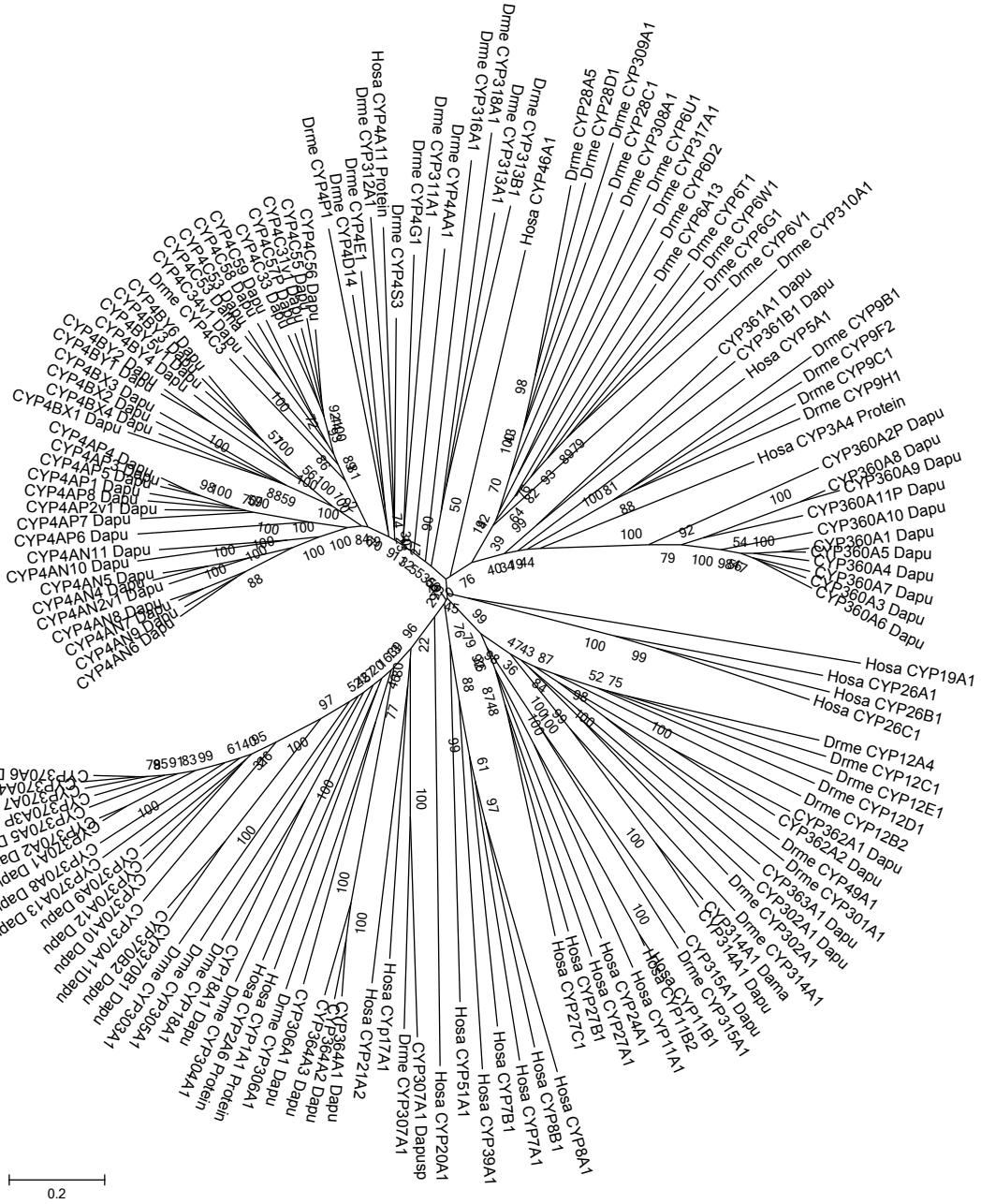
A) *Hydra*



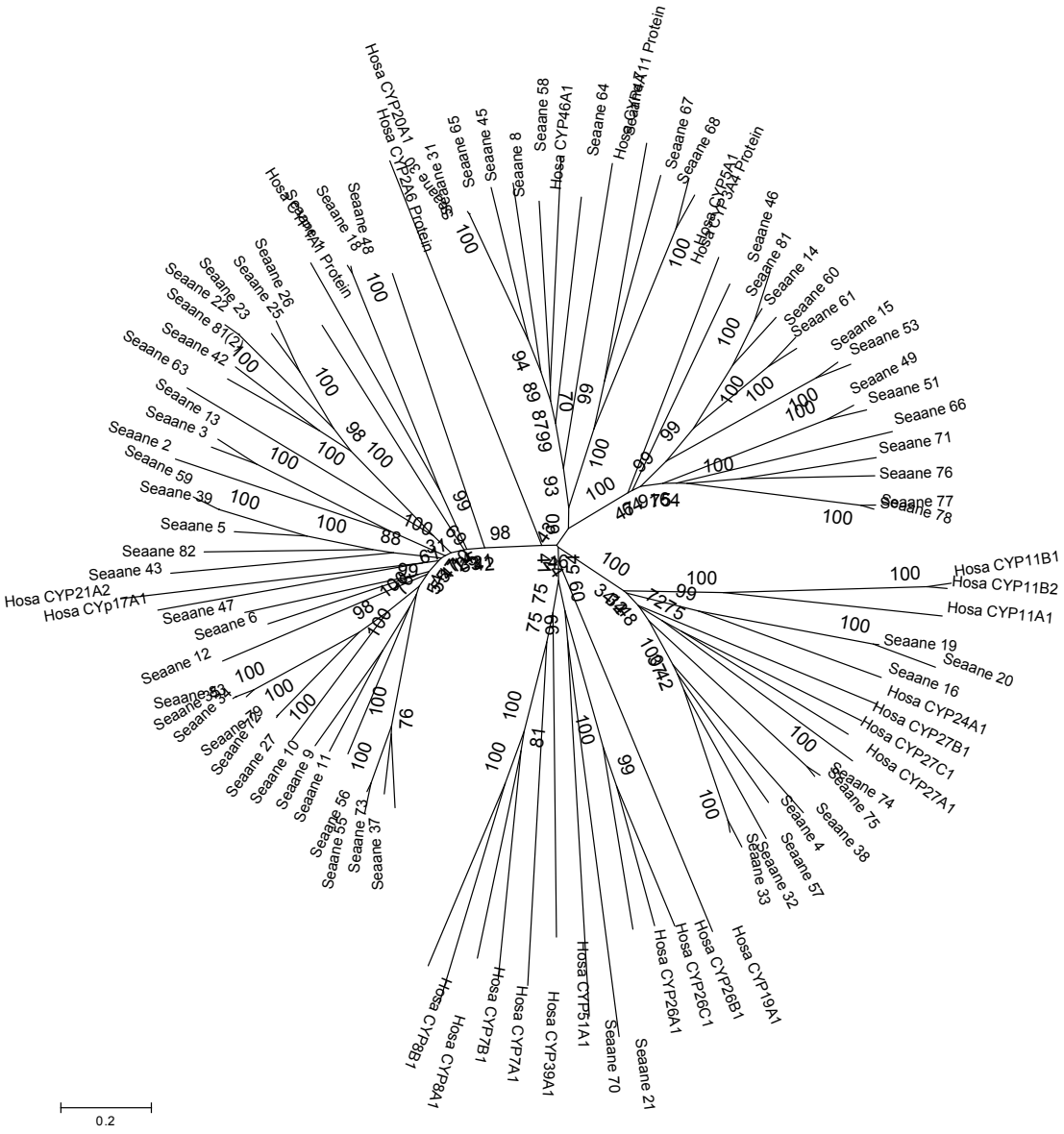
B) *Daphnia*


Hosa CYP5A1
Hosa CYP51A1 Protein
Hosa CYP39A1
Hosa CYP7A1
Hosa CYP7B1
Hosa CYP8A1
Hosa CYP8B1
Hosa CYP20A1
Hosa CYP26A1
Hosa CYP26B1
Hosa CYP26C1
Hosa CYP27B1
Hosa CYP27C1
Hosa CYP24A1
Hosa CYP11A1
Hosa CYP11B1
Hosa CYP11B2
Hosa CYP17A1
Hosa CYP21A2
Hosa CYP1A1 Protein
Hosa CYP2A6 Protein
CYP315A1 Dapu
CYP314A1 Dama
CYP314A1 Dapu
CYP363A1 Dapu
CYP302A1 Dapu
CYP362A1 Dapu
CYP362A2 Dapu
CYP307A1 Dapusp
CYP364A1 Dapu
CYP364A2 Dapu
CYP364A3 Dapu
CYP18A1 Dapu
CYP306A1 Dapu
CYP370B1 Dapu
CYP370B2 Dapu
CYP370A11Dapu
CYP370A9 Dapu
CYP370A10 Dapu
CYP370A12 Dapu
CYP370A8 Dapu
CYP370A13 Dapu
CYP370A1 Dapu
CYP370A5 Dapu
CYP370A2 Dapu
CYP370A3P Dapu
CYP370A7 Dapu
CYP370A4 Dapu
CYP370A6 Dapu
CYP360A1 Dapu
CYP360A5 Dapu
CYP360A4 Dapu
CYP360A7 Dapu
CYP360A3 Dapu
CYP360A6 Dapu
CYP360A8 Dapu
CYP360A9 Dapu
CYP360A10 Dapu
CYP360A11P Dapu
CYP360A2P Dapu
CYP4C57P Dapu
CYP4C33 Dapu
CYP4C31v1 Dapu
CYP4C56 Dapu
CYP4C55 Dapu
CYP4C59 Dapu
CYP4C58 Dapu
CYP4C53 Dapu
CYP4C53 Dama
CYP4C34v1 Dapu
CYP4BY6 Dapu
CYP4BY3 Dapu
CYP4BY5v1 Dapu
CYP4BY2 Dapu
CYP4BY1 Dapu
CYP4BY4 Dapu
CYP4BX3 Dapu
CYP4BX2 Dapu
CYP4BX4 Dapu
CYP4BX1 Dapu
CYP4AP5 Dapu
CYP4AP3 Dapu
CYP4AP4 Dapu
CYP4AP2v1 Dapu
CYP4AP1 Dapu
CYP4AP8 Dapu
CYP4AP7 Dapu
CYP4AP6 Dapu
CYP4AN11 Dapu
CYP4AN10 Dapu
CYP4AN5 Dapu
CYP4AN4 Dapu
CYP4AN2v1 Dapu
CYP4AN8 Dapu
CYP4AN7 Dapu
CYP4AN9 Dapu
CYP4AN6 Dapu
0.2

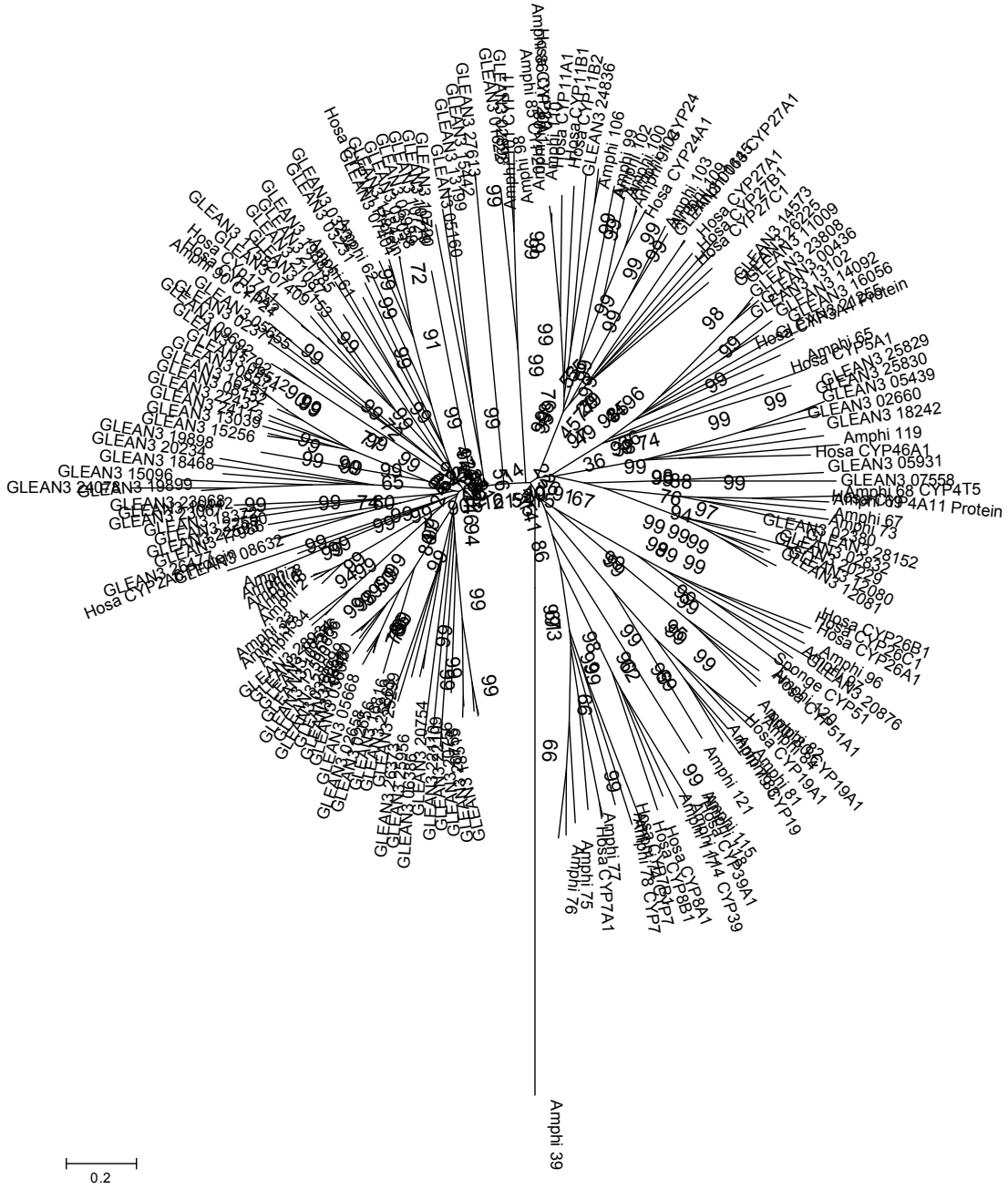
C) *Drosophila* and *Daphnia*



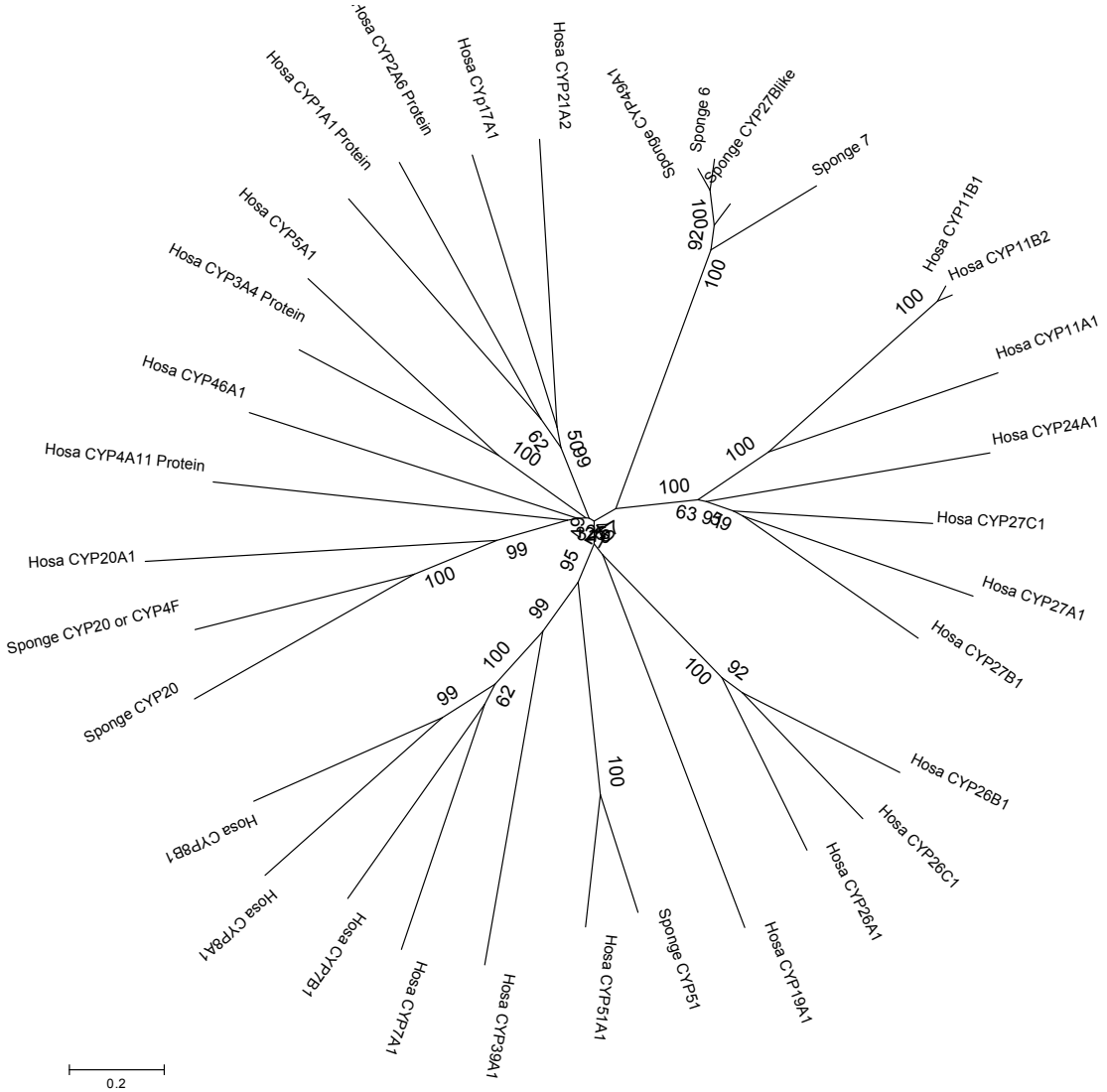
D) sea anemone



E) sea urchin and *Amphioxus*



F) sponge


Hosa CYP2A6 Protein
Hosa CYP17A1
Hosa CYP21A2
Sponge CYP4A9A1
Sponge 9
Sponge CYP27Blike
Sponge 7
Hosa CYP1A1 Protein
Hosa CYP5A1
Hosa CYP3A4 Protein
Hosa CYP46A1
Hosa CYP4A11 Protein
Hosa CYP20A1
Sponge CYP20 or CYP4F
Sponge CYP20
Hosa CYP8B1
Hosa CYP8A1
Hosa CYP7B1
Hosa CYP7A1
Hosa CYP39A1
Hosa CYP51A1
Sponge CYP51
Hosa CYP19A1
Hosa CYP26A1
Hosa CYP26C1
Hosa CYP26B1
Hosa CYP27B1
Hosa CYP27A1
Hosa CYP27C1
Hosa CYP24A1
Hosa CYP11A1
Hosa CYP11B2
Hosa CYP11B1
0.2

G) Choanoflagellate

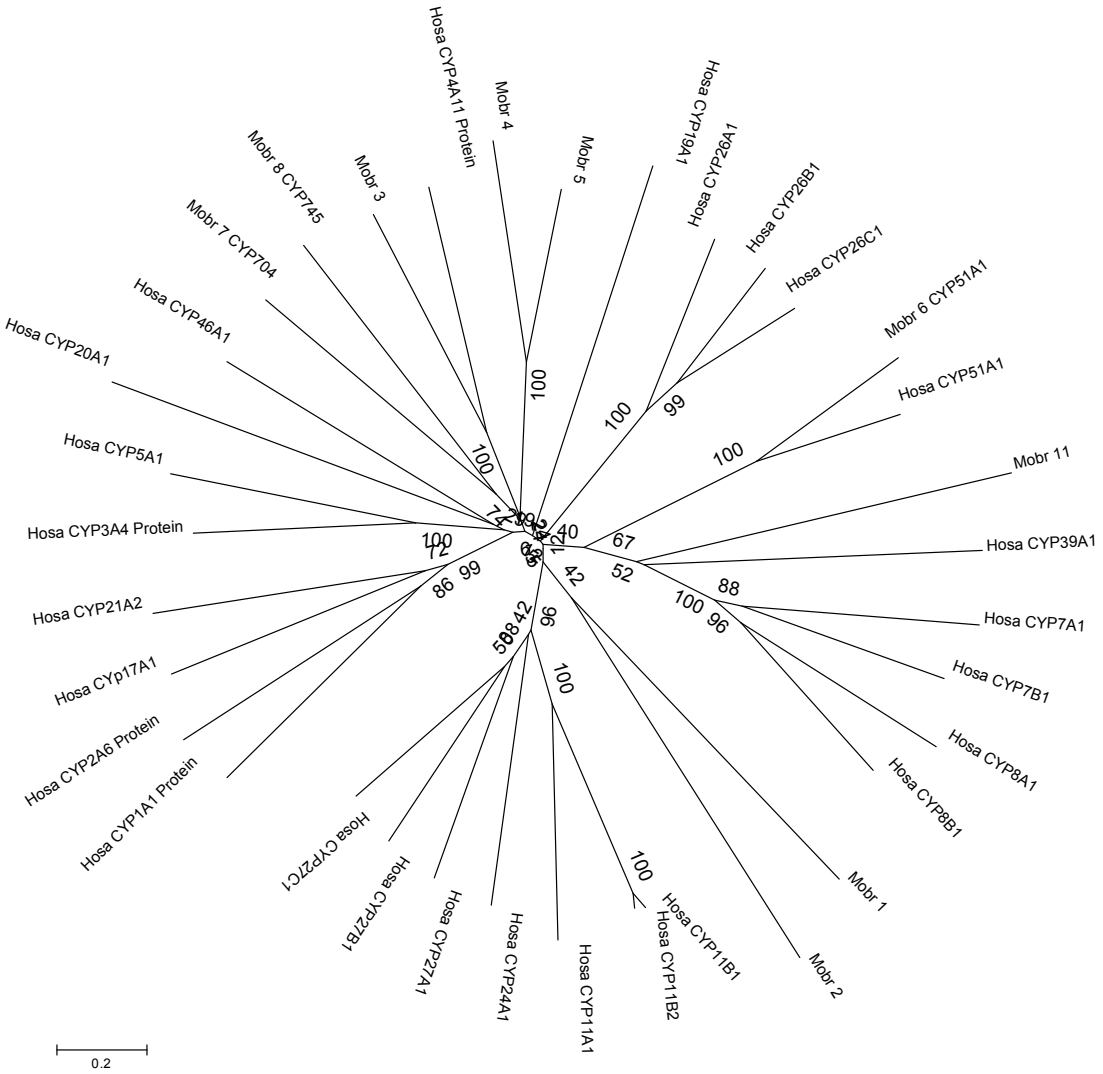


Figure 3-2-2. The estimated time of B- to D-type gene duplication.

The character of "B" and "D" represents B-type and D-type, respectively.

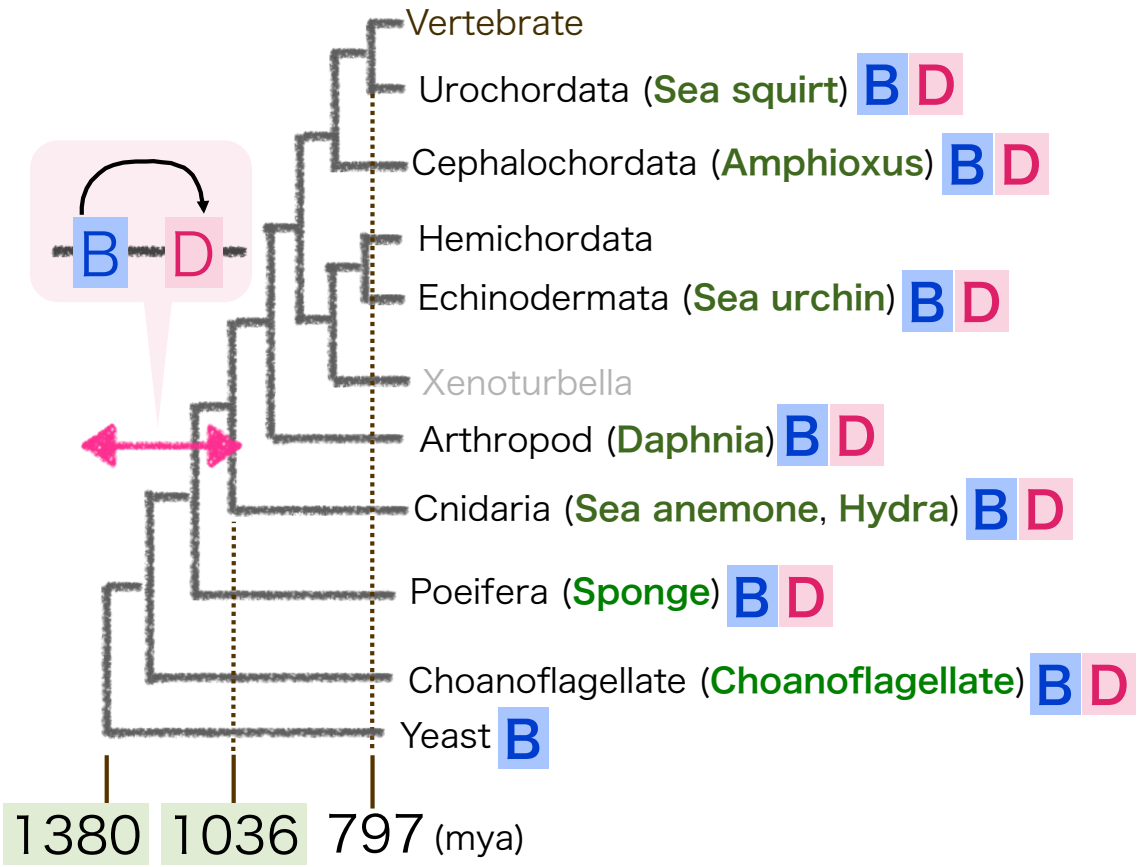


Figure 3-2-3. NJ tree of all invertebrate *CYP* genes.

D- and B-type *CYP* genes in humans are shown in red and blue text, respectively. The numbers near the brackets indicate clans. Orange character indicates the *Drosophila*-specific clan. Abbreviations and their color (in parentheses) are defined as follows: *Hosa*, *Homo sapien*s (red for D- and blue for B-type); *Brfl*, *Branchiostoma floridae* (dark brown); *Neve*, *Nematostella vectensis* (light blue); *Dapu*, *Daphnia pulex* (pink); *Stpu*, *Strongylocentrotus purpuratus* (aqua); *Ciin*, *Ciona intestinalis* (dark red); *Drme*, *Drosophila melanogaster* (brown); *Sace, Saccharomyces cerevisiae*; *Scpo*, *Schizosaccharomyces pombe* (purple). Gene names of *CYP6D2* and *6U1* in *D. melanogaster* are shown in clan 19. Bootstrap values supporting nodes of clusters mentioned in the text are shown.

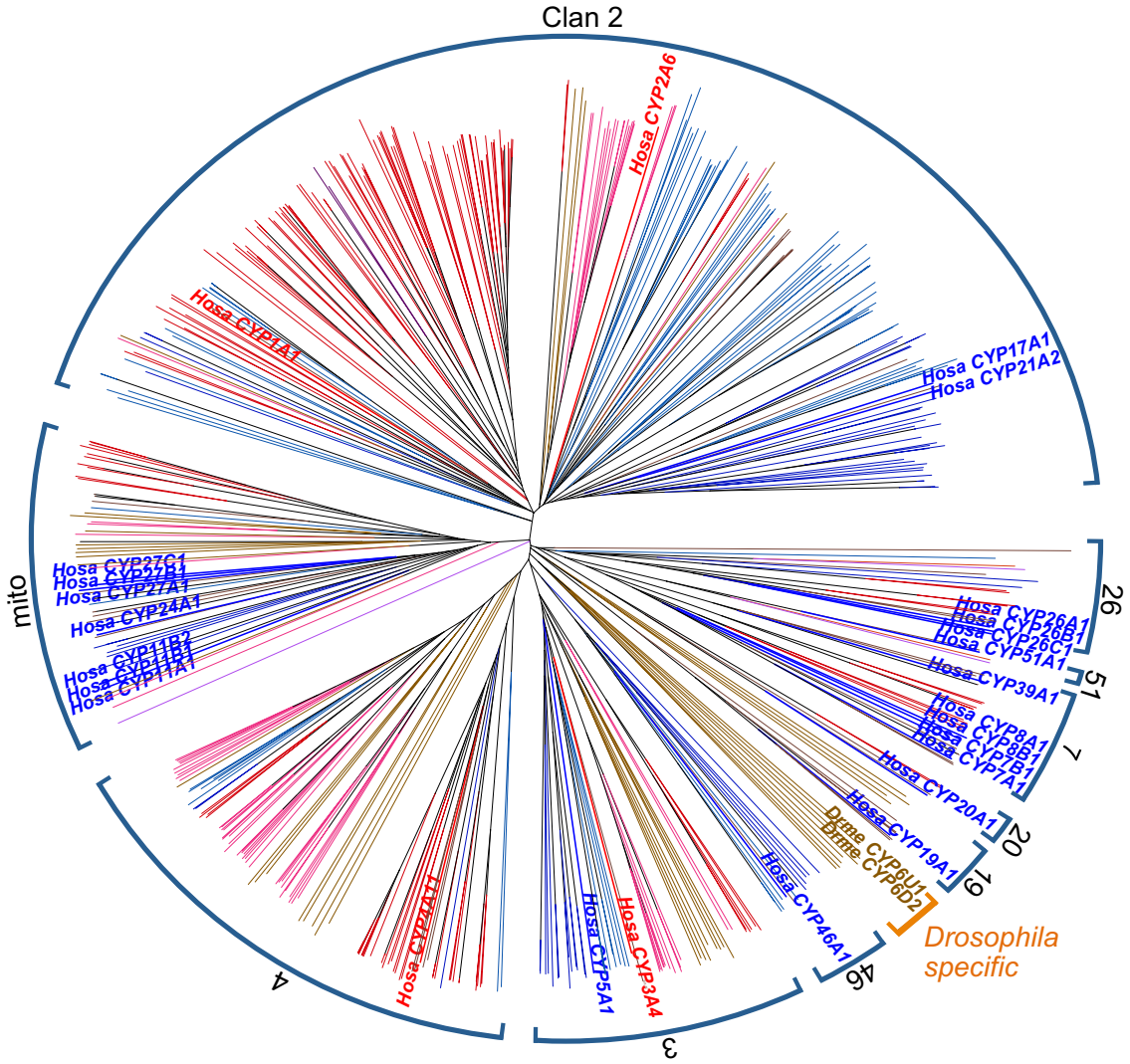
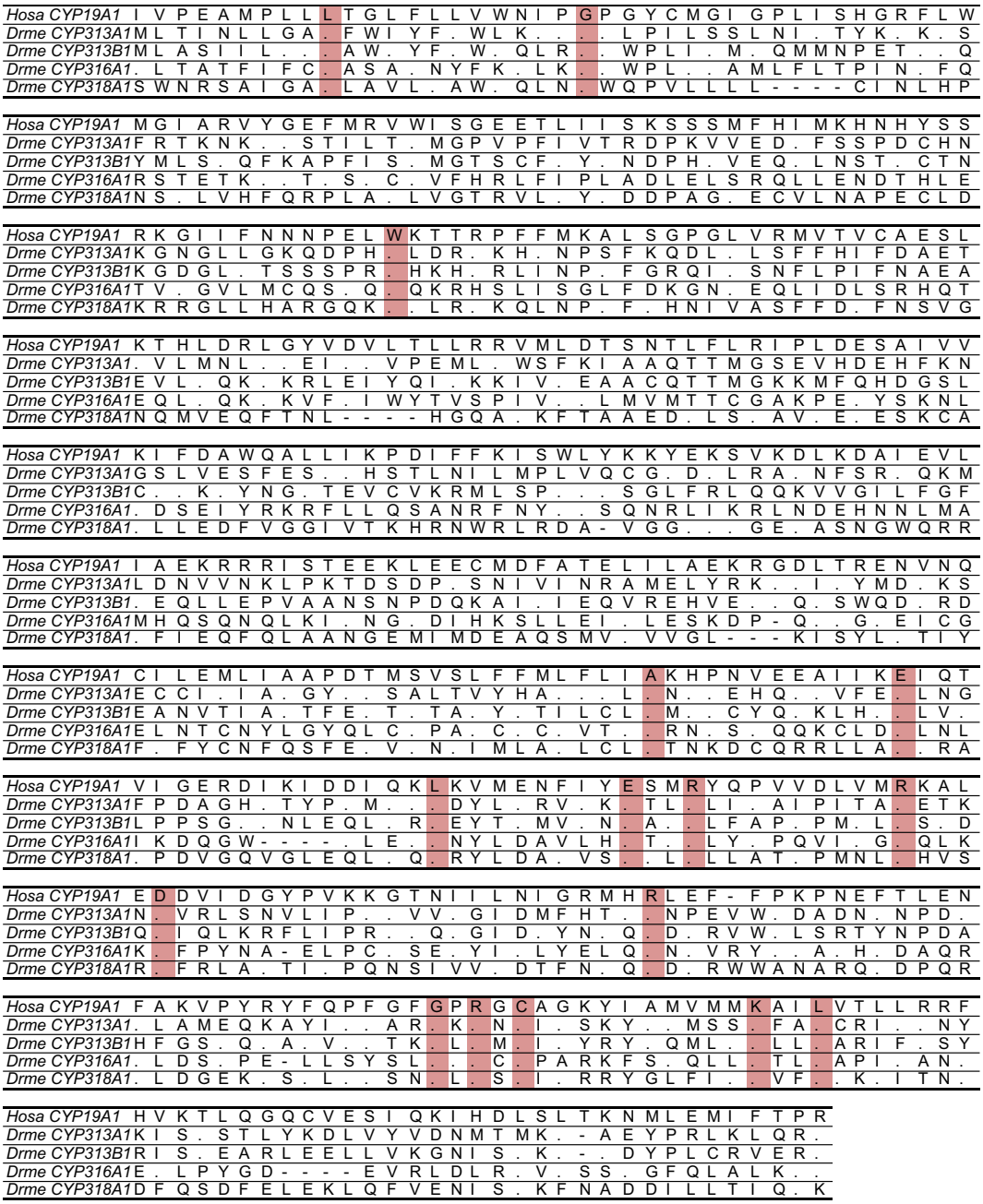

Clan 2
Hosa CYP2A6
Hosa CYP1A1
Hosa CYP17A1
Hosa CYP21A2
mito
Hosa CYP27C1
Hosa CYP27B1
Hosa CYP27A1
Hosa CYP24A1
Hosa CYP11B2
Hosa CYP11B1
Hosa CYP11A1
26
Hosa CYP26A1
Hosa CYP26B1
Hosa CYP26C1
Hosa CYP51A1
51
Hosa CYP39A1
7
Hosa CYP8A1
Hosa CYP8B1
Hosa CYP7B1
Hosa CYP7A1
20
Hosa CYP20A1
19
Hosa CYP19A1
Dme CYP6U1
Dme CYP6D2
Drosophila specific
46
Hosa CYP46A1
Hosa CYP3A4
Hosa CYP5A1
Hosa CYP2A13
4
3

Figure 3-2-4. Conserved amino acids between human and *Drosophila* genes in a *CYP19* clan.

Human (*CYP19A1*) and *Drosophila* (*CYP313A1*, *313B1*, *316A1*, and *318A1*) genes share the same amino acids at 16 of 423 aligned sites. The shared sites are shown in red.



3. 3 Origins of D-type genes are different between vertebrate and invertebrate

Clan 2 included human *CYP17A1* and *CYP21A2* (B-type) as well as members of the *CYP1* and *CYP2* families (D-type). Similarly, clan 3 included both types of *CYP* genes: *CYP5A1* (B-type) and *CYP3A* subfamilies (D-type). These two cases indicate that the emergence of D-type from B-type genes occurred after the emergence of the clan. However, clan 4 included only the *CYP4* family from humans but not *CYP46A1*, an ancestor of the *CYP4* family. This is the only case where the emergence of the D-type predates clan emergence. In addition, clan 4 included both vertebrate and invertebrate genes. Vertebrate *CYP4* likely acquired its detoxification function in the stem lineage of vertebrates when invertebrate sequences were B-type; alternatively, the ancestor of clan 4 may have already possessed D-type functions when invertebrate genes in clan 4 encoded D-type enzymes.

Fruit flies are known to possess two D-type *CYP* genes, *CYP6D2* and *CYP6U1*, which function in insecticide metabolism. In the tree generated in this study, these *CYP* genes were distantly related to human D-type genes, suggesting that the D-type genes in fruit flies emerged independently from those in vertebrates.

3. 4 Gene duplications and losses in the B- and D-type lineages during vertebrate evolution

Nearly all of the 14 vertebrate genomes examined here contained 21 orthologs to the 22 functional human B-type genes. On the basis of the presence or absence of *CYP* genes in each vertebrate genome, I parsimoniously estimated the number of genes

in each ancestor of amniotes, mammals, eutherian mammals, primates, catarrhini, and hominoids, as well as the number of gains of genes in each taxonomic lineage. The number of ancestral genes remained stable throughout the evolution of vertebrates: the number of genes in each vertebrate ancestor did not change over the course of evolution until the emergence of a primate ancestor. A gene-duplication event occurred in the primate ancestor, generating *CYP11B2.* In the ancestor of hominoids, emergence of new genes occurred twice, generating the ancestors of the *CYP51P1* and *CYP51P2* genes (Figure 3-4-1A).

In contrast to the rather stable mode of evolution observed in the stem, lineage-specific gains and losses of genes occurred relatively frequently. For instance, a shared duplication of *CYP19A1* occurred in the lineage leading to the common ancestor of zebrafish and medaka. In addition, lineage-specific gene duplications occurred in the zebrafish (*CYP8B2*, *CYP8B*3, *CYP17A1*, *CYP27A1*, and *CYP46A1*), medaka (*CYP46A1*), frog (*CYP8B1*, *CYP27A1*, and *CYP46A1*), green anole (*CYP24A1*), and opossum (*CYP8B1*) lineages. Interestingly, gene duplications of *CYP8B*, *CYP46A1*, and *CYP27A1* occurred independently several times in a species-specific manner. Similarly, lineage-specific gene losses (deletions) were observed; for instance, deletions occurred in a lineage leading to the medaka (*CYP7B1, CYP11B1*, and *CYP39A1*), frog (*CYP11B1*), green anole (*CYP11A1*, *CYP21A2*, *CYP26A1*), chicken (*CYP27B1*), zebra finch (*CYP11B1*, *CYP21A2* and *CYP27*), and opossum (*CYP17A1* and *CYP26B1*). Several deletions affecting the same genes (*CYP11B1* and *CYP21A2*) occurred independently in medaka, frog, green anole, and zebrafinch.

Although only a limited number of genomic sequences are available, I identified 19 gene gains and 16 losses among the 15 available genomes, including the human genome. Assuming that the total branch length in the vertebrate tree is 2,685 myr (for individual species divergence times, see Figure 3-4-2), I estimated the rate of gene gains and losses to be 0.7 and 0.6 per 100 myr, respectively.

Using a similar analysis, I also examined the 14 vertebrate genomes for the presence of paralogs and orthologs of 35 human D-type genes. This analysis revealed that the number of genes varies from 15 to 31 in ancestral species, and from 18 to 63 in extant species (Figure 3-4-1B). In contrast to the relatively stable evolutionary mode of B-type genes, D-type genes underwent more frequent gene duplications and pseudogenization (Table 3-1-3).

One important difference between D- and B-type genes is that D-type genes cluster on chromosomes, and these clusters are composed of closely related genes. This difference is reflected in the phylogeny, which shows that the genes in each cluster are monophyletic (Figure 3-1-5). In the human genome, five clusters have been identified: the *CYP2* family cluster on chromosome 19q (Hoffman SMG *et al*., 2006, Hu S *et al*., 2008), the *CYP2C* subfamily cluster on chromosome 10q, the *CYP3A* subfamily cluster on chromosome 7q, the *CYP4* family cluster on chromosome 1p, and the *CYP4F* subfamily clusters on chromosome 19p. Each cluster region occupies approximately 500 kb, with the exception of *CYP3A*, which occupies 250 kb. Each cluster included the following number of genes: 12 for *CYP2*, four for *CYP2C*, six for *CYP3A* and *CYP4*; and seven for *CYP4F* (Figure 3-4-3).

Using the phylogenic analysis of each *CYP1–4* family in vertebrates, I identified several species-specific gene duplications. The phylogenetic tree for the *CYP1* family revealed four subfamilies (*1A*, *1B*, *1C,* and *1D*), and showed that these subfamilies diverged in the ancestor of vertebrates. The *CYP1A* and *1B* subfamilies were conserved from fish to humans, whereas primates lacked *CYP1D*, and mammals lacked *CYP1C* (Figure 3-4-4A). In the tree of *CYP1A* and *1B*, the topology of each tree was same as species tree. The *CYP2* family was shown to be composed of 16 subfamilies (*CYP2A*, *2B*, *2C*, *2D*, *2E*, *2F*, *2G*, *2H*, *2J*, *2K*, *2R*, *2S*, *2T*, *2U*, *2W,* and *2AC*), three of which (*CYP2B*, *2E* and *2S*) were specific to mammals, while the *2A*/*G* and *F* subfamilies were present only in mammals and reptiles. These five subfamilies (except the *CYP2E* subfamily) diverged successively to form the *CYP2* cluster in an ancestor of mammals (Figure 3-4-4B). However, *CYP2U* and *2R* were shown to be common to all vertebrates. The *CYP3* family tree contained only two subfamilies, *CYP3A* and the fish-specific *3C* family (Figure 3-4-4C). *CYP3A* comprised amphibian-, bird-, and mammal-specific clades. In each taxonomic group, members of the *CYP3A* subfamily appear to have been duplicated independently. The tree constructed for the *CYP4* family included six subfamilies (*4A, 4B*, *4F*, *4V*, *4X*, and *4Z*) (Figure 3-4-4D). *CYP4A* and *4X/Z* were specific to mammals, whereas the other three subfamilies (*4B*, *F*, and *V*) were common to all vertebrates. In particular, the members of the *4F* subfamily formed several species-specific clusters, except *CYP4F22*. It is unclear, however, whether these species-specific clusters resulted from gene conversion or from recent duplication of the subfamily in each species. The evolution of D-type genes has

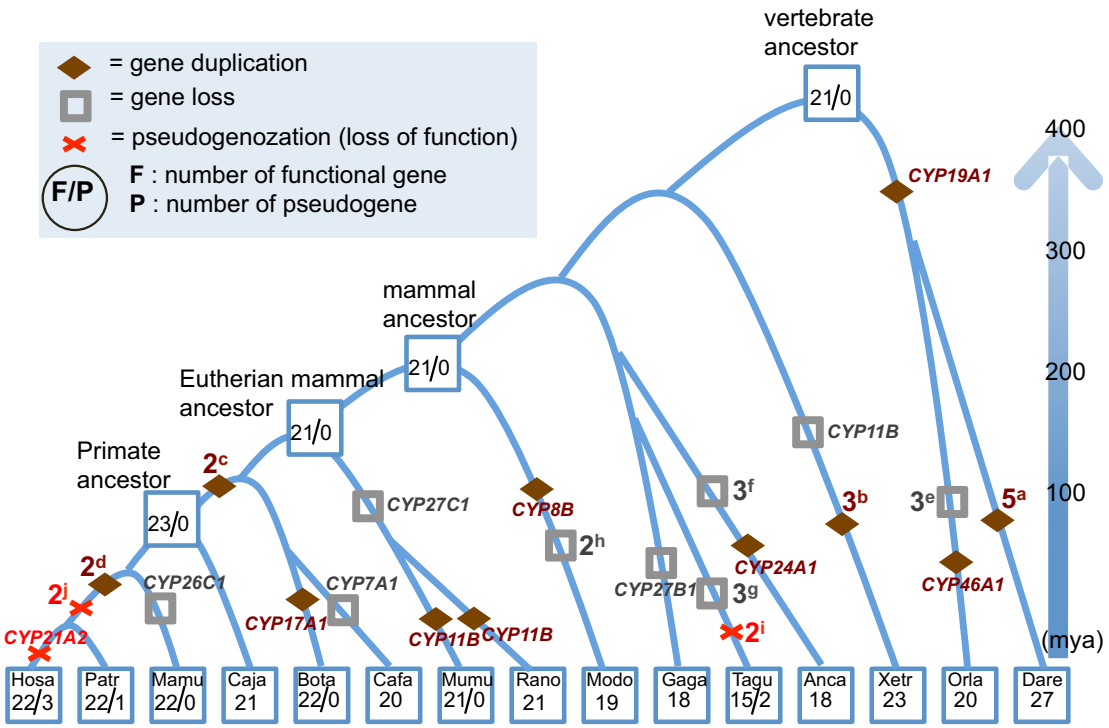
involved frequent species-specific gene duplications, compared to B-type genes (Figure 3-4-1B). The divergence time of each D-type cluster gene was estimated to be 350 million years ago for *CYP2* family, 248 mya for *CYP3A* subfamily, 217 mya for *CYP4* family, 144 mya for *CYP2C* subfamily and 106 mya for *CYP4F* subfamily, and they are shown in Figure 3-4-5A. Summarizing above results, the evolutionary scheme of D-type *CYP* gene is proposed as shown in Figure 3-4-5B, C, D, E.

In D-type genes, it is unclear how many gene duplications occurred before eutherian divergence. I estimated the rate of duplication subsequent to the eutherian radiation, which revealed 53 duplications in 432 myr, or the rate of 12.7 duplications per 100 myr. No deletion was observed. These results are in contrast to the results for B-type genes.

Figure 3-4-1 The birth and death processes of *CYP* genes in vertebrates.

A) B-type *CYP* genes and B) D-type *CYP* genes. In both figures, numbers inside squares represent the number of functional genes and pseudogenes in each species and its ancestors. Diamonds, crosses, and rectangles indicate gene duplication, pseudogenization, and deletion events, respectively. The number adjacent to each symbol represents the number of events. The letter adjacent to the number indicates the list of *CYP* genes, as follows. a: *CYP8B2*, *CYP8B3*, *CYP17A1*, *CYP27A1,* and *CYP46A1*, b: *CYP8B*, *CYP27A1,* and *CYP46A1*, c: *CYP11B* and *CYP21A2*, d: *CYP51* (two genes), e: *CYP7B*, *CYP11B,* and *CYP39A1*, f: *CYP11A*, *CYP21A2,* and *CYP26*, g: *CYP11B*, *CYP21A2,* and *CYP27* (A or B), h: *CYP17A1* and *CYP26B1*, i: *CYP24A1* and *CYP27A1*, j: *CYP21A1P* and two *CYP51*, k: *CYP4F9P*, *CYP4F23P,* and *CYP4F24P*, l: *CYP4A11* and *CYP2F1P*, m: *CYP2T2P*, *CYP2T3P,* and *CYP2G1P*, n: *CYP4A11*, *CYP4B1*, *CYP4F22*-like (two genes), and *CYP4F23P,* o: *CYP2A7P1*, *CYP2A13*, *CYP2B6P,* and *CYP4F11*, p: *CYP2B7P1*, *CYP2D8P1*, *CYP2F1P*, *CYP4F9P*, *CYP4F23P,* and *CYP4F24P*, **CYP1D1P* were found in *Pan paniscus*, but were absent from *Pan troglodytes*, q: *CYP2B6* and *CYP2C18*, r: *CYP2A7P1*, *CYP2G2P,* and *CYP4Z2P*.

A


= gene duplication
= gene loss
= pseudogenozation (loss of function)
F/P
F : number of functional gene
P : number of pseudogene
vertebrate ancestor
21/0
mammal ancestor
21/0
Eutherian mammal ancestor
21/0
Primate ancestor
23/0
CYP19A1
CYP11B
CYP27C1
CYP8B
CYP24A1
CYP27B1
CYP46A1
CYP26C1
CYP7A1
CYP17A1
CYP11B
CYP11B
CYP21A2
400
300
200
100
(mya)
Hosa 22/3
Patr 22/1
Mamu 22/0
Caja 21
Bota 22/0
Cafa 20
Mumu 21/0
Rano 21
Modo 19
Gaga 18
Tagu 15/2
Anca 18
Xetr 23
Orla 20
Dare 27

B

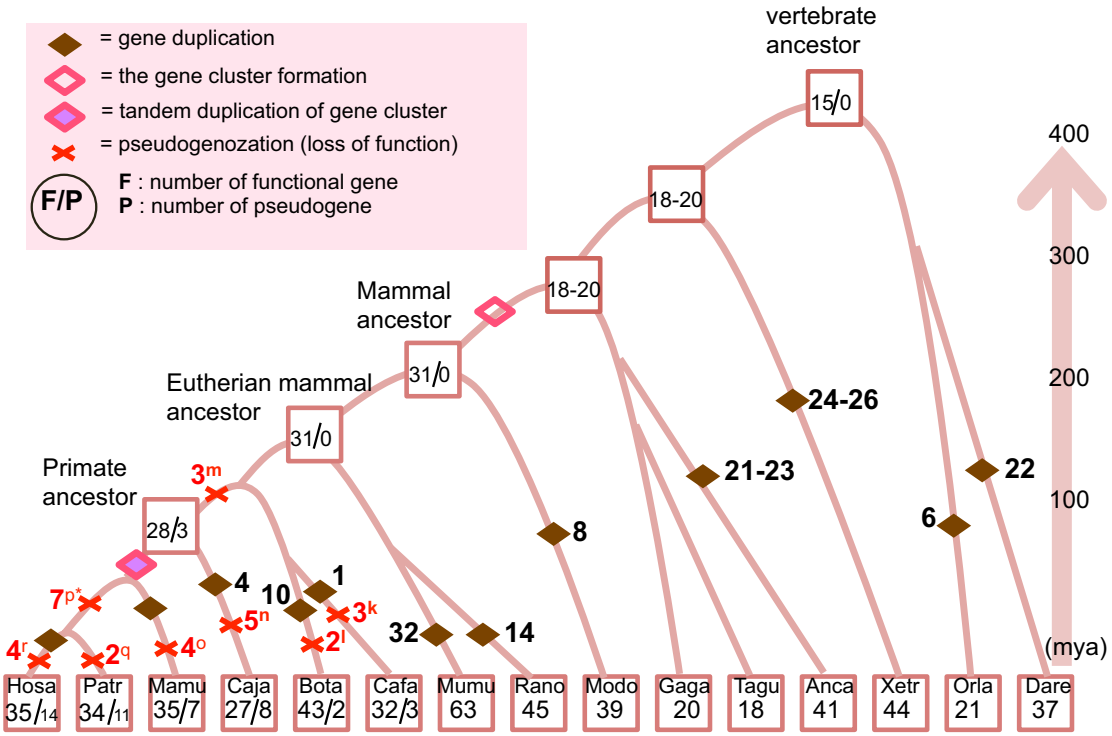

= gene duplication
= the gene cluster formation
= tandem duplication of gene cluster
= pseudogenozation (loss of function)
F/P
F : number of functional gene
P : number of pseudogene
vertebrate ancestor
15/0
18-20
Mammal ancestor
18-20
31/0
Eutherian mammal ancestor
31/0
Primate ancestor
28/3
24-26
21-23
22
8
6
400
300
200
100
(mya)
Hosa 35/14
Patr 34/11
Mamu 35/7
Caja 27/8
Bota 43/2
Cafa 32/3
Mumu 63
Rano 45
Modo 39
Gaga 20
Tagu 18
Anca 41
Xetr 44
Orla 21
Dare 37

Figure 3-4-2. Phylogenetic tree and species-divergence time.

The phylogenetic tree was constructed on the basis of the divergence time of 15 species (time tree). The species names at the tip of the tree are abbreviated as in Table 3-1-3. *Hosa*, *Homo sapiens*. The number at each node represents species divergence time in mya. The scale under the tree indicates time in myr.

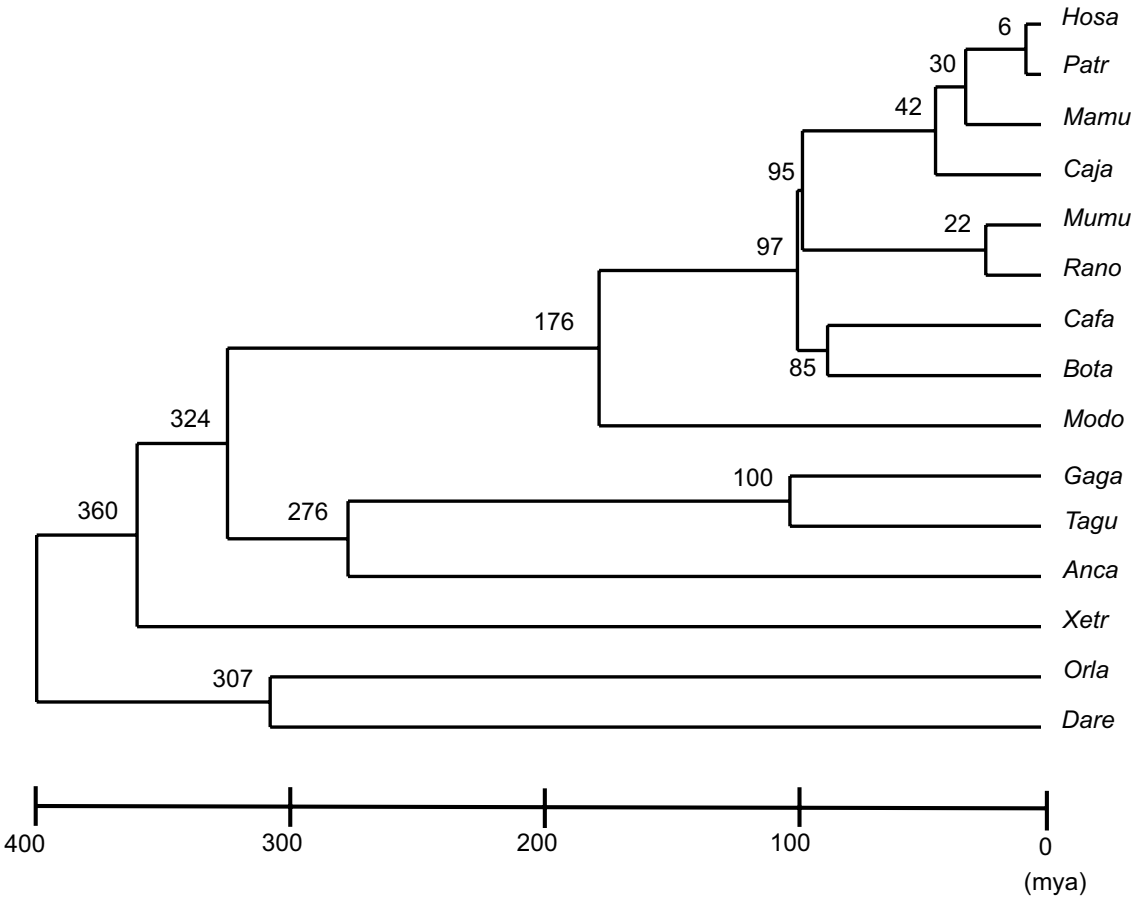


Table3-4-1. Conserved B-type *CYP* genes in vertebrates.

gene	function
CYP5A1	Thromboxane
CYP19A1	aromatase, estrogene
CYP20A1	unknown
CYP24A1	Vitamin D3
CYP26C1	Retinoic acid
CYP27A1	Bile acid, Vitamin D3
CYP51A1	Ranosterole

Figure 3-4-3. *CYP* gene clusters in the human genome.

A striped arrow represents an anchor gene in a syntenic region of each cluster. Black arrows and dotted arrows represent functional *CYP* genes and pseudogenes, respectively. The length of each gene cluster is approximately 500 kb, except the *CYP3A* cluster (250 kb). The number of total genes, functional genes (F), and pseudogenes (P) in each cluster are shown after the cluster name.

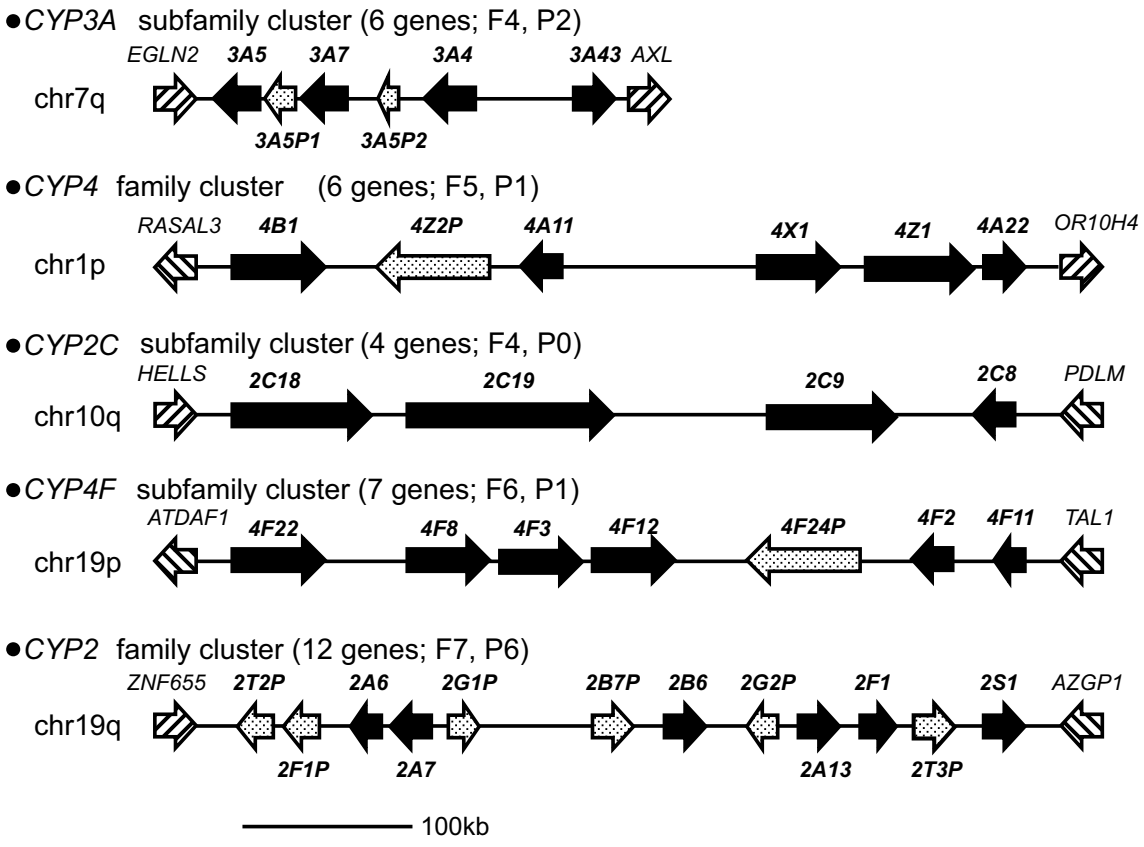
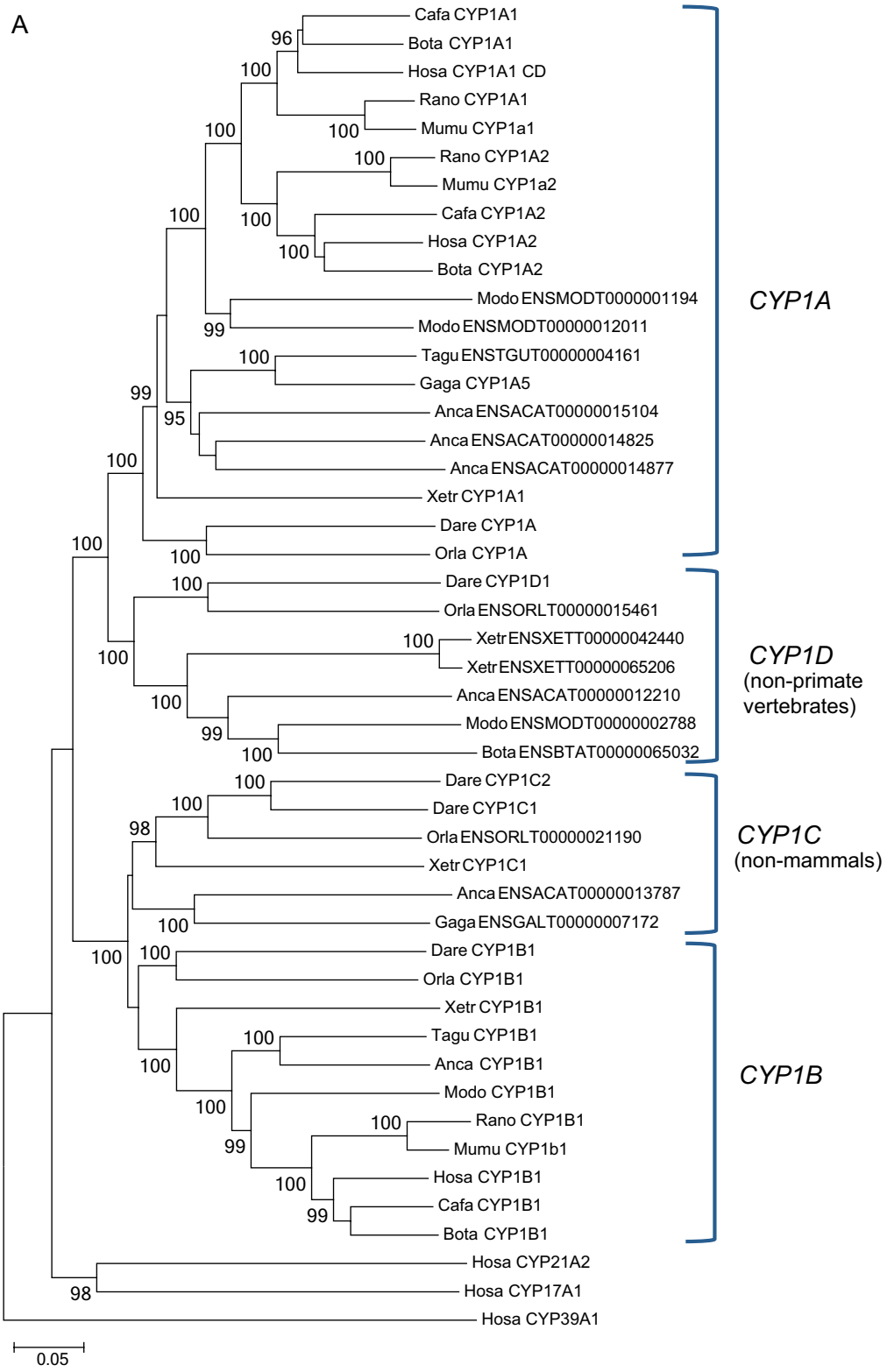
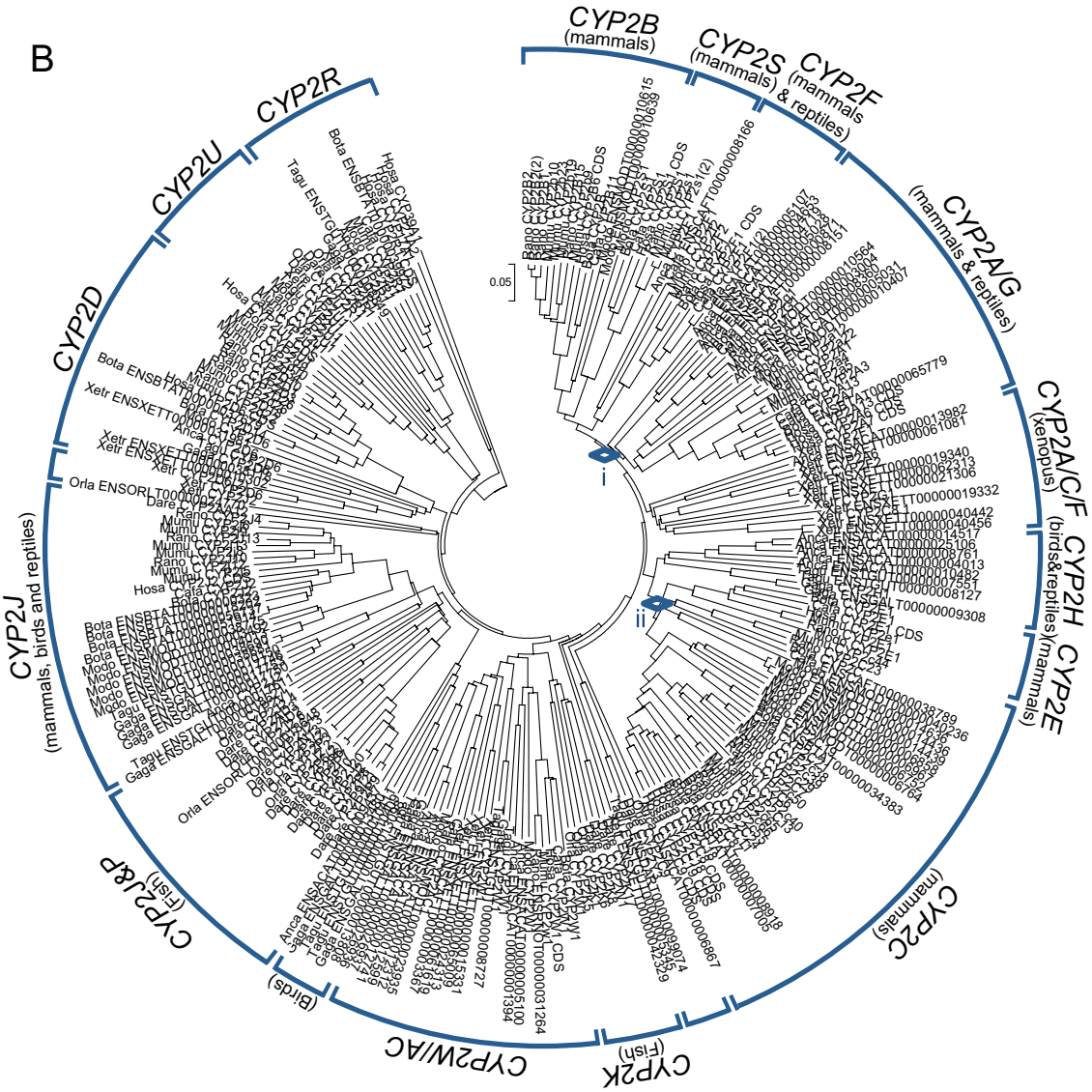
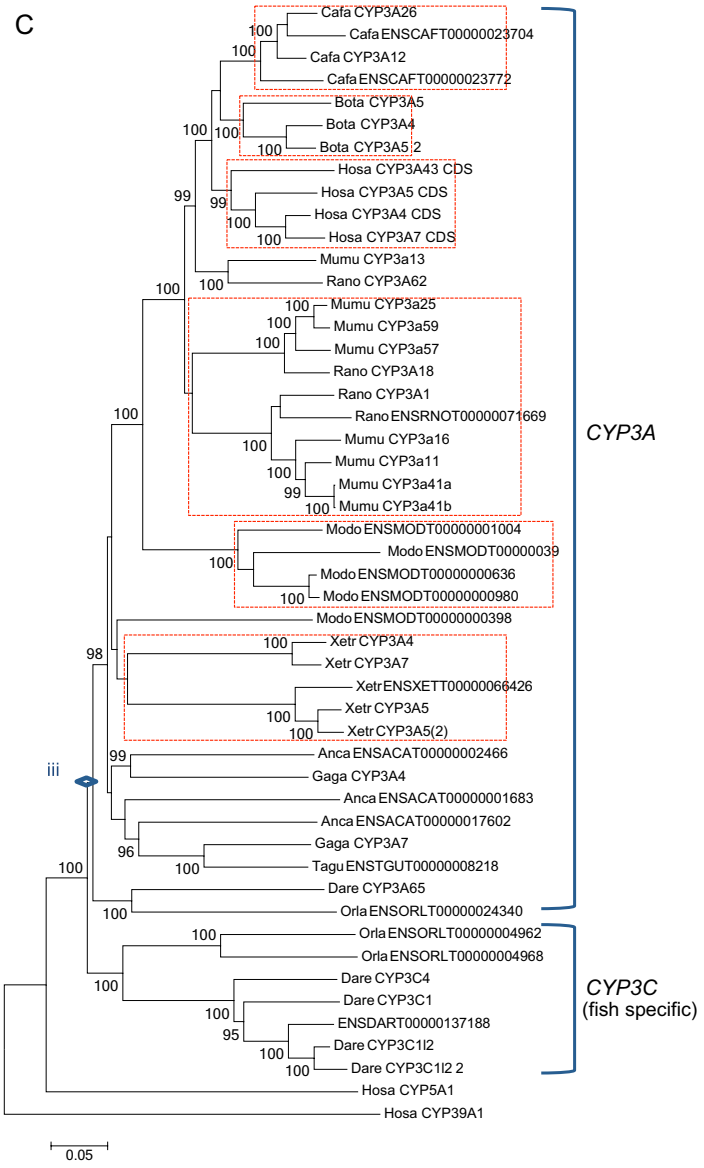


Figure 3-4-4. Phylogenetic tree of the D-type family.

A) *CYP1* family, B) *CYP2* family, C) *CYP3* family, and D) *CYP4* family. Each NJ tree was based on the total nucleotide substitutions among members. The origin of each of the five clusters (corresponding to i–v in Figure 3-1-3) is indicated with a diamond in Figure B–D. Each subfamily is indicated by a bracket. In Figure B, the *CYP2T* subfamily is not shown because no functional gene belonging to this subfamily is present in the human genome. In Figure C and D, the red dashed rectangle outlines a specific clade.


A
Cafa CYP1A1
Bota CYP1A1
Hosa CYP1A1 CD
Rano CYP1A1
Mumu CYP1a1
Rano CYP1A2
Mumu CYP1a2
Cafa CYP1A2
Hosa CYP1A2
Bota CYP1A2
Modo ENSMODT00000001194
Modo ENSMODT00000012011
Tagu ENSTGUT00000004161
Gaga CYP1A5
Anca ENSACAT00000015104
Anca ENSACAT00000014825
Anca ENSACAT00000014877
Xetr CYP1A1
Dare CYP1A
Orla CYP1A
CYP1A
Dare CYP1D1
Orla ENSORLT00000015461
Xetr ENSXETT00000042440
Xetr ENSXETT00000065206
Anca ENSACAT00000012210
Modo ENSMODT00000002788
Bota ENSBTAT00000065032
CYP1D
(non-primate
vertebrates)
Dare CYP1C2
Dare CYP1C1
Orla ENSORLT00000021190
Xetr CYP1C1
Anca ENSACAT00000013787
Gaga ENSGALT00000007172
CYP1C
(non-mammals)
Dare CYP1B1
Orla CYP1B1
Xetr CYP1B1
Tagu CYP1B1
Anca CYP1B1
Modo CYP1B1
Rano CYP1B1
Mumu CYP1b1
Hosa CYP1B1
Cafa CYP1B1
Bota CYP1B1
CYP1B
Hosa CYP21A2
Hosa CYP17A1
Hosa CYP39A1
0.05


B
CYP2B
(mammals)
CYP2S
(mammals)
CYP2F
(mammals
& reptiles)
CYP2A/G
(mammals & reptiles)
CYP2A/C/F
(xenopus)
CYP2H
(birds&reptiles)
CYP2E
(mammals)
CYP2C
(mammals)
CYP2K
(Fish)
CYP2W/AC
(Birds)
CYP2J&P
(Fish)
CYP2J
(mammals, birds and reptiles)
CYP2D
CYP2U
CYP2R
0.05
i
ii


C
Cafa CYP3A26
Cafa ENSCAFT00000023704
Cafa CYP3A12
Cafa ENSCAFT00000023772
Bota CYP3A5
Bota CYP3A4
Bota CYP3A5 2
Hosa CYP3A43 CDS
Hosa CYP3A5 CDS
Hosa CYP3A4 CDS
Hosa CYP3A7 CDS
Mumu CYP3a13
Rano CYP3A62
Mumu CYP3a25
Mumu CYP3a59
Mumu CYP3a57
Rano CYP3A18
Rano CYP3A1
Rano ENSRNOT00000071669
Mumu CYP3a16
Mumu CYP3a11
Mumu CYP3a41a
Mumu CYP3a41b
Modo ENSMODT00000001004
Modo ENSMODT00000039
Modo ENSMODT00000000636
Modo ENSMODT00000000980
Modo ENSMODT00000000398
Xetr CYP3A4
Xetr CYP3A7
Xetr ENSXETT00000066426
Xetr CYP3A5
Xetr CYP3A5(2)
iii
Anca ENSACAT00000002466
Gaga CYP3A4
Anca ENSACAT00000001683
Anca ENSACAT00000017602
Gaga CYP3A7
Tagu ENSTGUT00000008218
Dare CYP3A65
Orla ENSORLT00000024340
Orla ENSORLT00000004962
Orla ENSORLT00000004968
Dare CYP3C4
Dare CYP3C1
ENSDART00000137188
Dare CYP3C1l2
Dare CYP3C1l2 2
Hosa CYP5A1
Hosa CYP39A1
0.05
CYP3A
CYP3C
(fish specific)

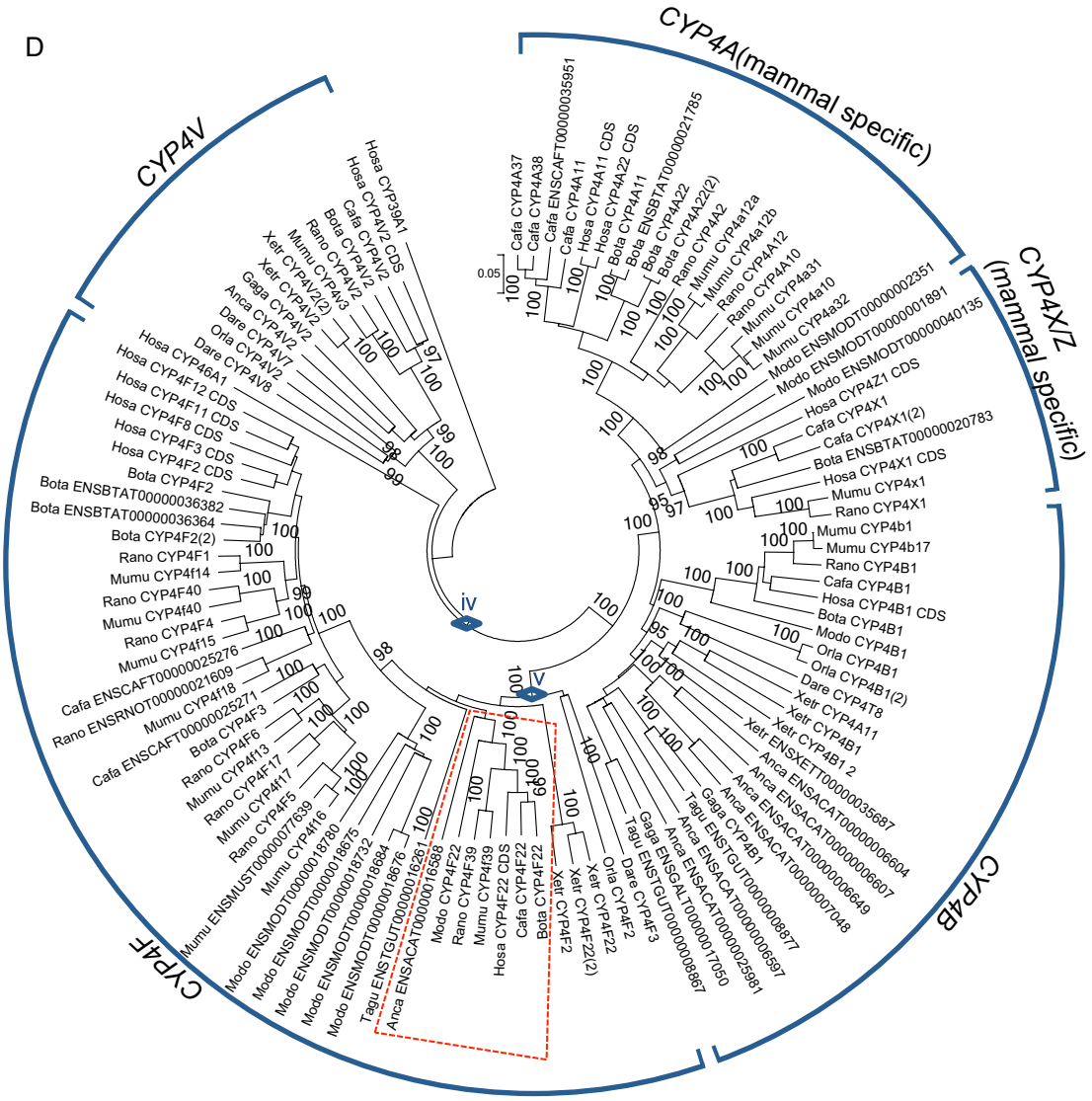
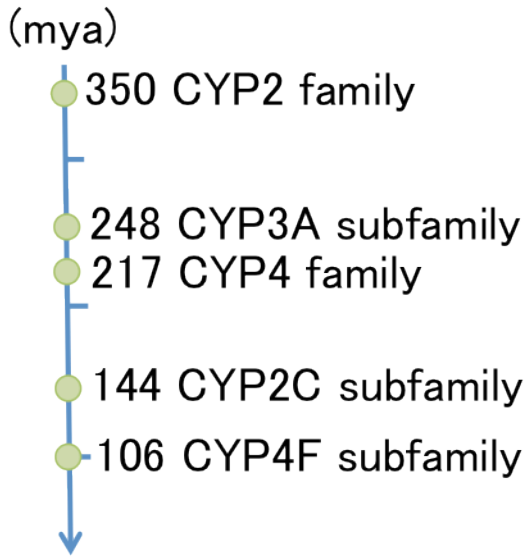
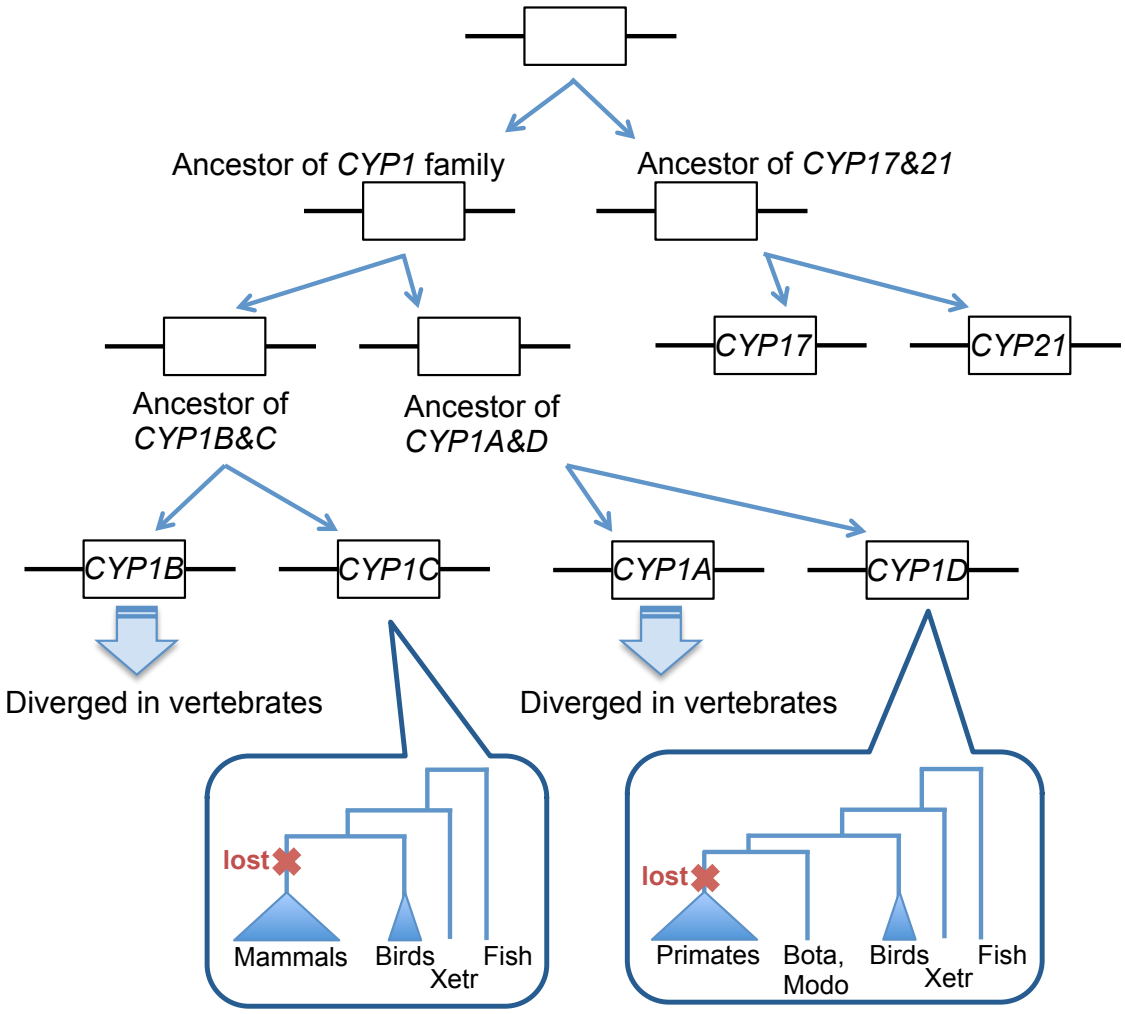

D
CYP4A(mammal specific)
CYP4X/Z
(mammal specific)
CYP4B
CYP4F
CYP4V
0.05
iv
v
Cafa CYP4A37
Cafa CYP4A38
Cafa ENSCAFT00000035951
Cafa CYP4A11
Hosa CYP4A11 CDS
Hosa CYP4A22 CDS
Bota CYP4A11
Bota ENSBTAT00000021785
Bota CYP4A22
Rano CYP4A22(2)
Mumu CYP4A2
Mumu CYP4a12a
Mumu CYP4a12b
Rano CYP4A12
Rano CYP4A10
Mumu CYP4a31
Mumu CYP4a10
Mumu CYP4a32
Modo ENSMODT00000002351
Modo ENSMODT00000001891
Modo ENSMODT00000040135
Hosa CYP4Z1 CDS
Hosa CYP4X1
Cafa CYP4X1(2)
Bota ENSBTAT00000020783
Hosa CYP4X1 CDS
Mumu CYP4x1
Rano CYP4X1
Mumu CYP4b1
Mumu CYP4b17
Rano CYP4B1
Cafa CYP4B1
Hosa CYP4B1 CDS
Bota CYP4B1
Modo CYP4B1
Orla CYP4B1
Orla CYP4B1(2)
Dare CYP4T8
Xetr CYP4A11
Xetr CYP4B1
Xetr ENSXETT00000035687
Xetr CYP4B1 2
Anca ENSACAT00000006604
Anca ENSACAT00000006607
Anca ENSACAT00000006649
Anca ENSACAT00000007048
Gaga CYP4B1
Tagu ENSTGUT00000008877
Gaga ENSTGUT00000008867
Anca ENSACAT00000017050
Anca ENSACAT00000025981
Anca ENSACAT00000006597
Tagu ENSTGUT00000008867
Orla CYP4F2
Dare CYP4F3
Xetr CYP4F22
Xetr CYP4F22(2)
Bota CYP4F22
Cafa CYP4F22
Hosa CYP4F22 CDS
Mumu CYP4f39
Rano CYP4F39
Modo CYP4F22
Anca ENSACAT00000016588
Tagu ENSTGUT00000016261
Anca ENSACAT00000018676
Modo ENSMODT00000018684
Modo ENSMODT00000018732
Modo ENSMODT00000018780
Modo ENSMODT00000018675
Mumu ENSMUST00000077639
Mumu CYP4f16
Rano CYP4F5
Mumu CYP4f17
Rano CYP4f13
Bota CYP4F6
Bota CYP4F3
Mumu CYP4f18
Rano ENSRNOT00000025271
Cafa ENSCAFT00000021609
Mumu CYP4F4
Rano CYP4f15
Mumu CYP4f40
Rano CYP4F40
Mumu CYP4f14
Rano CYP4F1
Bota CYP4F2(2)
Bota ENSBTAT00000036364
Bota ENSBTAT00000036382
Bota CYP4F2
Hosa CYP4F2 CDS
Hosa CYP4F3 CDS
Hosa CYP4F8 CDS
Hosa CYP4F11 CDS
Hosa CYP4F12 CDS
Hosa CYP4A6A1
Dare CYP4V8
Orla CYP4V2
Dare CYP4V7
Anca CYP4V2
Gaga CYP4V2
Xetr CYP4V2(2)
Xetr CYP4V2
Mumu CYP4v3
Rano CYP4V2
Bota CYP4V2
Cafa CYP4V2
Hosa CYP4V2 CDS
Hosa CYP39A1

Figure 3-4-5. The divergence time of D-type gene cluster (A) and the evolutionary scheme of D-type *CYP* gene (B, C, D and E).

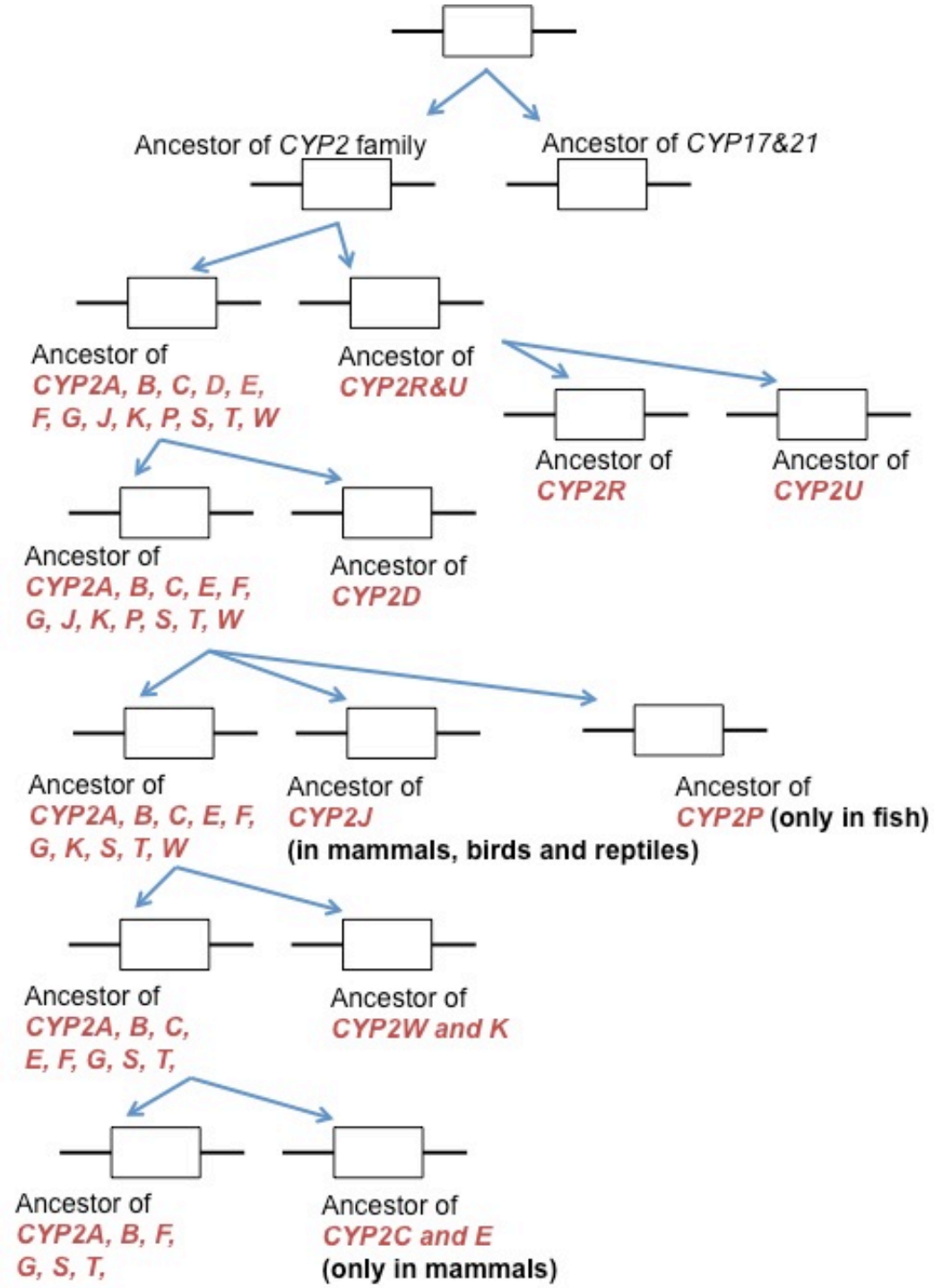
A) The divergence time of D-type gene cluster



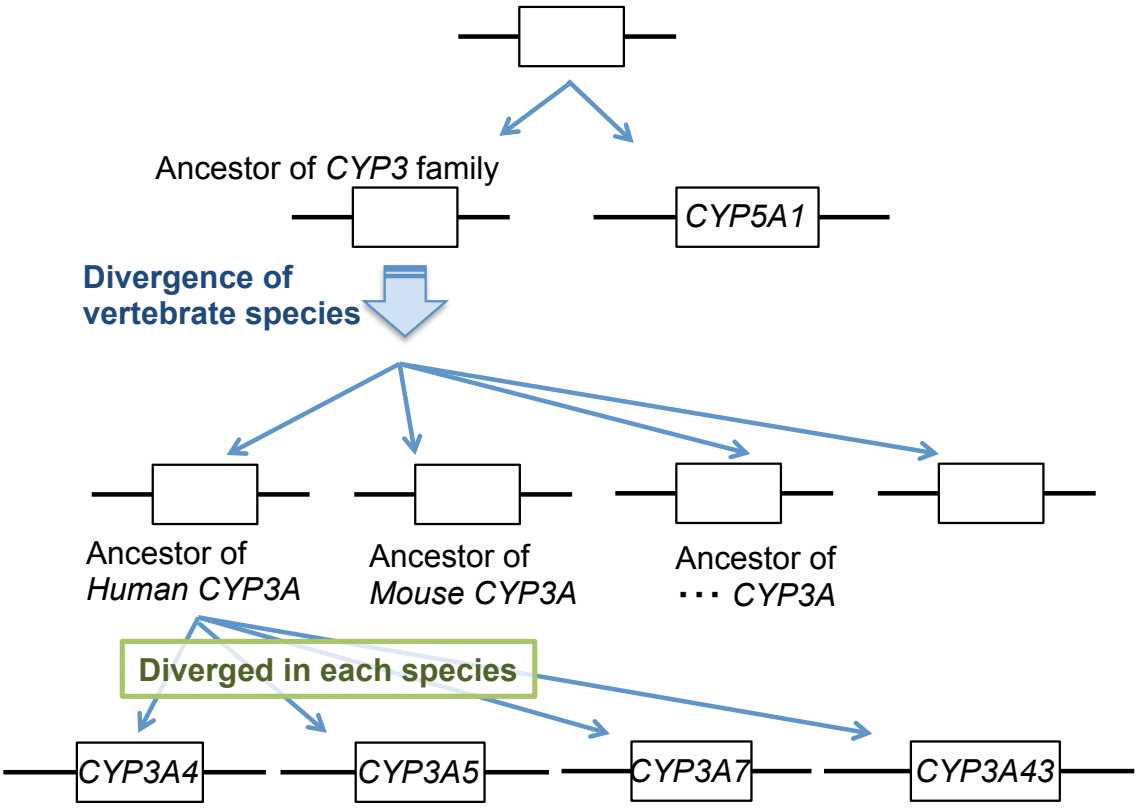
B) The evolutionary scheme of *CYP1* family genes



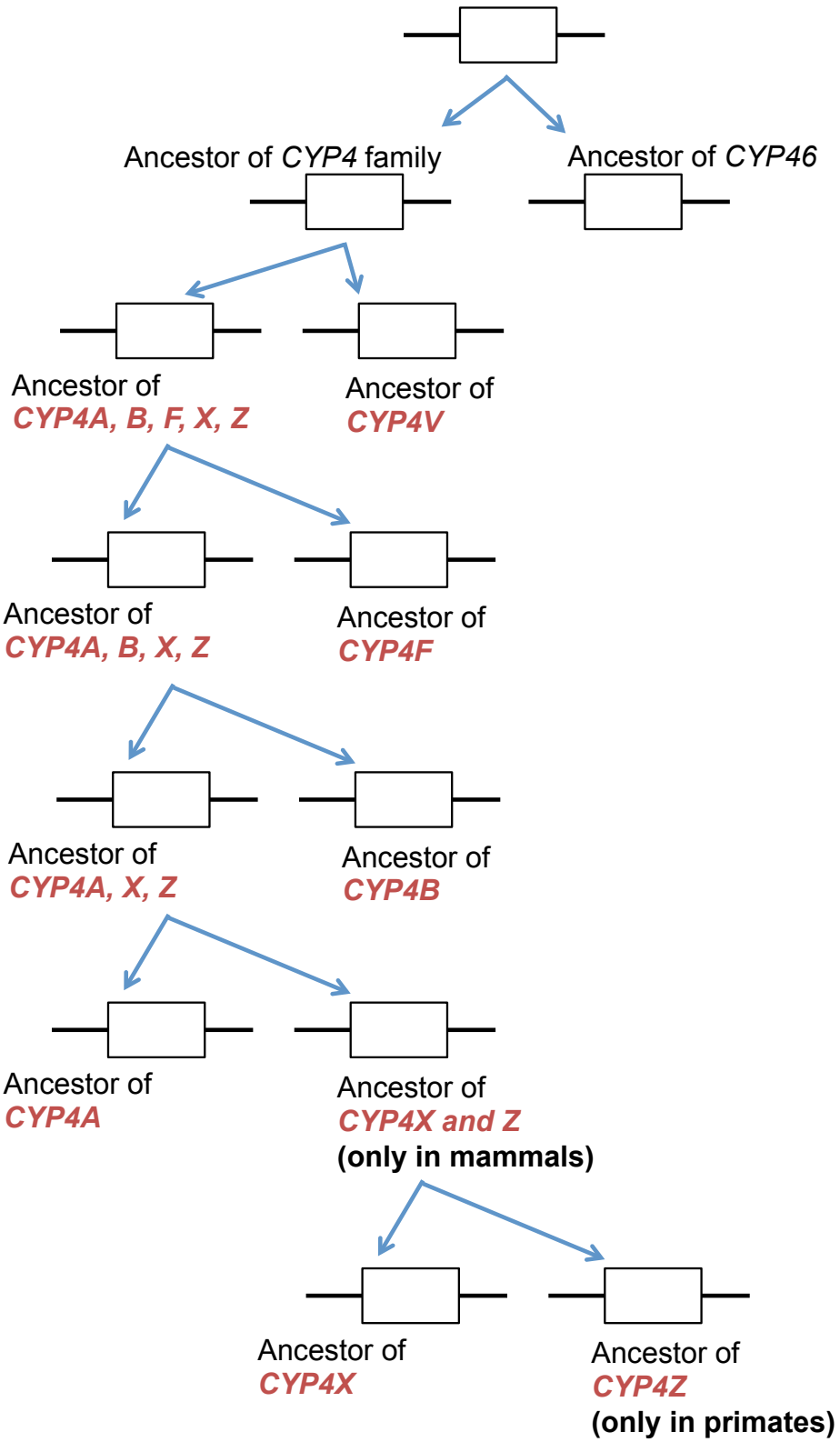
C) The evolutionary scheme of *CYP2* family genes



D) The evolutionary scheme of *CYP3* family genes



E) The evolutionary scheme of *CYP4* family genes



3.5 *CYP* gene clusters in human genome

It has been reported that there are five *CYP* clusters in the human genome (Wang H *et al*., 2008). They are *CYP2* family gene (*CYP2* cluster), *CYP2C* subfamily gene (*CYP2C* cluster), *CYP3A* subfamily gene (*CYP3A* cluster), *CYP4* family gene (*CYP4* cluster) and *CYP4F* subfamily gene (*CYP4F* cluster). The location of each cluster was shown in Figure 3-5-1.

To reveal the genome structure of each cluster, I analyzed them by genome matcher. The inverted repeat-like structures were found in *CYP2* (Figure 3-5-2A) and *CYP4* cluster (data is not shown). I examined whether similar structure was found in other five vertebrate species (chimpanzee, macaque, marmoset, cow and dog) (Figure 3-5-2B, C, D). The result shows chimpanzees and macaques have this structure, while marmoset, cow and dog do not have. In addition, this inverted repeat-like structure is composed of similar two units of which direction is opposite to each other. Human, chimpanzee and macaque have this structure. On the other hand, the length of *CYP2* cluster in marmoset is about half (250 Kb) of human, chimpanzee or macaque (500 Kb) and contains the single unit. In addition, dog and cow also have only one unit. Therefore, the unit was likely constructed before the divergence of mammal by *CYP2* gene duplication, then the duplication of unit with inverted direction had occurred on the genome of catarrhine ancestor.

Several short repeats were detected by the genome matcher analysis. To know the nature of repeats, I tried to examine whether they were retrotransposons (LINE: long interspersed nuclear element and SINE: short interspersed nuclear element) by using the

program of RepeatMasker. Maps of retrotransposons identified in clusters were shown in Figure 3-5-3A-E and their numbers as well as their frequencies per 1Kb were shown in Figure 3-5-3F. In the human genome, there are about 1,090,000 *Alu*s and their average frequency is 0.3 *Alu*s/kb. The range of frequencies in the *CYP* clusters is 0.23/kb for *CYP4* cluster to 1.22/kb for *CYP2* cluster. Compared with the genome average, the frequency of *Alu*s in a *CYP* cluster is higher. Moreover, the frequencies vary largely even within a cluster: in *CYP3A* cluster, the region for *CYP3A5P1* contains no retrotransposons whereas the region for *CYP3A43* contains 22. The density of retrotransposons in a *CYP* gene is not equal among genes in a cluster. Insertion of *Alu*s was highly dependent on the GC contents of the region, and the cause of unequal distribution of retrotransposons could be due to the difference of GC contents between *CYP* genes. However, there was no correlation between GC contents and *Alu* numbers at least in the *CYP* cluster. The GC contents of each gene, intron and exon were shown in Table 3-5-1. GC contents in an exon were generally higher than that in an intron.

Figure 3-5-1. The position of 5 *CYP* gene clusters in the human genome.

Five *CYP* gene clusters were shown by blue rectangles. The number at the top of each chromosome represents chromosome number. This figure was modified from CYP Home Page.

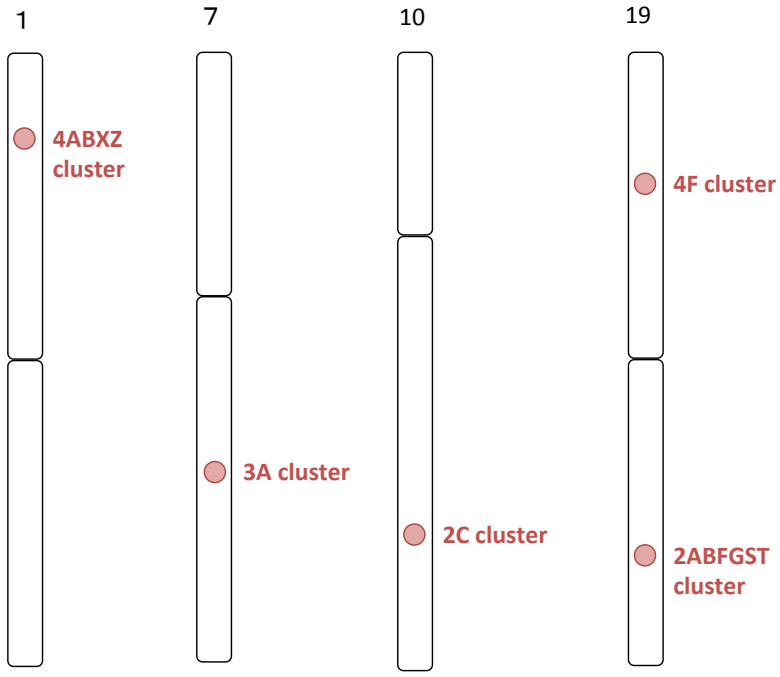
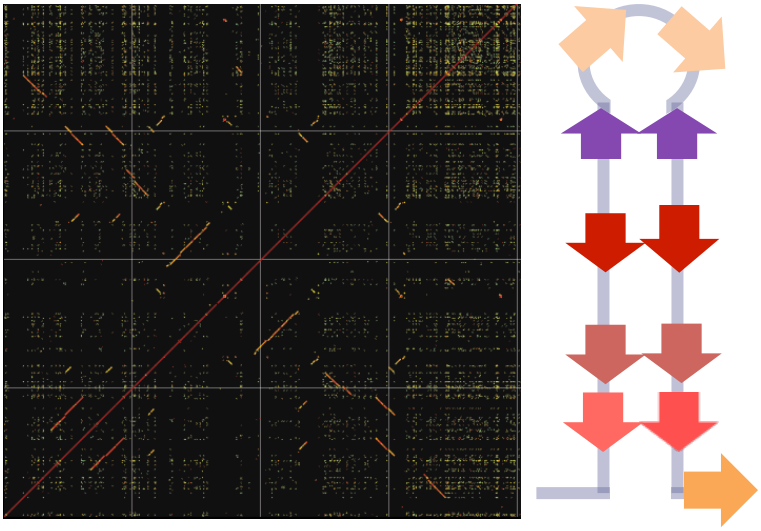


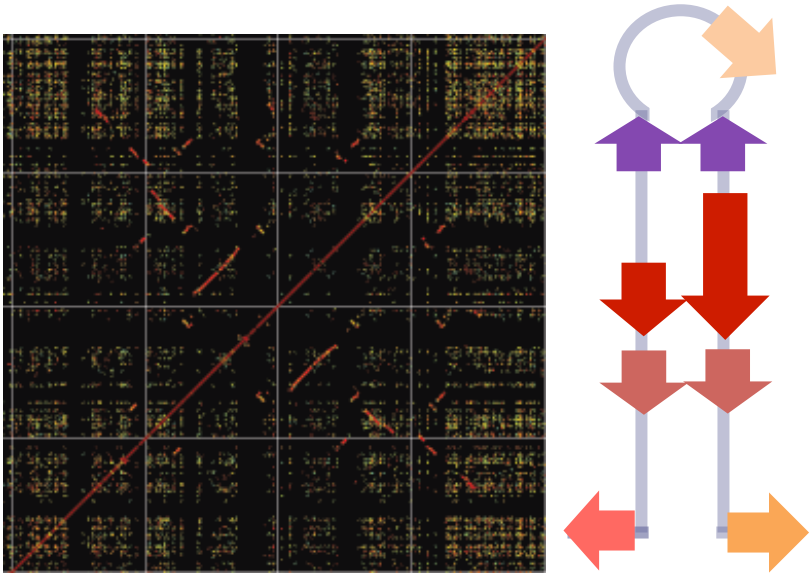
Figure 3-5-2. The genome structure of *CYP2* family cluster

The genome structure in the region of *CYP2* family was illustrated by Genome matcher in human (A), chimpanzee (B), macaque (C) and marmoset (D). The diagonal red line was made by high homology of comparison for own sequence. Each short unit line shows tandem repeat (parallel to the center diagonal line) and inverted repeat (vertical to the center diagonal line), respectively. The right illustration represents the image of genome structure.

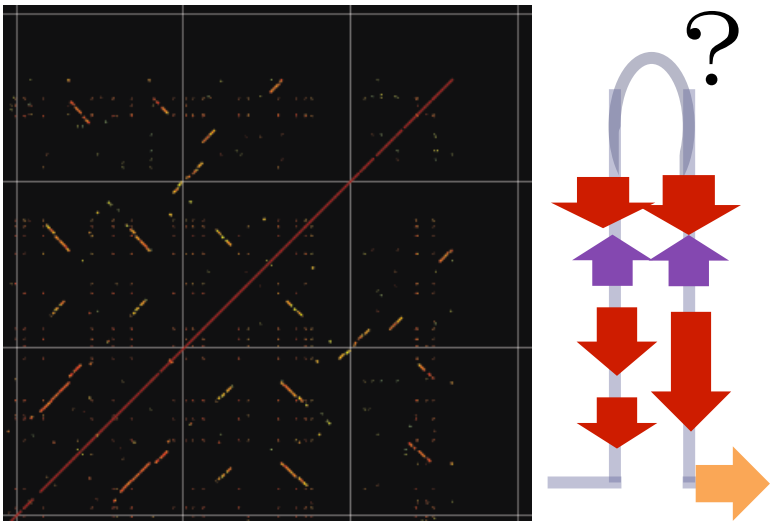
A) Human genome (chr.19_41.3-41.7)



B) Chimpanzee genome (chr.19_46.4-46.8)



C) Macaque genome (chr.19 47.1-47.5)



D) Marmoset genome (chr22_34.1-34.4)

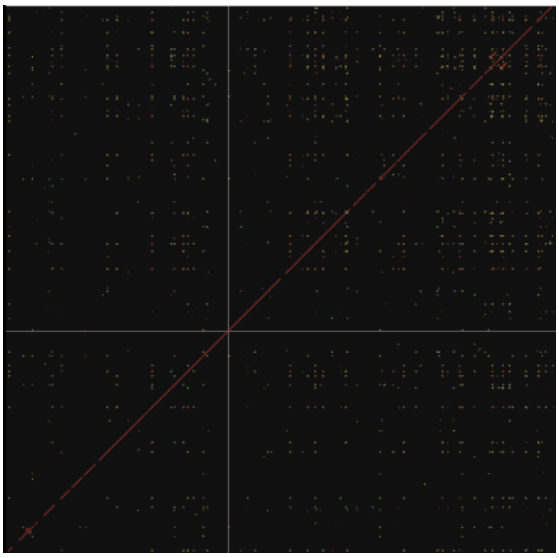
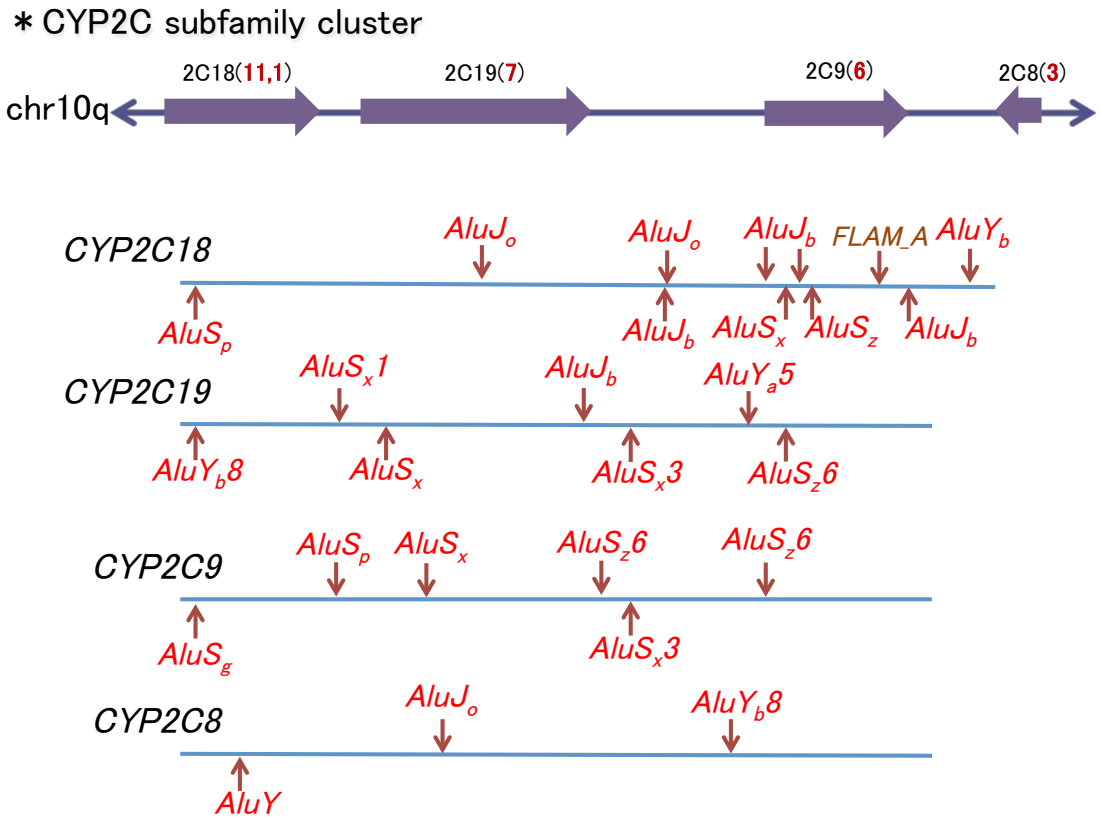


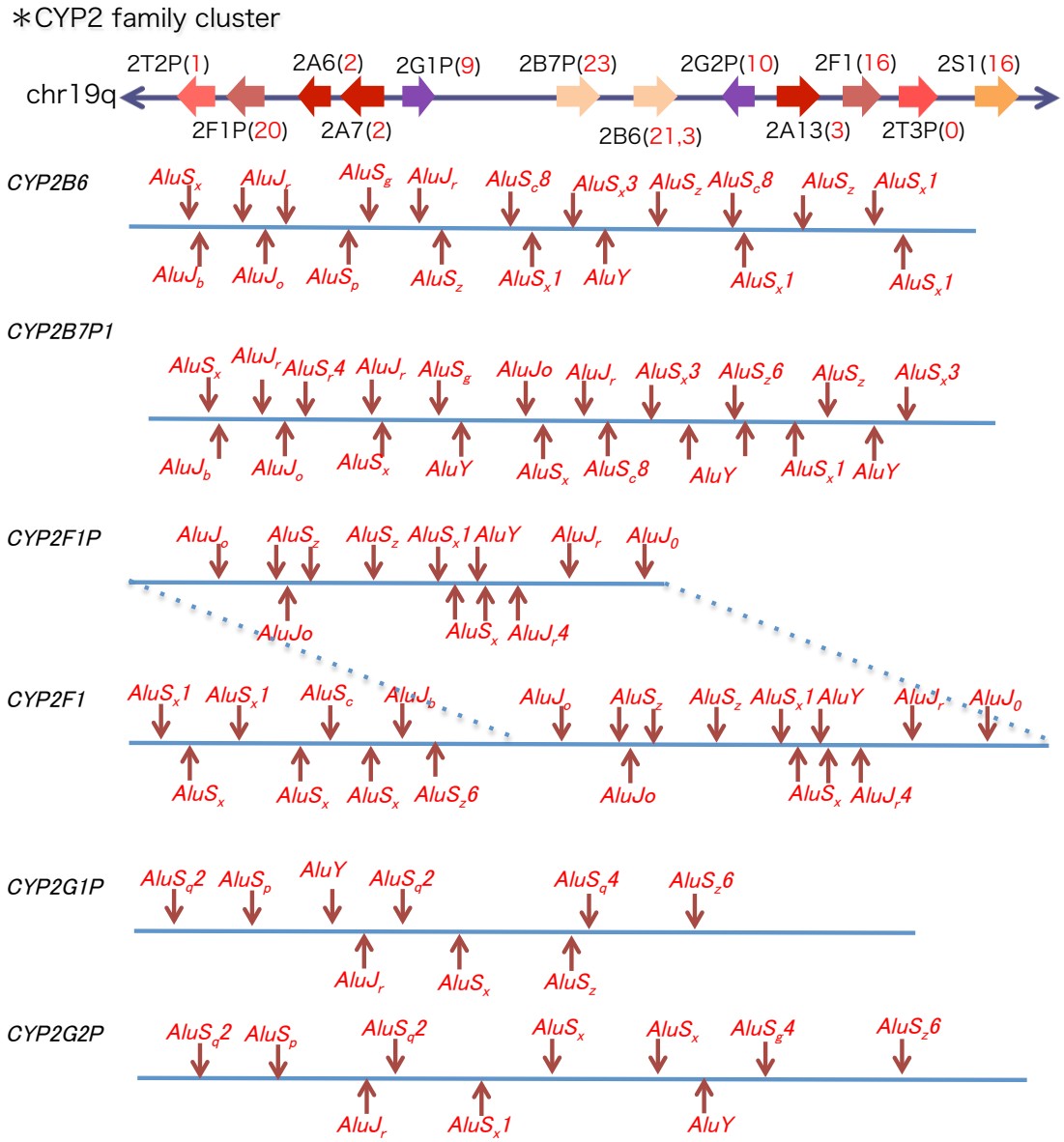
Figure 3-5-3. The map of retrotransposons in CYP clusters in the human genome.

Positions of *Alu* repeats within a *CYP* gene A) *Alu* position in each gene on *CYP2C* cluster, B) on *CYP2* cluster, C) on *CYP3A* cluster, D) on *CYP4* cluster, and E) on *CYP4F* cluster F) The numbers and frequencies of *Alu*s in a cluster. *CYP4F22* have 57 *Alu*s, and their order and kind of repeat is listed below.

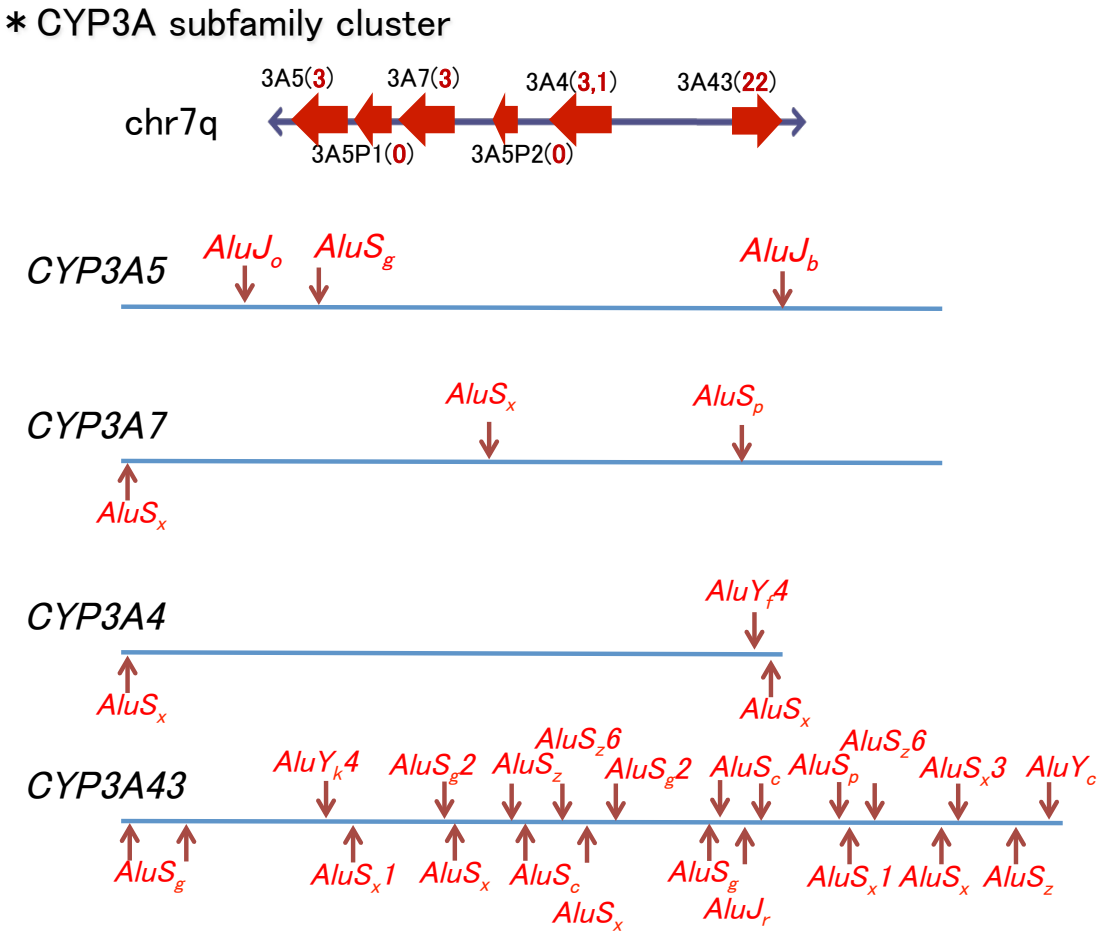
A)



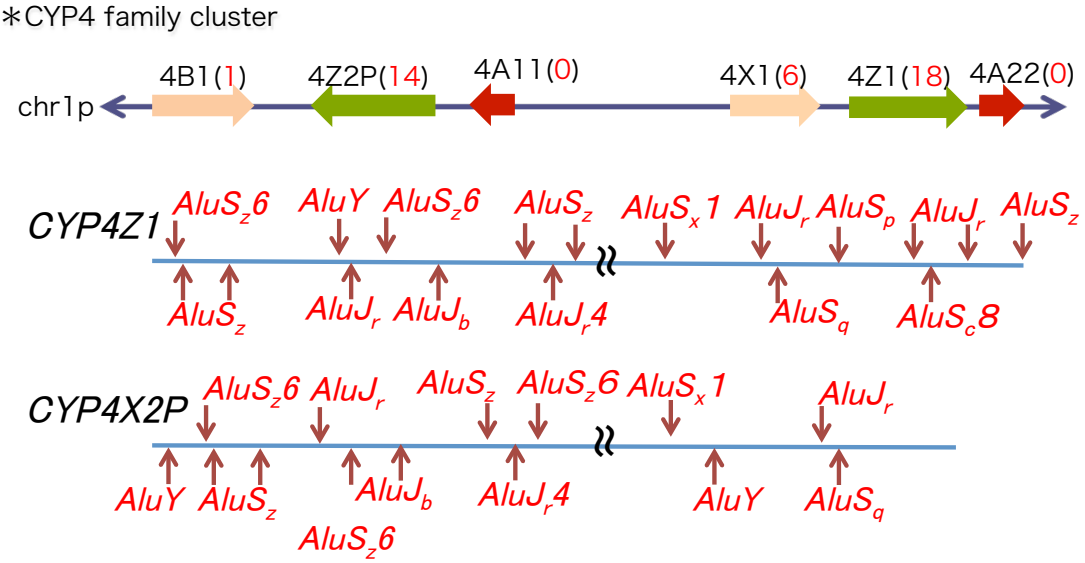
B)


＊CYP2 family cluster
chr19q
2T2P(1)
2A6(2)
2G1P(9)
2B7P(23)
2G2P(10)
2F1(16)
2S1(16)
2F1P(20)
2A7(2)
2B6(21,3)
2A13(3)
2T3P(0)
CYP2B6
CYP2B7P1
CYP2F1P
CYP2F1
CYP2G1P
CYP2G2P

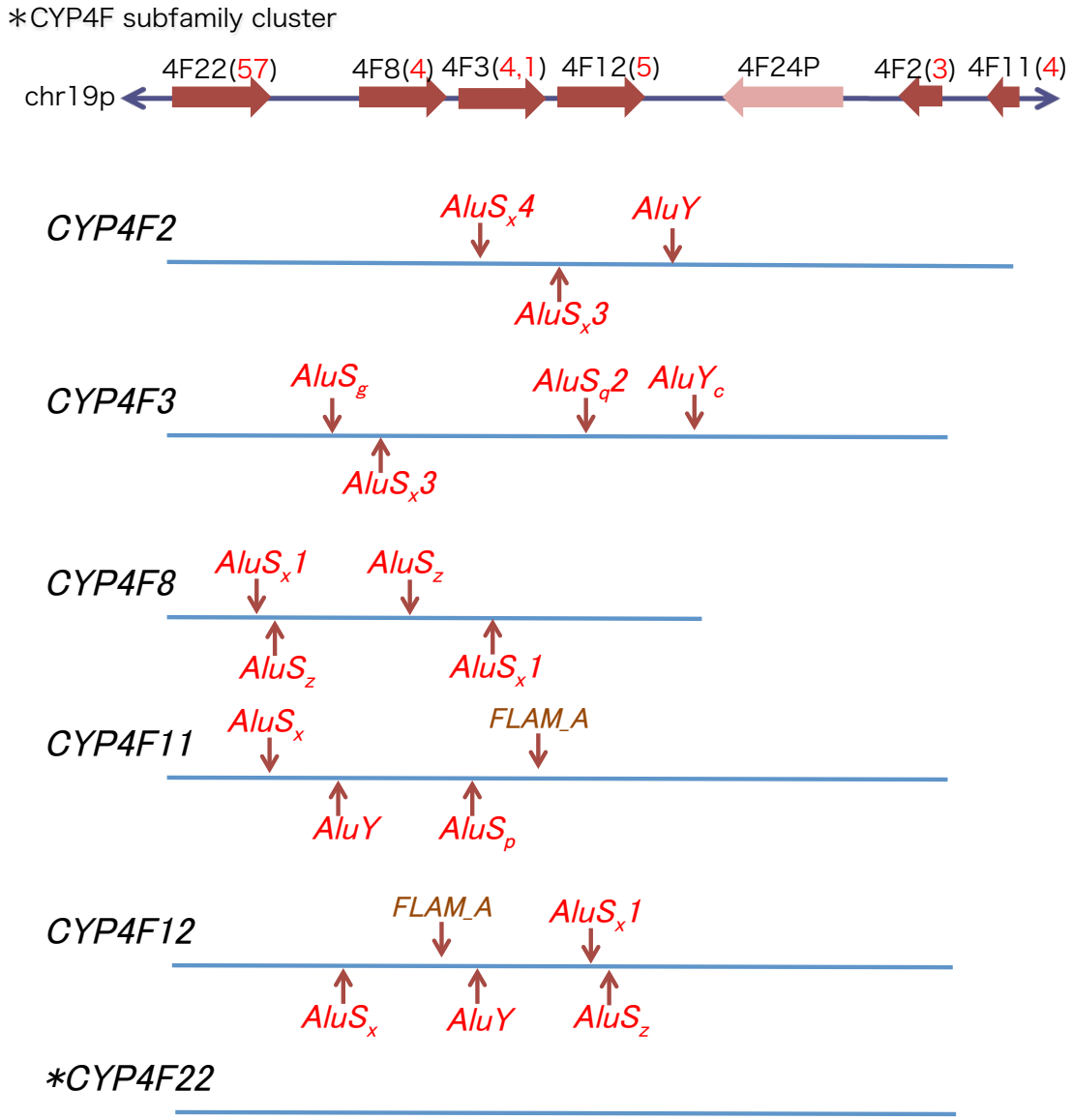
C)


* CYP3A subfamily cluster
chr7q
3A5(3)
3A7(3)
3A4(3,1)
3A43(22)
3A5P1(0)
3A5P2(0)
CYP3A5
AluJo
AluSg
AluJb
CYP3A7
AluSx
AluSp
AluSx
CYP3A4
AluYf4
AluSx
AluSx
CYP3A43
AluYk4
AluSg2
AluSz
AluSz6
AluSg2
AluSc
AluSp
AluSz6
AluSx3
AluYc
AluSg
AluSx1
AluSx
AluSc
AluSx
AluSg
AluJr
AluSx1
AluSx
AluSz

D)


＊CYP4 family cluster
chr1p
4B1(1)
4Z2P(14)
4A11(0)
4X1(6)
4Z1(18)
4A22(0)
CYP4Z1
AluSz6
AluY
AluSz6
AluSz
AluSx1
AluJr
AluSp
AluJr
AluSz
AluSz
AluJr
AluJb
AluJr4
AluSq
AluSc8
CYP4X2P
AluSz6
AluJr
AluSz
AluSz6
AluSx1
AluJr
AluY
AluSz
AluJb
AluJr4
AluY
AluSq
AluSz6

E)



1. *AluJb*, 2. *AluJb*, 3. *AluSp*, 4. *AluSx1*, 5. *FLAM_C*, 6. *AluSx*, 7. *AluY*, 8. *AluSg7*, 9. *AluSx*, 10. *AluY*, 11. *AluSx*, 12. *AluJr*, 13. *AluY*, 14. *AluSx1*, 15. *AluY*, 16. *AluSc8*, 17. *AluSc8*, 18. *AluSx1*, 19. *AluJr*, 20. *AluSg*, 21. *AluJb*, 22. *AluSz*, 23. *AluSq*, 24. *AluJb*, 25. *AluJb*, 26. *AluJr4*, 27. *AluJo*, 28. *FLAM_A*, 29. *AluSz*, 30. *AluJo*, 31. *AluJb*, 32. *AluSq2*, 33. *AluJr*, 34. *AluJr*, 35. *AluJo*, 36. *AluSz*, 37. *AluJo*, 38. *AluJo*, 39. *AluJo*, 40. *AluSx1*, 41. *AluJr*, 42. *AluJr*, 43. *AluJo*, 44. *AluSp*, 45. *AluSq*, 46. *AluJb*, 47. *AluSz6*, 48. *AluSq*,

49. *AluJb*, 50. *AluJo*, 51. *AluSz*, 52. *AluJo*, 53. *AluSz6*, 54. *AluSq*, 55. *FLAM_C*, 56. *AluJb*, 57. *AluSx*

F) The location and frequency of *Alu*s in a cluster.

The numbers after the cluster name represent the number and the frequency (per kb) of *Alu*s. The numbers in red in the parentheses represent the numbers of *Alu*s in a given gene.

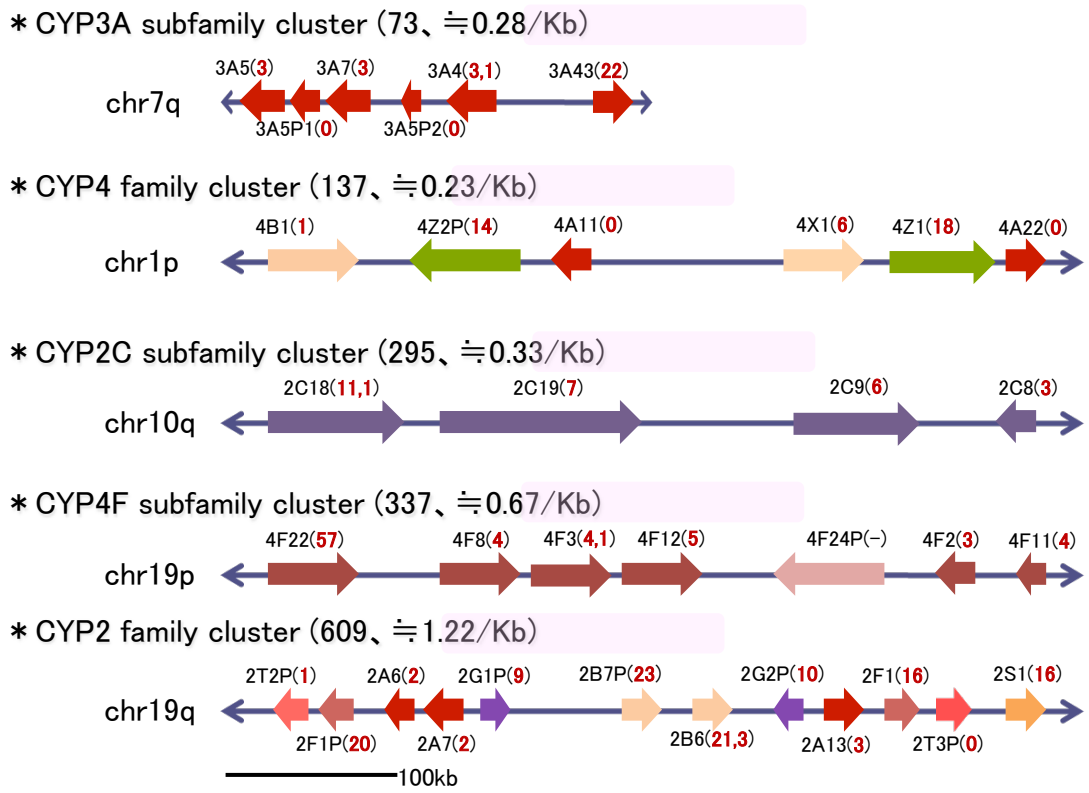


Table 3-5-1. GC contents in a gene on *CYP* clusters

The numbers of *Alu*s as well as GC contents(%) in the entire gene, exon and intron are shown.

A) *CYP2* cluster

Gene	Number of *Alu*s	Inside of gene (%)	Exon (%)	Intron (%)
CYP2T2P	1	59.1	60.8	57.9
CYP2F1P	20	52.7	59.6	51.8
CYP2A6	2	53.4	56.6	
CYP2A7	2	53.4	56.4	
CYP2G1P	9	44.8	50.6	43.6
CYP2B7P	23	48.6		
CYP2B6	21,3	50.0		
CYP2G2P	10	43.5	52.0	42.3
CYP2A13	3	53.1	57.4	
CYP2F1	16	58.5		
CYP2T3P	0	60.6		
CYP2S1	16	59.5		

B) *CYP4F* subfamily cluster

Gene	Number of *Alu*s	Inside of gene (%)	Exon (%)	Intron (%)
CYP4F22	57	49.3	57.2	
CYP4F8	4	48.2	57.1	
CYP4F3	4, 1	46.3	57.7	
CYP4F12	5	45.8	56.3	
CYP4F24P				
CYP4F2	3	46.1	57.8	
CYP4F11	4	46.0	55.9	

C) *CYP2C* subfamily cluster

Gene	Number of *Alu*s	Inside of gene (%)	Exon (%)	Intron (%)
CYP2C18	11, 1	38.7	46.3	
CYP2C19	7	39.5	45.8	
CYP2C9	6	37.8	45.15	
CYP2C8	3	37.6	44.0	

D) *CYP4* family cluster

Gene	Number of *Alu*s	Inside of gene (%)	Exon (%)	Intron (%)
CYP4B1	1		56.4	
CYP4Z2P	14			
CYP4A11	0		55.6	
CYP4X1	6		49.5	
CYP4Z1	18		47.0	
CYP4A22	0		55.7	

E) *CYP3A* subfamily cluster

Gene	Number of *Alu*s	Inside of gene (%)	exon (%)	intron (%)
CYP3A5	3	40.5	44.0	
CYP3A5P1	0	42.6		
CYP3A7	3	41.4	43.3	
CYP3A5P1	0	43.8		
CYP3A4	3, 1	39.6	43.7	
CYP3A43	22	40.6	41.7	

3.6 B- and D-type *CYP* pseudogenes

The evolutionary modes of D- and B-type *CYP* genes differed also in pseudogenization that was defined as a loss of gene function. Among the 58 pseudogenes present in the human genome, more than half (41 of 58: Figure 3-6-1) are fragmented, with few exons and introns remaining. The total length of pseudogenes was less than one-tenth of that of functional *CYP* genes, which prevented identification of functional paralogs (Figure 3-6-1). Functional paralogs were identified for 17 pseudogenes, among which 3 were B-type (*CYP21A1P*, *CYP51P1*, and *CYP51P2*) and 14 were D-type (*CYP1D1P*, *CYP2A7P1*, *CYP2B7P1*, *CYP2D7P1*, *CYP2D8P1*, *CYP2F1P*, *CYP2G1P*, *CYP2G2P*, *CYP2T2P*, *CYP2T3P*, *CYP4F9P*, *CYP4F23P*, *CYP4F24P*, and *CYP4Z2P*) (Table 3-1-1). Of the 3 B-type pseudogenes, *CYP51P1* and *CYP51P2* are processed pseudogenes, and the biological causes of their pseudogenization are not related to a relaxation of functional constraints. In this sense, *CYP21A1P* is only a pseudogene due to relaxation of functional constrains. Rhesus macaques, orangutans, and humans have two copies of *CYP21A*, and chimpanzees have three (Figure 3-6-2). However, a pseudogene for *CYP21A* is present only in humans, and the time of pseudogenization is estimated to be 6.7 mya, around the divergence of humans from chimpanzees. The presence of this pseudogene is clinically significant in humans: partial gene conversion from a pseudogene to a functional gene causes 21-hydroxylase deficiency; furthermore, copy number variation has been observed in the region containing *CYP21A* and the neighboring *C4A* in the *HLA* region of human chromosome 6 (Urabe K *et al*., 1990).

In contrast, among the 14 D-type pseudogenes, four (*CYP2G1P*, *2G2P*, *2T2P*, and *2T3P*) have been reported to be human-specific, on the basis of a comparison between humans and mice (Nelson DR *et al*., 2004). I searched for orthologs to the human pseudogenes in other primate genomes and found that all but *CYP2G2P* are pseudogenized in other primates as well, but are functional in non-primate vertebrates (Figure 3-6-3). The findings showed that *CYP2G1P*, *2T2P*, and *2T3P* are primate-specific pseudogenes, whereas *CYP2G2P* is a human-specific pseudogene. Using an accelerated non-synonymous substitution rate in pseudogenes (Sawai H *et al*., 2008), I calculated that *CYP2G2P* emerged 2.6 mya. In addition to *CYP2G2P*, further analysis revealed a single human-specific pseudogene, *4Z2P,* with a pseudogenization time of 6.4 mya. On the basis of the results of this analysis, *CYP2D7P1* also appeared to be a human-specific pseudogene. Interestingly, however, pseudogenization of this ortholog has also been found in orangutans, but the cause is different from that for humans (Yasukochi Y *et a*l., 2011). It appeared that this gene lost its function in humans and orangutans independently.

Nine human specific pseudogenes were previously identified in the human genome (Kim HL *et al*., 2009). The DNA sequence of *ZNF850* (Wang *et al*., 2006) and *SIGLEC13* (Angata *et al*., 2004) was, however, not found in NCBI. Therefore I used seven of them, namely *CMAH* (Hayakawa *et al*., 2006), *GLRA4* (IHGSC 2001), *KRT41* (Winter *et al*., 2001), *MBL1P1* (Wang *et al*., 2006), *myh16* (Stedman *et al*., 2004), *S100A15* (Hahn *et al*., 2007) and *TDH* (Edgar, 2002) to compare the time of loss of function and causes for thse pseudogenization with those of human specific *CYP*

pseudogenes (Figure 3-6-4). The result shows that *MLB1P1* and *KRT41* seem to have lost their function just after the divergence between human and chimpanzee. After that, *Myh16*, *CMAH*, *TDH*, *S100A15* and *GLRA4* became a pseudogene around 4 mya. Finally, *CYP2G2P* also have lost its function 2.6 mya.

In D-type genes, in addition to the 14 pseudogenes present in the human genome, 7 pseudogenes were identified in chimpanzees, macaques, marmosets, dogs, and cows. Among the seven, six were species-specific, one (*2C18*) to chimpanzees, two (*2A13* and *4F11*) to macaques, and three (*4B1* and two *4F22*-like genes) to marmosets. The remaining one, *CYP2B6P,* was pseudogenized independently in chimpanzees and macaques. Among the 11 pseudogenes, with the exception of the three human-specific pseudogenes, *CYP2A7P1* was pseudogenized in macaques and humans independently, at 28.4 mya and 5.9 mya, respectively. Pseudogenization of the remaining 10 genes occurred in the primate or hominoid stem lineage. Furthermore, other pseudogenization times of other vertebrate *CYP*s are listed in Table 3-6-2.

It is unclear how many times pseudogenization occurred in D-type genes before eutherian divergence. I estimated the rate after the eutherian radiation. It is 30 pseudogenizations over 432 myr, yielding a rate of 6.9 per 100 myr. In contrast, the number of pseudogenization events in B-type genes was estimated to be only five over 2,685 myr, yielding a rate of 0.19 per 100 myr. Regarding pseudogenization, D-type genes show more unstable or rapid turn over than B-type genes.

Table 3-6-1. The list of human 58 *CYP* pseudogenes

Human 58 *CYP* pseudogene and locus is listed in the table. Same locus is colored with the same color.

Gene name	locus
1A8P/1D1P	9q21.12
2A18PC	19q13.2
2A18PN	19q13.2
2B7P1	19q13.2
2C8-de6b	10q23.33
2C9-de1b	10q23.33
2C9-de2c3c	10q23.33
2C58P	10q23.33
2C62P	10q24.31
2C-se1[7]	2q24.3
2C-se2[1:2]	10q21.3
2C-se3[1]	21q21.2
2C-se4[1]	Xq28
2D7P	22q13.2
2D8P	22q13.2
2F1P	19q13.2
2G1P	19q13.2
2G2P	19q13.2
2T2P	19q13.2
2T3P	19q13.2
2AB1P	3q27.1
2AC1P	6p12.3
3A4-ie1b	7q22.1
3A5-de13c	7q22.1
3A5-de1b2b	7q22.1
3A5_v2	7q22.1
3A5_v3	7q22.1
3A7_del1b2b	7q22.1
3A43-de1b	7q22.1
3A43-de4c6c	7q22.1
3A-se1[2]	7q22.2
3A-se2[5]	7q22.2
4A-se1[12]	1p33
4A-se2[1]	1p33
4A-se3[12]	1p33
4A-se4[2]	1p33
4F2-de12b	19p13.12
4F9P	19p13.12
4F10P	19p13.12
4F23P	19p13.12
4F24P	19p13.12
4F-se5[5:8]	2q11.1
4F-se9[6:7:8]	2q21.1
4F-se10[6:7:8]	2q21.1
4F-se11[6:7:8]	2q21.1
4F-se3[6:7:8]	2q21.1
4F-se12[6:8]	8p11.1
4F-se13[6:8]	9p11.2
4F-se1[6:8]	9q13
4F-se2[6]	9q13.3
4F-se6[6]	9q13.1
4F-se7[6:7:8]	13q11
4F-se8[6:7:8]	18p11.21
4F-se4[6:7:8]	21q11.2
4Z2P	1p33
21A1P	6p21.3
46A-se1[12:13:14]	1p33
51P1	3q12.2
51P2	13q12.3
51P3	6q24.3

Figure 3-6-1. Categorization of the 58 human *CYP* pseudogenes

Among the 58 pseudogenes, paralogs were detected by a BLAST search.

a. The number of exons and introns is the same as in the paralogous genes (13 genes), b. They contain greater than half the number of exons and introns of their paralogs (6 genes), c. One or two exons or introns remained (18 genes), d. A portion of an exon remained (13 genes), e. The BLAST search returned no hits (6 genes). *Two pseudogenes were absent from the human genome databases. Approximately one-third (a and b) of the human *CYP* pseudogenes were used for phylogenetic analysis.

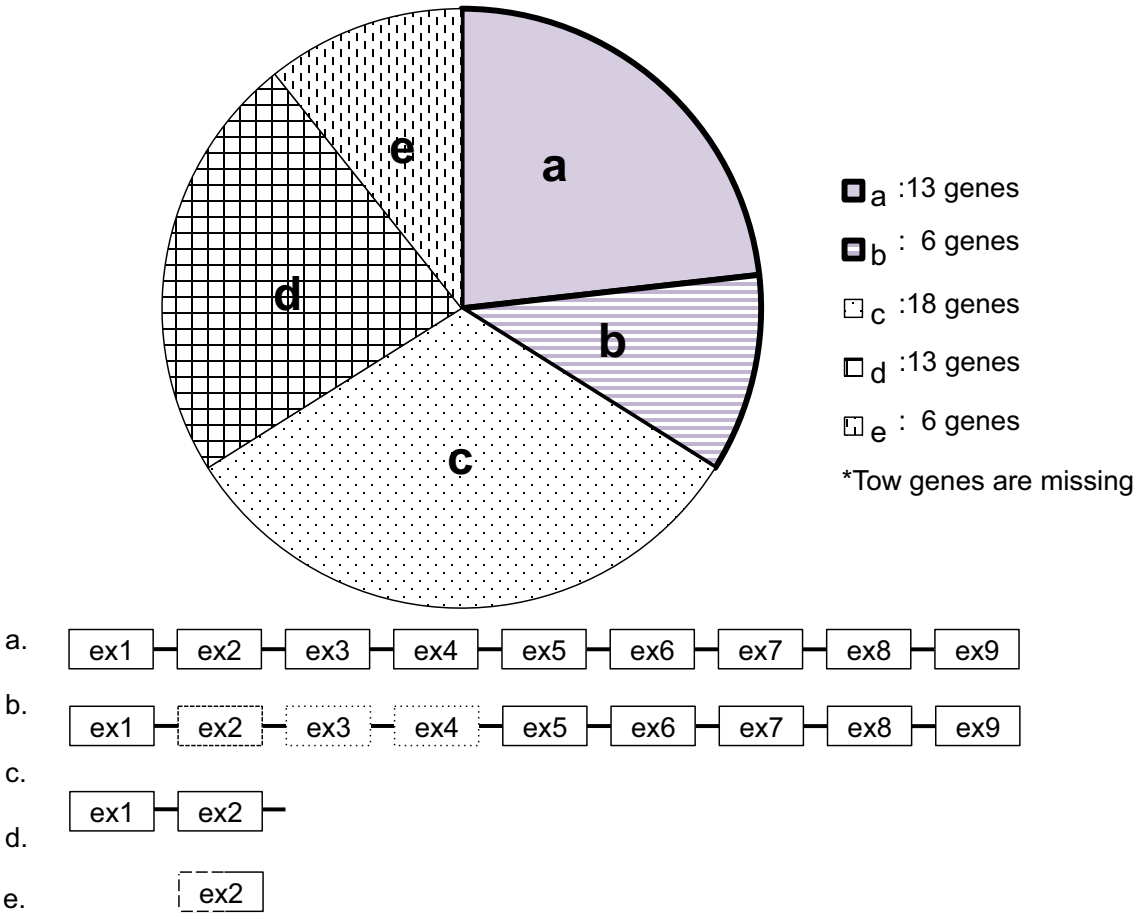


Figure 3-6-2. Time of pseudogenization of *CYP21A1P* in humans.

The phylogenetic tree was obtained using the NJ method, using the CDS. The pseudogene is represented by Ψ. The cross shows the time at which function was lost.

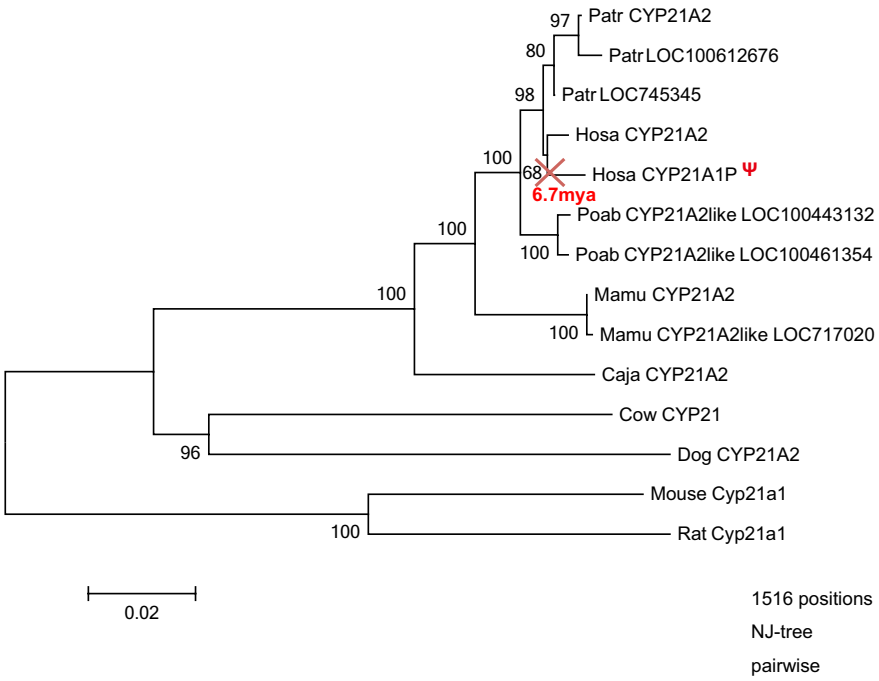


Figure 3-6-2. The cause of functional loss in human-specific *CYP* pseudogenes.

There are four human-specific *CYP* psuedogenes (*CYP2G1P*, *2P*, *CYP2T2P*, and *3P*).

Possible causal mutations, premature stop codons (red) and frame-shift mutations (blue), were identified in human and other primate CYP nucleotide and amino acid alignments. The row labeled "exon" for *CYP2G1P* and *2P* shows the number of exons in which mutations were found, and the row labeled "bp" indicates the nucleotide position of the CDS in functional genes from rat and mouse.

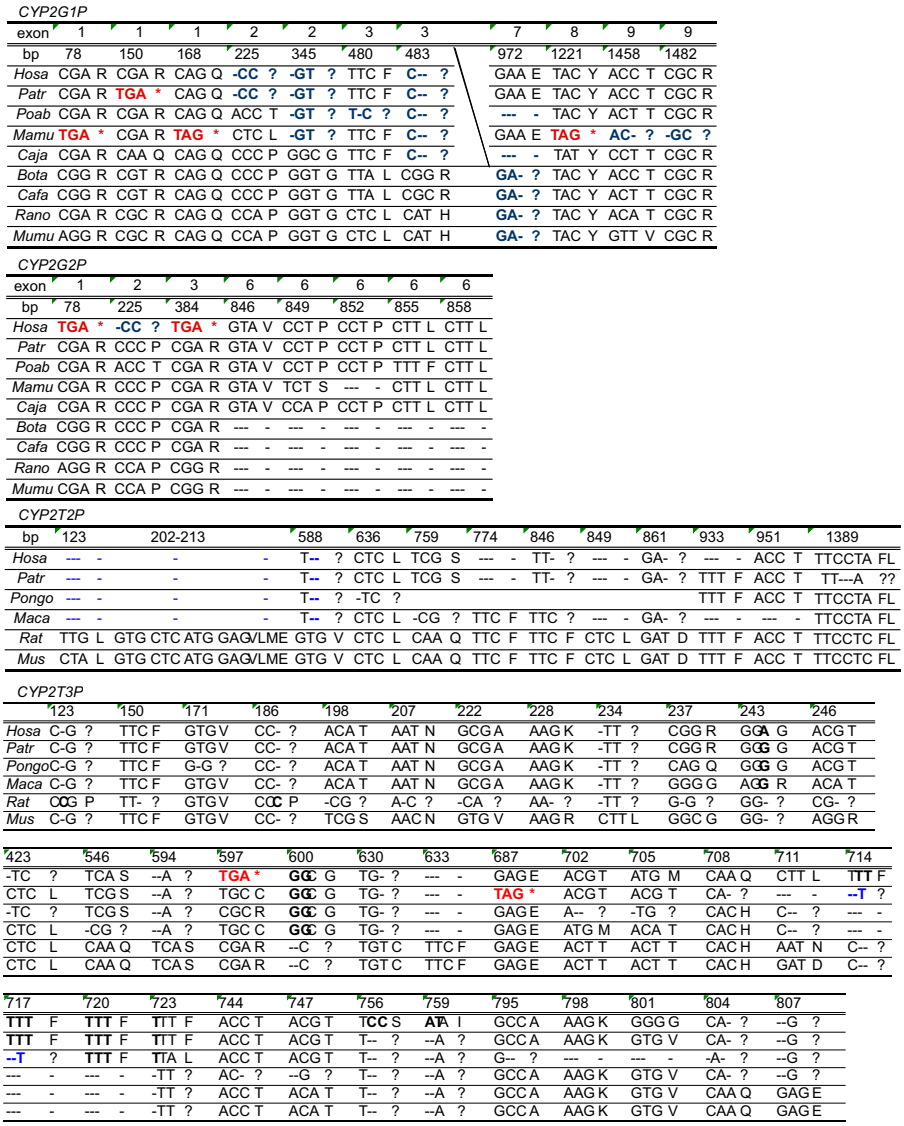


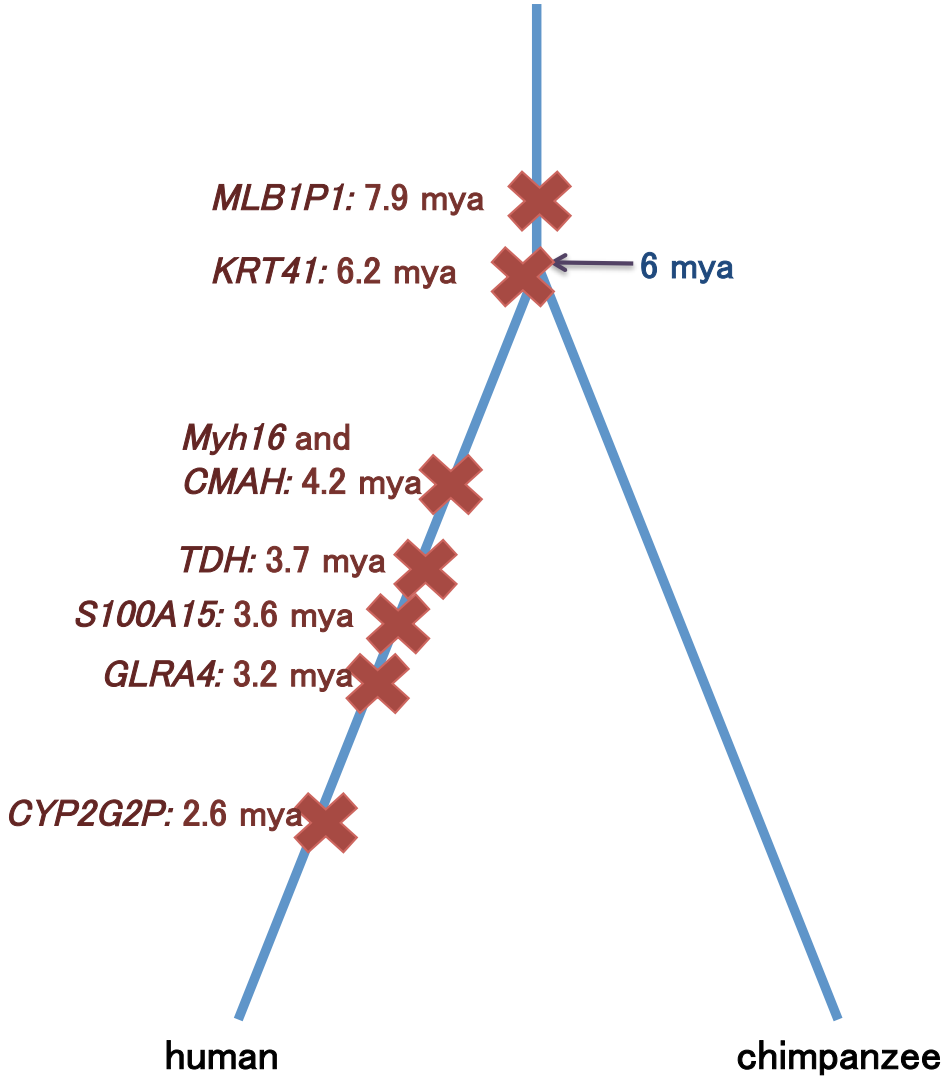
Table 3-6-3. Pseudogenization time in other vertebrate *CYP* pseudogene

Gene	Species	Time of pseudogenization (mya)
CYP2B6like	Patr	0.2
CYP4A22	Caja_1	8.0
	Caja_2	9.2

CYP4B1	**Poab**	**14.2**
	Caja	**19.5**
CYP4F9P	**Patr**	**10.2**
	Cafa	**8.4**
CYP4F23P	**Cafa**	**5.1**

Figure 3-6-4. The pseudogenization time of human specific pseudogenes

The cross mark represents the point of loss of function in each gene.

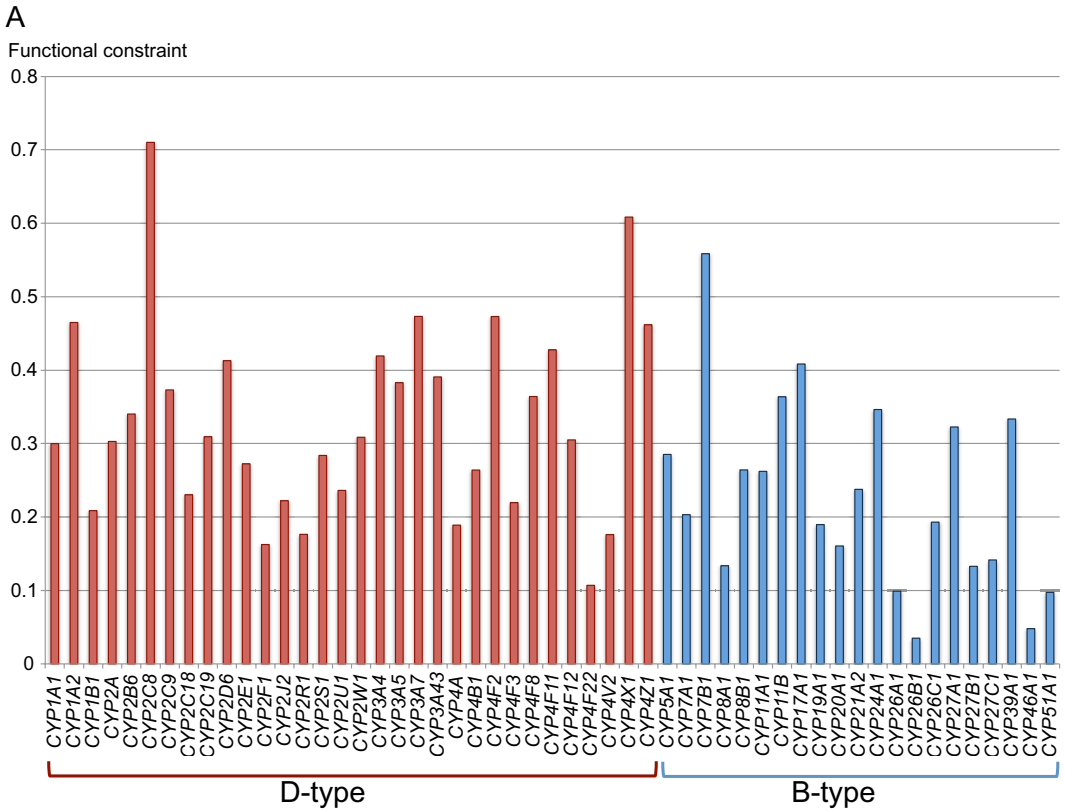


3.7 Evolutionary rate of B- and D-type genes

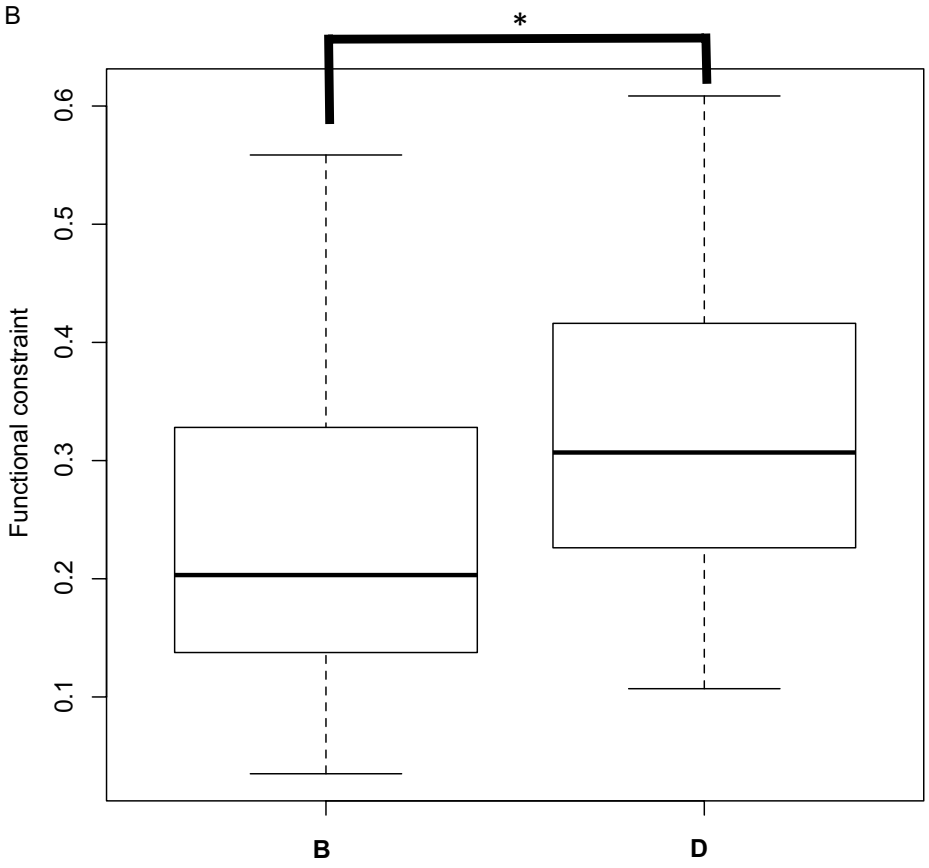
The results revealed that the births and deaths of genes was more frequent in D-type genes than in B-type genes. As such, it was important to compare the evolutionary rate of B- and D-type genes. For this comparison, the non-synonymous substitution rate was normalized to the synonymous rate, and the ratio (*f*) for each B- and D-type gene was calculated in primates (see Materials and Method, Figure 3-7-1A). The average *f*-values for B- and D-type genes were calculated to be 0.24 ± 0.14 and 0.33 ± 0.13, respectively, and the median values for B- and D-type genes were determined to be 0.23 and 0.31, respectively (Figure 3-7-1B). The average and median values for D-type genes were significantly greater than those for B-type genes (Wilcoxon's test, *P*-value = 0.0173), suggesting that the degree of functional constraint (1-*f*) is stronger in B-type than in D-type genes. These results are consistent with the rapid birth and death process of D-type genes.

Figure 3-7-1. Functional constraint of *CYP* genes.

A) Fraction of (1-*f*) nonsynonymous to synonymous substitution rate was estimated for each *CYP* gene. The *y*-axis shows the value of *f* obtained via the ratio of per-site non-synonymous substitutions to synonymous substitutions (D_N/D_S). Red bars indicate D-type genes, and blue bars indicate B-type genes.



B) Comparison of median values for the fraction (f) of nonsynonymous to synonymous substitution rate between primate B- and D-type genes. “B” stands for B-type genes, and “D” for D-type genes. The *P*-value was 0.01282 (significance was defined as $P < 0.05$). The *P*-value was calculated using the Mann–Whitney *U* test.



Chapter 4 General discussion and Perspectives

4.1 Evolutionary mode of *CYP* genes in vertebrates

4.1.1 The origin of D-type *CYP* genes

The origin of B-type genes is assumed to be single and ancient, because fission yeast possesses B-type genes and because a possible ortholog to the B-type gene *CYP51* is present even in prokaryotic genomes. However, D-type genes have different origins. The present phylogenetic analyses demonstrate that four D-type families are conserved among all vertebrates, and that the D-type families have been derived from three gene-duplication events of B-type genes in the stem lineage of vertebrates. Based on the molecular clock hypothesis, B- to D-type gene duplications are estimated to occurre before 600–700 mya, consistent with the phylogenetic analysis. D-type *CYPs* impart resistance to insecticides in invertebrates; in fruit flies, two such enzymes are CYP6U1 and CYP6D2. The phylogenetic analysis of both human and fruit fly *CYP* genes are shown to have an independent emergence from vertebrate D-type genes. Moreover, other invertebrate genomes contain human D-type-like genes, even though orthology has not been confirmed. It appears that D-type genes in vertebrates and insects evolved independently from different origins, which is consistent with the idea of a rapid turnover of D-type genes.

Here, I focused on an early stage of *CYP* gene diversification in vertebrates and showed the emergence of D-type from B-type genes. However, some exceptions

should be noted. For example, *CYP2R1* is categorized as a B-type gene on the basis of its function, but its nucleotide sequence is closely related to other D-type *CYP2* genes. From this observation, it appears that *CYP2R1* has been converted from a D-type to a B-type *CYP* gene. This is supported by the observation that the amino acid sequence of CYP2R1 is highly conserved in all vertebrates, reflecting the high degree of functional constraint on the gene.

4.1.2 The evolution of *CYP* genes is driven by substrate specificity

The birth and death (pseudogenization) rates of B- and D-type genes differed in magnitude: the rates in B-type genes were 0.7 and 0.2 per 100 myr, respectively, whereas those in D-type genes were 12.7 and 6.9 per 100 myr, respectively. Compared with D-type genes, the evolution of B-type genes was highly conserved with regard to their mode of birth and death processes as well as amino acid substitutions. The substrates of B-type enzymes are chemicals that play important roles in metabolism of vitamin D, steroids, and cholesterol. In contrast, the substrates of D-type enzymes are xenobiotics such as plant alkaloids. In light of this substrate specificity, I hypothesize that the conserved evolutionary pattern observed in B-type enzymes reflects the importance and conservation of their substrates, whereas the rapid evolution of D-type enzymes indicates that their substrates are flexible with adapting to changing environmental factors. Future studies on the evolution of substrate-recognition sites (SRS) will be required to examine the hypothesis. SRS is the major regions for the recognition of substrates in *CYP* genes. Six domains were identified for *CYP2* family

and SRSs of other CYP families remain to be examined by crystarographic analyses.

As Table 3-4-1 shows, there are seven conserved B-type genes in vertebrates. There is no gene loss among vertebrates in these genes and it suggests that these genes must be important for vertebrate. In fact, some deteriorated mutations in three of them (*CYP5A1*, *CYP19A1* and *CYP27A1*) are responsible for diseases in human (Table 4-1-1) and all are lethal. Diseases are so far only reported in humans, but even in other vertebrates, if it is predicted to be lethal if they have mutations on the gene.

Table 4-1-1. The disease caused by *CYP* mutations in human

substrate type	gene name	function	disease presentation
D	4F22	arachidonic acid, fatty acid	Lamellar ichthyosis and Non-Bullous congenital ichthyosiform erythroderma (NBCIE)
B	5A1	Thromboxane	Diaphyseal dysplasia with anemia (Ghosal)
	7B1	Bile Acid	Hereditary spastic paraplegia, Congenital bile acid synthesis defect
	11A1	steroid	Congenital adrenal hyperplasia, 46,XX disorders of sex development
	11B1	steroid	Congenital adrenal hyperplasia, 46,XX disorders of sex development etc, aldosteronism
	11B2	aldosterone	Aldosterone synthase deficiency, Glucocorticoid-remediable aldosteronism
	17A1	steroid	Congenital adrenal hyperplasia, 46,XX disorders of sex development
	19A1	aromatase, estrogen	46,XX disorders of sex development, Aromatase excess syndrome
	21A2	steroid	Congenital adrenal hyperplasia, 46,XX disorders of sex development
	27A1	Bile Acid	Cerebrotendinous xanthomatosis
	27B1	Vitamine D3	Type 1 diabetes mellitus, Vitamin D-dependent rickets

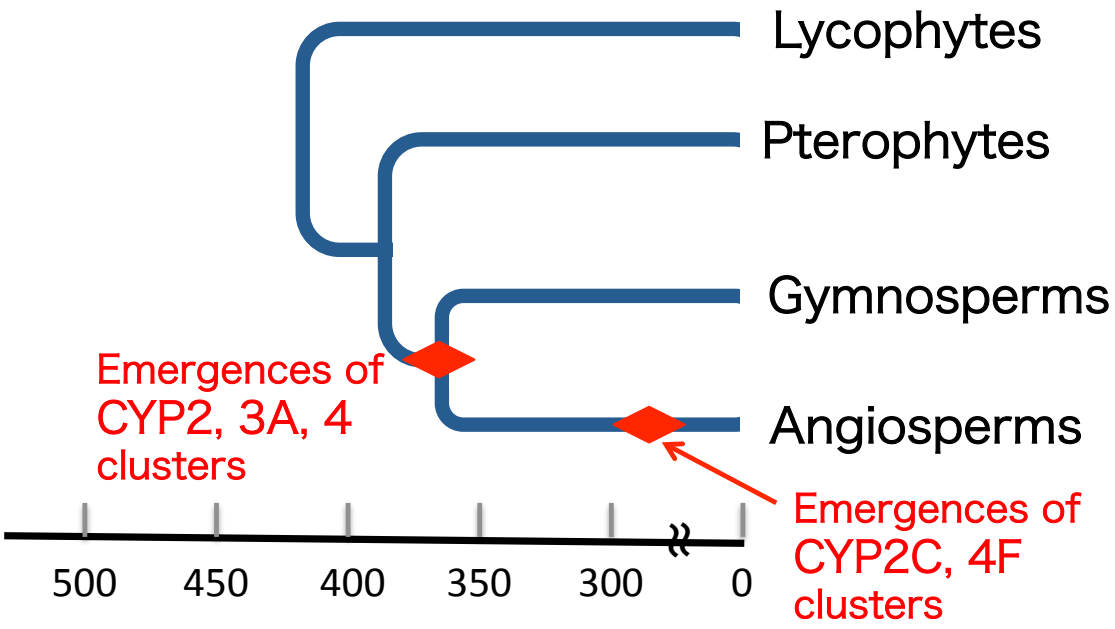
D and B represent D-type and B-type *CYP* gene, respectively. Function and associated disease are shown. Gray shaded rows indicate a gene conserved in all vertebrates.

4.1.3 Evolution of D-type *CYP* genes and the prosperity of plants

As Figure 3-4-1 shows, there are five points of cluster divergence in vertebrates. At those times, land plant species also evolved and expanded the number of species (Figure 4-1-3). Gymnosperms expanded around 360 mya, and *CYP2* cluster in vertebrates had also appeared at that time. After that, *CYP3A* subfamily and *CYP4* family cluster was emerged in 248 and 217 mya, respectively. Just after the divergence of these *CYP* clusters, mammal had been arisen. Most of gymnosperms had already appeared at this era. Moreover, *CYP2C* and *4F* subfamily clusters had diverged 144 and 106 mya, respectively and angiosperm evolved dramatically around this time. Then, primate species diverged around 65 mya and the one unit of *CYP2* family cluster had inverted before the divergence of Haplolhini. At the same period, Dicotyledoneae flourished. As seen above, the evolution of D-type *CYP* genes in vertebrates must be closely rerated to the prosperity of the plants. This must imply the main foods of vertebrate ancestor are plants. D-type *CYP* genes in vertebrates may had been diverged for the detoxification of many alkaloids in plants.

Figure 4-1-3. The evolution of land plants.

Green rectangle shows the emergence of each cluster.



This figure was modified Peason Education.

4.2 Perspectives

According to my analysis, both B- and D-type *CYP* genes exist before the divergence of vertebrates. After the divergence of vertebrates, D-type genes had duplicated frequently in each species and the main biological causes for this divergence must be foods or habitats. On the other hand, B-type genes are conserved among vertebrate species.

Mammals in ocean (whale, killer whale and dolphin) are diverged from the ancestor of elephant or hippopotamus in 105 mya (Figure 4.2.1, Time Tree). These land animals are herbivorous and must have many D-type genes. But marine mammals do not eat land plants in most cases. For example, main foods for Monodontidae are fishes, crustacean, shellfish and segmented worms. It means that they are carnivorous. Phocoenidae also eat fishes, crustacean and squid. These Cetacea changed their food from plants to animals, because their living environment has few plants (Armfield BA *et al*, 2013). For these reasons, it is to be expected that these Cetacea species (whales, dolphins and porpoises) had D-type genes in ancestors but most of these may have been lost their function or lose their functional constraint. But B-type genes must be retained in their genomes as in land mammal genomes. In Cetacea, the DNA sequences in seven species (*Orcinus orea*, *Tursiopus truncate*, *Stenella coeruleoalba*, *Lagenorbynchus actus*, *Phocoenoides dalli*, *Balaenoptera acutorostrata*, *Subalaena glacialis*) are available in my preliminary analysis. The analysis shows that *O. orea* and *T. truncate* have almost the same number *CYP* genes as humans, but other cetacean does not have enough sequence data. According to the phylogenetic tree including humans, cow,

O.orea and *T.truncate*, each gene shows the same tree as species tree. Further analysis about pseudogenization in cetacean is needed to clarify the adaptive evolution for oceans in these species.

Moreover, it is interesting that there is the mammal that eats plants in ocean or in waterfront. They are Sirenia including manatees, dugong and sea cow. They had been diverged from the ancestor of elephants or mammoths in 61.1 mya (Time tree). Manatees eat plants in ocean or waterfront and dugong eats seaweed (Christopher D *et al*., 2000). They eat plants, but ocean plants may have different alkaloids from land plants. This implies that they have different D-type *CYP* gene from other mammals. As far as I searched, 99 *CYP* genes are available in these aquatic mammals. It seems interesting to compare these CYPs with those in land mammals together with foods or habitats.

Comparison of the *CYP* genes in Carnivore is also worthwhile, because they do not take plants. In fact, cats or dogs lost some function of detoxification genes. If they take a piece of onion, they will be die. Therefore I should include CYPs in these species in the future analysis.

In the CYP Homepage, DNA sequences of 17 insect species (*Drosophila pseudoobscura*, *Drosophila melanogaster*, *Anopheles*, *Trialeurodes vaporariorum*, *Bombyx mori*, *Monarch butterfly*, *Lepidoptera*, honeybee, parasite wasp, leaf cutter ant, Argentin ant, seed-harvester ant, fire ant, *Tribolium castaneum*, aphid, Two-spotted spider mite, *Ixodes scapularis*) are available. Their foods or habitats are restricted to

plants compared to other large animals. Foods of most insects are limited to several varieties of plants that are not suitable to vertebrates. Therefore it will be interesting to compare CYPs in these insects with those in vertebrate species. In addition, these insects have D-type *CYP* genes that have different origin from vertebrates. It may be interesting also to examine whether or not the mode of evolution in D-type genes in insects are similar to that in vertebrate D-type gene.

There are many types of CYPs in the world and the number of CYPs identified increases day by day. Even though not all CYPs are included in this study, the comparisons of CYPs between different categories, B- and D-type have unveiled the major trends and illustrated the differences or similarities in the evolutionary mode between them with contributing further and deeper understanding evolution of CYPs.

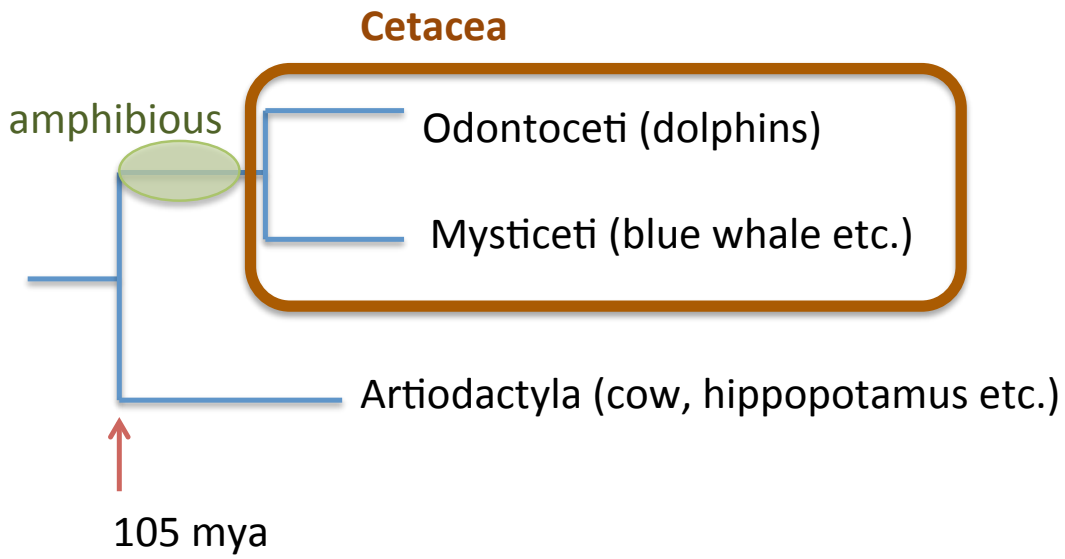


Figure 4-2-1. The phylogenetic tree of Cetartiodactyla

The divergence time of Artiodactyla and Cetacea is 105 mya. Brawn rectangle shows Cetacea.

References

Angata, T., Margulies, E., Green, E. & Varki, A. (2004) Large- scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms. Proc. Natl Acad. Sci. USA 101, 13251–13256.

Aoyama Y, Horiuchi T, Gotoh O, Noshiro M, Yoshida Y (1998) CYP51-like gene of Mycobacterium tuberculosis actually encodes a P450 similar to eukaryotic CYP51. J Biochem 124: 694-696.

Armfield BA, Zheng Z, Bajpai S, Vinyard CJ, Thewissen J (2013) Development and evolution of the uniwue cetacean dentition. Peer J. 1:e24

Christopher D. Marshall, Paul S. Kubilis, Glenn D. Huth, Virginia M. Edmonds, Deborah L. Halin, and Roger L. Reep. (2000) Food-handling ability and feeding-cycle length of manatees feeding on several species of aquatic plants. Mammalogy. 81:1542-1545

Debeljak N, Fink M, Rozman D (2003) Many facets of mammalian lanosterol 14α-demethylase from the evolutionarily conserved cytochrome P450 family CYP51. Arch Biochem Biophys 409: 159-171.

Edgar, A. (2002) The human L-threonine 3-dehydrogenase gene is an expressed pseudogene. BMC Genet. 3, 18.

Estabrook RW, Hildebrandt AG, Baron J, Netter KJ & Leibman K (1971) Biovhem. Biophys, Res. Commun. 42:132-139

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-76

Feyereisen R (1999) Insect P450 enzymes. Annu. Rev. Entomol. 44:507-533

Feyereisen R (2011) Arthropod CYPomes illustrate the tempo and mode in P450 evolution. Bioch et Biophys Acta 1814: 19-28.

GenomeMatcher: A graphical user interface for DNA sequence comparison. BMC Bioinformatics 2008, 9:376 (16 September 2008) Ohtsubo Y1, Ikeda-Ohtsubo W, Nagata Y, Tsuda M.

Gotoh O (2012) Evolution of cytochrome p450 genes from the viewpoint of genome informatics. Biol Pharm Bull 35: 812-817.

Hahn, Y., Jeong, S. & Lee, B. (2007) Inactivation of MOXD2 and S100A15A by exon deletion during human evolution. Mol. Biol. Evol. 24, 2203 – 2212.

Hashimoto Y, Yamano T and Mason HS (1962) An Electron Spin Resonance Study of Microsomal Electron Transport. J. Niol. Chem. 237:3843-3844

Hayakawa, T., Aki, I., Varki, A., Satta, Y. & Takahata, N. (2006) Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics 172, 1139 – 1146

Hedges SB, Dudley, Kumar S (2006) Time Tree: a public knowledge-base of divergence times among organisms. Bioinformatics 23:2971-2.

Hoffman SMG, Hu S (2006) Dynamic evolution of the CYP2ABFGST gene cluster in primates. Mutation Res 616: 133-138.

Hu S, Wang H, Knisely AA, Raddy S, Kovacevic D et al. (2008) Evolution of the CYP2ABFGST gene cluster in rat, and a fine-scale comparison among rodent and primate species. Genetica 133: 215-226.

Jones D, Taylor WR, Thronton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275-282.

Karl Walter Mock (2003) Vertebrate UDP-glucuronosyltransferases: functional and evolutionary aspects. Biochem. Pharmacol. 66:691-696

Kato R, Yokoi K, Yamazoe Y (2010) Drug metabolism, 2nd edition, pp

Klingenberg M. (1958) Pigments of rat liver microsomes. Arch Biochem Biophys 75: 376-386.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, Mcwilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948

Malloty JC, Crudden G, Johnson BI, Mo C, Pierson CA, Bard M and Craven RJ (2005) Dap1p, a Heme-Binding Protein that Regulates the Cytochrome P450 Protein Erg11p/Cyp51p in Saccharomyces cerevisiae. Mol Cell Biol. 25:1669-1679.

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 18: 50-60.

Mao G, Seebeck T, Schrenker D, Yu O (2013) CYP709B3, a cytochrome P450 monooxygenase gene involved in salt tolerance in Arabidopsis thaliana. BMC Plant Biol 13:169

Meunier B, de Visser SP, Shaik S (2004) Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. Chem Rev 104: 3947-3980.s

Munro AW, Lindsay G (1996) Bacterial cytochromes P-450. Mol Microbiol 20: 1115-1125.

Nebert DW, Dalton TP (2006) The role of cytochrome P450 enzymes in endogenous signaling pathways and environmental carcinogenesis. Nature Rev Cancer 6: 947-960.

Nei M. and Kumar S. (2000). Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nelson D, Werck-Reichhart D (2011) A P450-centric view of plant evolution. Plant J. 66:194-211

Nelson DR (1998) Metazoan cytochrome P450 evolution. Comp Biochemi and Phys Part C: Pharm Toxic Endocr 121: 15-22.

Nelson DR (1999) Cytochrome P450 and the individuality of species. Arch Biochem Biophys 369: 1-10.

Nelson DR (2009) The cytochrome p450 homepage. Human Genomics 1:59-65.

Nelson DR, Goldstone JV, Stegeman JJ (2013) The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s. Phil Trans R Soc B 368: 20120474.

Nelson DR, Koymans L, Kamataki T, Stegeman JJ, Feyereisen R, Waxman DJ, Waterman MR, Gotoh O, Coon MJ, Estabrook RW, Gunsalus IC, Nebert DW (1996)

P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics, 6(1): 1-42

Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14(1): 1-18

Omura T, Sato R (1962) A New Cytochrome in Liver Microsomes. J Biol Chem 237: PC1375-PC1376.

Omura T, Ishimura Y and Fujii-Kuriyama Y (1993) Cytochrome P450, 2nd ed., Kodansha.

Omura T, Ishimura Y and Fujii-Kuriyama Y (2009) Cytochrome P450, 2nd ed., Kodansha.

Poulos TL, Finzel BC, Gunsalus IC, Wagner GC & Kraut J (1985) The 2.6-A crystal structure of Pseudomonas putida cytochrome P-450. J. Biol. Chem. 260: 1612216130

Qi X, Bakht S, Qin B, Leggett M, Hemmings A, et al. (2006) A different function for a member of an ancient and highly conserved cytochrome P450 family: From essential sterols to plant defense. Proc Natl Acd Sci USA 103: 18848-18853.

Quaderer R, Omura S, Ikeda H, Cane DE (2006) Pentalenolactone biosynthesis. Molecular cloning and assignment of biochemical function to PtlI, a cytochrome P450 of Streptomyces avermitilis. J Am Chem Soc 128: 13036-13037.

Rewitz KF, O'Connor MB & Gilbert LI (2007) Insect Biochem. Mol. Biol. 37:741-753

Rezen T, Debeljak N, Kordis D, Rozman D (2004) New aspects of lanosterol 14a-demethlase and cytochrome P450 evolution: Lanosterol/cycloartenol diversification and lateral transfer. J Mol Evol 59: 51-58.

Rzhetsky A, Nei M (1992) Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J Mol Evol 35:367-75

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

Sandstrom P, Welch WH, Blomquist GJ and Tittiger C (2006) Insect Biochem. Mol. Biol. 36:835-845

Sawai H, Go Y, Satta Y (2008) Biological implication for loss of function at major histocompatibility complex loci. Immunogenetics 60: 295-302.

Scott JG & Wen Z (2001) Pest Manag. Sci. 57:958-967

Scott JG (1999) Insect Biochem. Mol. Biol. 29:757-777

Sea urchin genome sequencing consortium (2006) The genome of the sea urchin Strongylocentrotus purpuratus. Science 314: 941-952.

Stedman, H. H. et al. (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428, 415–418.

Sutherland TD, Unnithan GC, Andersen JF, Evans PH, Murataliev MB, Szabo LZ, Mash EA, Bowers WS & Feyereisen R (1998) Proc. Natl. Acad. Sci. USA 95:12884-12889

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. Mol Biol Evol 28: 2731-2739.

Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genetics 3: e67

Urabe K, Kimura A, Harada F, Iwanaga T, Sasazuki T (1990) Gene conversion in steroid 21-hydroxylase genes. Am J Hum Genet 46: 1178-1186.

Wang, X., Grus, W. E. & Zhang, J. (2006) Gene losses during human origins. PLoS Biol. 4, 366–377.

Winter, H., Langbein, L., Krawczak, M., Cooper, D. N., Jave-Suarez, L. F., Rogers, M. A., Praetzel, S., Heidt, P. J. & Schweizer, J. (2001) Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. Hum. Genet. 108, 37 – 42.

Yoshida Y, Aoyama Y, Kumaoka H and Kubota S (1977) A highly purified preparation of cytochrome P-450 from microsomes of anaerobically grown yest. Biochem Biophys. Res. Commun. 78:1005-1010.

Yoshida Y, Aoyama Y, Noshiro M, Gotoh O (2000) Sterol 14-demethylase P450 (CYP51) provides a breakthrough for the discussion on the evolution of cytochrome P450 gene superfamily. Biochem Biophys Res Comm 273: 799-804.

Yasukochi Y, Satta Y (2011) Evolution of the CYP2D gene cluster in humans and four non-human primates. Genes Genet Syst 86:109-116