YABBI 3097
TYPED / 15/10/02
ARTICLE IN PRESS
No. of pages: 7
DTD 4.3.1/SPS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ACADEMIC PRESS

Archives of Biochemistry and Biophysics xxx (2002) xxx–xxx

ABB

www.academicpress.com

Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution

David R. Nelson*

Department of Molecular Sciences, University of Tennessee, 858 Madison Avenue, Memphis, TN 38163, USA

Received 3 June 2002, and in revised form 17 September 2002

Abstract

The fugu (pufferfish) genome has been sequenced, and a second genome assembly was released 17 May 2002. Exhaustive searches were made to identify all P450 genes and pseudogenes from the earlier release of 26 October 2001. P450 genes assembled as completely as possible from these data were used to do additional searches of the newer assembly and all P450 genes and pseudogenes in the available fugu sequence data have been identified, compared to human P450s, and assigned names. There are 54 P450 genes in fugu and 1 nearly intact pseudogene (CYP3A50P). CYP1A is missing much of its N-terminal half; however, 45 P450 genes are completely assembled. Eight others are lacking only one or two exons or less. CYP2X4 is known only from an EST. This may be a 55th P450 gene if it represents an accurate sequence. In addition to 2X4, there are 16 other pseudogene fragments or small pieces of P450 genes. At the P450 family level, 17 of 18 mammalian families are found in fugu. CYP39 is the only CYP family missing and it is not seen in any other fish sequence data either. The CYP2 family shows the largest degree of divergence. In the CYP2 family, only CYP2R1 and CYP2U1 are conserved as recognizable subfamilies across species. Intron–exon boundaries are largely preserved across 420 million years of evolution.

Keywords: Cytochrome P450; Evolution; CYPs; *Fugu* genome; Comparative genomics

Genomic comparisons in eukaryotes are beginning to move out of the well-characterized genetic model organisms like yeast, fly, worm, mouse, and *Arabidopsis* and into the realms of unexplored territory. Eukaryotes are ancient. Molecular fossil steranes, such as cholestane, have been found in 2.7-billion-year-old Archaen shales from Australia [1]. These are interpreted as signature molecules of eukaryotic life, suggesting a long history for eukaryotes. Recent genome comparisons were limited to mouse and human (96 million years of divergence [2]), followed by human and fly or worm (protostome–deuterostome divergence, 670 million years ago; mya [3]). This left a 574 million year gap in the history of our lineage. One of the major dichotomies in animal evolution that falls in this gap is the ray-finned fish and tetrapod divergence. This has been dated to about 420 mya [4]. The release of the fugu genome assembly by the International Fugu Sequencing Consortium provides data for following the evolution of vertebrate animals over that time frame.

Our interest lies in the cytochrome P450 superfamily and its evolution. It is the goal of the Committee for a Standardized Cytochrome P450 Nomenclature to name all P450 genes as they are discovered to aid accurate communication in the research community and to foster better understanding of this gene family. In this pursuit, 1925 P450 genes have been named as of 1 June 2002. This number should grow to over 2500 by the end of 2002 as all 478 rice P450 genes and about 150 *Phanerochaete chrysosporium* (white rot fungus) P450 genes are named. About 60 more P450s from each of two *Ciona* (sea squirt) genomes are also in the process of being assembled and named now. All of this information is posted at the Cytochrome P450 home page (http://drnelson.utmem.edu/CytochromeP450.html) for unrestricted access by all.

P450 genes are named based on their evolutionary relationships, so it is imperative to correctly assemble

* Fax: 1-901-448-7360.
E-mail address: dnelson@utmem.edu

0003-9861/02/$ - see front matter © 2002 Published by Elsevier Science (USA).
PII: S0003-9861(02)00553-2

the protein sequences from new genomes to be able to build phylogenetic trees and name the genes. Automated programs employed by genome annotation projects are often unable to correctly assemble P450 genes. These programs often miss N-terminal exons, or skip exons, producing truncated P450 assemblies. They also fuse P450s that appear in gene clusters, resulting in chimeric sequences. This is especially a problem in plants that frequently have clusters of 10 or more P450s in a row. The *Phanerochaete* genome has P450s with three- and five-amino-acid exons that will probably not be detected by gene-finding programs. These programs offer a starting point, but there is no substitute for expert human annotation.

Here is described the cytochrome P450 complement of the Japanese pufferfish *Takifugu rubripes*, as well as the process used to find and assemble the genes. Many aspects of this method can be automated and it is our goal to streamline the gene discovery part; however, final assembly will probably still require the human touch.

Methods

It is a fairly simple matter to do a BLAST search and find hits to a P450 sequence in a database. It is not so simple to find all members of a large gene family in a genome. A systematic search procedure needs to be implemented. With the fugu genome, two BLAST servers are available to search several data sets. At the Joint Genome Institute in Walnut Hill, California, the server http://bahama.jgi-psf.org/fugu/bin/fugu_search is able to search the 26 October 2001 genome assembly using five BLAST formats, including TBLASTN, which was the most useful for this work. The "expect value" could be increased only to 10^{-3}, which might not pick up exons with low sequence conservation often seen in the middle of P450s. An expect value of 1 or 10 is often useful to find these regions. The MRC Human Genome Mapping Consortium Resource Center Fugu BLAST server http://fugu.hgmp.mrc.ac.uk/blast/ offers multiple data sets, including cosmids, cosmid ends, BAC ends, cDNA, and scaffolds. It used to include zebrafish, but that has moved to a separate server and it has been replaced by 320 Mb of *Tetraodon nigroviridis* (freshwater puffer) sequence data. This server allows expect values of up to 1000 and the option to turn off the low-complexity filter, which is best when BLASTing with proteins. Both genome assemblies (26 October 2001 and 17 May 2002) are BLAST searchable.

It was necessary to find all P450-containing fragments from all of these databases. An assumption was made that fish P450s would fall into family groups (greater than 40% sequence identity), similar to mammalian P450s. Therefore, it would be necessary only to search the All Fugu database with 18 sequences (one from each mammalian family) to find all members of P450 in fugu. Since some mammalian subfamilies are nearly in distinct families (they are on the border of the 40% definition), these subfamilies were also included in the search. Since positive hits were often in the mid-20% range and they could still be identified as valid P450s, this strategy was deemed sufficient to pick up most vertebrate P450 members, even if they were in a new family not seen in mammals. It should also be able to detect a wide variety of pseudogenes. This strategy was validated, as no P450s outside the 18 mammalian families were discovered.

Phase 1: BLAST searching

The process of finding all members of the P450 family in fugu can be divided into three stages. The first is the identification of all accession numbers in the data set that have P450s or parts of P450s on them. The BLAST searches are done with each member of the query set and the output is examined to see which are legitimate hits. The accidental matches are thrown out. An expect value of 10 is used in this process. This will give some false matches, but there are true matches that are found very close to the expect value of 10, they are usually from the middle region of the P450s, which is poorly conserved. For automation of this process, it might be possible to make the expect value 1 instead of 10. This would make the false positives very rare, but it would miss some true hits. The rationale for doing this would be that the true hits would have an expect value lower than 1 when another more similar query sequence was used.

Once the first search is done and all false hits have been discarded, a file is made of the accession numbers for the true hits. This is sorted alphanumerically and saved for comparison to the next search. Search 2 is done and the same procedure is followed. Now the two lists of accession numbers are compared, and any duplicate hits are deleted from the list. In an automated version of this process, it might be useful to keep a tally of the number of times an accession number is found and the percentage identity of the BLAST result for the best match. The rare hits might be in new families. The process is repeated with all members of the query set. In an automated version of this process, there is no reason every human sequence could not be used instead of the 18–20 used in the manual search. This would be more comprehensive and might pick up a few more accession numbers that would be missed otherwise.

Phase 2: sorting into gene families and individual genes

A BLAST server has been set up on a Linux server at the University of Tennessee in the Bioinformatics Suite. This server has all P450 members from 12 different species, including human, rat, and mouse (http://

132.192.64.52/p450.html). This is a curated data set that is nonredundant and comprehensive. As new members are found, they are added to these databases. The latest sequences are 188 contigs from the sea squirt *Ciona*. The server is linked from the Cytochrome P450 home page and is available to the world as a service. The BLAST search results from phase 1 are compared to the complete human set of P450s to identify the best match. This is the process of family or subfamily identification for each accession number. This procedure can be shortened somewhat by early identification of multiple accession numbers from the same gene. Once identified, these accession numbers do not have to be searched again. The results from the phase 1 searches resulted in 332 accession numbers (http://drnelson.utmem.edu/fugu.alpha.html). These were sorted into 17 gene families and numerous subfamilies by BLASTing against the human set.

Individual genes were placed into gene bins for later assembly. These often had multiple exons, but were not yet assembled at the level of GT-AG boundaries. The exons were put in the order in which they occurred in the genomic DNA and these rough gene translations were added to the P450 BLAST server. To identify all unique protein contigs from these data, each sequence was BLAST searched against all other P450 protein sequences from fugu, and overlapping pieces were sorted into the same gene bins. This process reduced the number of contigs from 332 accessions to 75 nonoverlapping gene contigs (http://drnelson.utmem.edu/fugu.fasta.html). Further refinement has lowered the number of contigs to 71.

Phase 3: assembly

The genes were still not assembled to identify the intron–exon boundaries. This step required comparison to mammalian gene models and, as the process created complete fish genes, to fish gene models. Many of the exon boundaries were in the same place between humans and fugu. The phase was also the same at these boundaries. That made the process of gene assembly easier. Several gene clusters were found on the same scaffold and these genes tended to be highly similar, making assembly by comparison a possibility. The few pseudogenes that were found were recognized as pseudogenes by multiple frameshifts, in-frame stop codons, and missing exons.

Results and discussion

The cytochrome P450 set from fugu is not yet complete. The genome project claims that the second release has about 90% coverage of the nonrepetitive part of the genome in the 17 May 2002 assembly. There are 45 full-length genes and 8 more that are missing only a small portion (7 amino acids up to one or two exons). CYP1A is missing about 200 amino acids after the first coding exon. In addition to these 54, there is one EST sequence, CYP2X4, that is not found in the assembly of the genome and this may be a 55th P450 gene. 2X4 codes for the last 114 amino acids from the PKG motif to the end of the gene. One gene is a full-length pseudogene, CYP3A50P. This gene has a deletion and frameshift in the heme signature and a few other small defects. It is probably a very recent pseudogene. There are 15 other short fragments. Most or all of these partials are pseudogenes and they are as complete as they can get. Some of the incomplete genes had orthologs in *Tetraodon*, zebrafish, or other fish, like medaka or trout. The best match was used to search once more in fugu for the missing pieces, but they have not been found. These parts of the gene are missing from the available data. Table 1 lists all 71 Fugu P450 genes and pseudogenes. The scaffold numbers are from the 17 May 2002 assembly. Some of the smaller fragments were not found in this assembly and they have different numbers. Those beginning with Fc are cosmid end sequences. CYP8A3P is from a cosmid. The FE sequence (CYP2X4) is an EST. The actual P450 protein sequences assembled from these sequences are posted on the P450 home page under pufferfish.

One family, CYP39, is missing in fugu. This sequence is missing from zebrafish and every other fish in GenBank, both in the EST database and in mRNA or genomic sequences. This gene probably is unique to mammals or at least has arisen in the tetrapod lineage after it diverged from fish. This sequence has not been found outside of mammals yet (as in birds or reptiles). The function is oxysterol 7α-hydroxylase with a preference for 24-hydroxycholesterol [5]. This gene provides an alternative pathway to the synthesis of 7α-hydroxylated bile acids. The other gene that does this is another P450 CYP7B1, with very little sequence similarity. The two genes are sexually dimorphic, with CYP39 being expressed to a higher level in females, while CYP7B1 is higher in males.

Fig. 1 shows a UPGMA phylogenetic tree of 54 fugu P450s compared with 60 human P450s (includes 3 pseudogenes) and 8 other fish P450s. Only full-length or nearly full-length protein sequences were used, so the fugu 1A gene is not in this tree. The other fish sequences were included when there was no human ortholog or the fish sequence was a much better match (see 4T2 from seabass, 2X1 from catfish, and 2P3 from killifish). Five prominent clans are labeled. A P450 clan is a cluster of P450 families that consistently are grouped together and are similar to a clade in a species tree. The 122 sequence alignment (22 pages) is posted on the P450 home page under pufferfish (http://drnelson.utmem.edu/fuguhum.aln.html).

Table 1
Fifty-four scaffolds including 54 different genes and 17 pseudogenes or fragments

1A1	FS:S006359 Scaffold_6359
1B1	FS:S007281 Scaffold_7282
1C1, 1C2	FS:S000289 Scaffold_289
2K9	FS:S003854 Scaffold_3854
2K10, 2K13P/2K14P; contig may be out of order	FS:S007893 Scaffold_7894
2K10 exons 1 and 2	FS:S000255 Scaffold_255
2K11, 2K15P	FS:S006775 Scaffold_6776
2K12P	FS:S003006 Scaffold_3006
2K15P	FS:S008343 Scaffold_8344
2N9, 2N10, 2N11, 2N12, 2P4 exons 1, 2, 3, 5, 6	FS:S000805 Scaffold_805
2P4 exons 7, 8, 9	FS:S001425 Scaffold_1425
2P5P	FS:c060F24y1 LPC.22843.y1
2R1	FS:S000037 Scaffold_37
2R2P	Fc:c104I03x1 LPC.39565.x1
2R3P	Fc:c068L08y2 LPC.26046.y2
2U1	FS:S002188 Scaffold_2188
2X2	FS:S003334 Scaffold_3334
2X3 exons 1, 3, 4	FS:S005546 Scaffold_5546
2X3 exons 5–11	FS:S006918 Scaffold_6919
2X4 (EST)	FE:EFRy002apsE4
2X5P	FS:S002630 Scaffold_2630
2Y1, 2Y2	FS:S000830 Scaffold_830
2Z1, 2Z2	FS:S002784 Scaffold_2784
3A47	FS:S000150 Scaffold_150
3A48, 3A49, 3A50P	FS:S000035 Scaffold_35
3B1, 3B2	FS:S001762 Scaffold_1762
4F28	FS:S001453 Scaffold_1453
4T5	FS:S009035 Scaffold_9036
4V5	FS:S002098 Scaffold_2098
4V5-like lone PERW exon	FS:S000209 Scaffold_209
5A1	FS:S000031 Scaffold_31
7A1	FS:S002374 Scaffold_2374
7C1	FS:S005652 Scaffold_5652
8A1	FS:S002619 Scaffold_2619
8A2	FS:S000268 Scaffold_268
8A3P	Fc:c61O19bD10 Fc:c061O19bD6
8B1, 8B3P	FS:S002259 Scaffold_2259
8B2 C-terminal	FS:S006668 Scaffold_6669
8B2 N-terminal	FS:S008552 Scaffold_8553
11A1	FS:S001739 Scaffold_1739
11B1	FS:S001181 Scaffold_1181
17A1	FS:S002402 Scaffold_2402
17A2	FS:S005320 Scaffold_5320
17A fragment a	Fc:c028I22x1 LPC.10549.x1
17A fragment b	Fc:c028I22x2 LPC.10549.x2
19A1	FS:S005700 Scaffold_5700
19A2	FS:S000694 Scaffold_694
20	FS:S000069 Scaffold_69
21	FS:S000061 Scaffold_61
24	FS:S002393 Scaffold_2393
26A1	FS:S009376 Scaffold_9377
26B1	FS:S000306 Scaffold_306
26C1	FS:S004331 Scaffold_4331
27A1	FS:S000554 Scaffold_554
27A2	FS:S000138 Scaffold_138
27A3	FS:S002565 Scaffold_2565
27A exon 4	Fc:c110F07x1 LPC.42075.x1
27B1	FS:S000063 Scaffold_63
27C1	FS:S000106 Scaffold_106
39	Not detected in fish
46A1, 46A2P, 46A3P	FS:S000256 Scaffold_256
51	FS:S000877 Scaffold_877

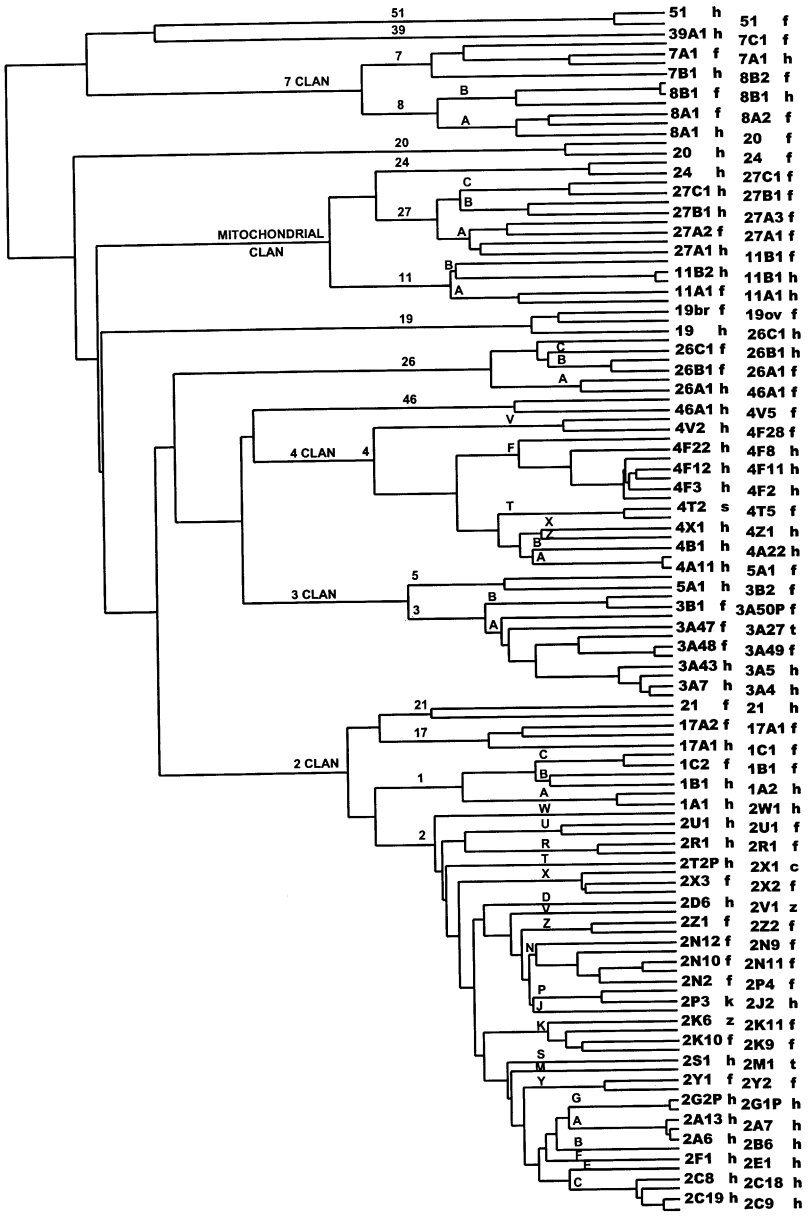

Fig. 1. UPGMA tree of 54 fugu, 60 human, and 8 other fish P450s. Species are indicated by f, h, z, c, k, s, and t for fugu, human, zebrafish, catfish, killifish, seabass, and trout, respectively. The sequence alignment used to make this tree can be downloaded from http://drnelson.utmem.edu/fuguhum.aln.html. More information on fugu P450 exon structure can be obtained from the pufferfish section of the cytochrome P450 home page http://drnelson.utmem.edu/puffer.html.

As mentioned above, only 1 P450 family of 18 was missing in fugu and probably in fish in general. All 17 other P450 families have clear orthologs present. This means that the diversity of P450 families seen in mammals predated the tetrapod–ray finned fish divergence (about 420 mya [4]). When individual families are examined in more detail, the intron–exon boundary structure is also conserved. The CYP2 family members have nine exons in fish and humans and the boundaries are in the same positions. A preliminary look at the *Ciona* P450s shows many CYP2-like sequences, but they would not be named CYP2. They would probably belong to the 2 clan. These sequences have nine exons also, but none of the eight boundaries are in the same place.

An overview of the relationships of the P450s found in human and fugu is shown in Fig. 2. Sixty-one human sequences (57 functional genes and 4 pseudogenes) are linked to their orthologs in 71 fugu contigs. The left side shows the CYP2 family. This is the least conserved family, with most members falling into new subfamilies, except CYP2R1 and 2U1, which show clear orthologous relationships. The function of 2R1 and 2U1 are not known, but I propose they act on endogenous substrates rather than exogenous substrates based on their con-

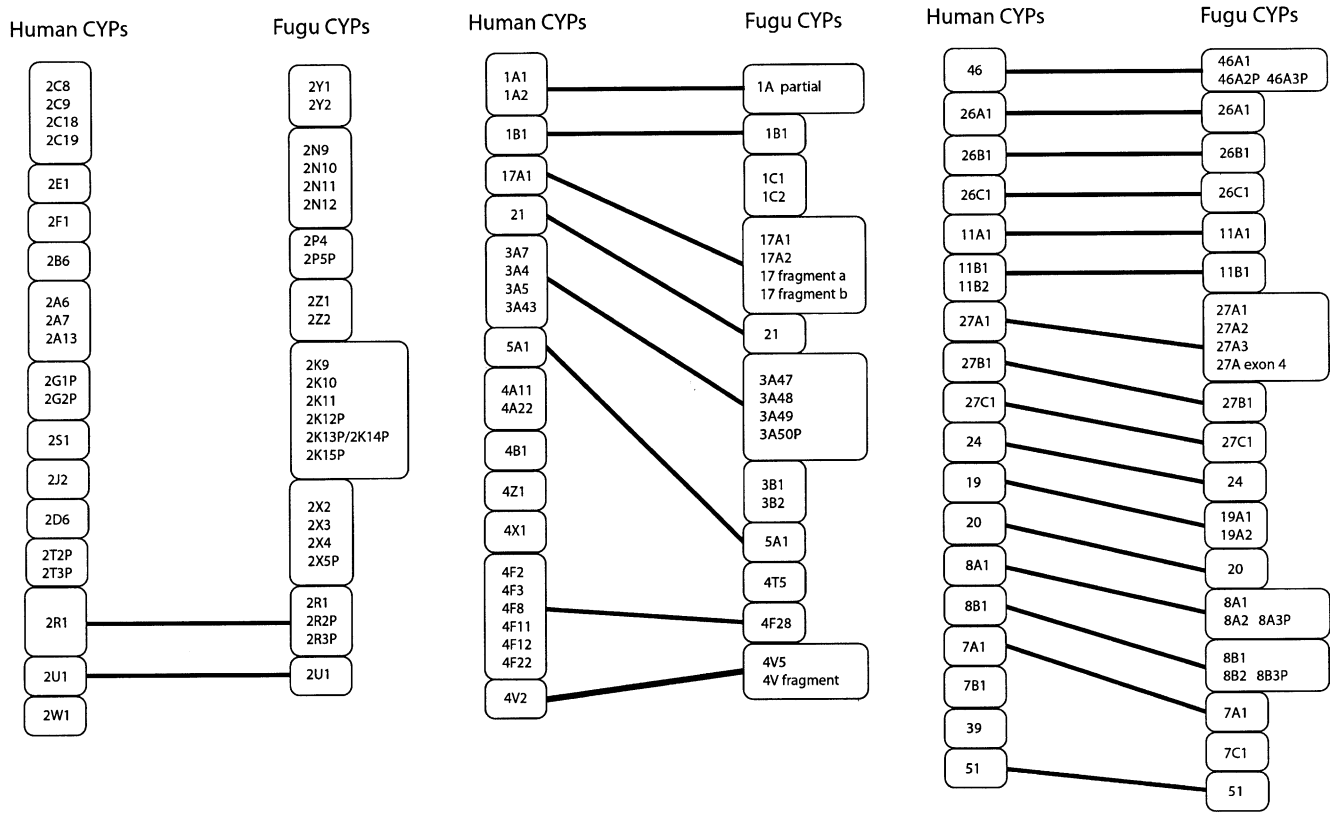

Fig. 2. A graphical representation of the relationships between P450 subfamilies from fugu and human. All fugu P450s and pseudogenes are included, but not all human pseudogenes are included. CYP2G and CYP2T subfamilies are known only as pseudogenes in human, but they are functional in rodents, so they are included. CYP39 is the only family without an ortholog in fish.

servation. The CYP2 family in general is involved in metabolism of foreign compounds and so it is not surprising that it is highly variable over 420 million years. Fig. 1 shows that CYP2N, P, V, and Z are related to CYP2D and CYP2J in mammals.

The middle of Fig. 2 shows families 1, 3, 4, 5, 17, and 21. All six of these families are present in both species, but there are subfamilies that are new in each species, such as 1C and 3B in fugu and 4A, 4B, 4X, and 4Z in human. The tree in Fig. 1 suggests that human 4A, 4B, 4X, and 4Z might be derived from the 4T subfamily of fish. The right side of Fig. 2 shows nearly a one-to-one correspondence among the remaining 11 families. Here we may be seeing some remnants of a genome duplication event that took place in teleost fish but not in tetrapods [6,7]. Fugu has two CYP19s, CYP8As, and CYP8Bs; three CYP27As, and two CYP17s (Fig. 2, right). There are only single copies of these genes in most mammals. Pig is an exception that has three CYP19 genes, but they arose after the mammalian radiation [8,9]. The CYP19s are responsible for synthesis of estrogen from testosterone by aromatization of the A ring. Fugu has a brain form and an ovary form (br and ov in Fig. 1), showing that there could be specialization of the production and use of estrogen in these tissues. This has been reported in zebrafish [10] and goldfish [11]. CYP8B1 and CYP8B2 are so similar that these had to be created very recently and could not be from a more distant genome duplication. The fish genome duplication had to take place after the 420-mya split between fish and tetrapods. The time window is estimated to be between 420 and 300 mya. That means the branch points on the tree between duplicated fugu sequences should be less than, but similar to, human–fugu orthologs. This is true for 8A1, 8A2, 19ov and 19br, 17A1 and 17A2, and the CYP27As.

The results show that the defining characteristics of vertebrate P450s have not changed much in 420 million years. Only one new family, CYP39, is seen in mammals, and this acts in a pathway after CYP46. This new step in an alternate pathway to bile acids [5] is a testament to the apparent value of P450s in biology. New functions are probably occurring in many different lineages. Analyzing whole genome sets of gene families will identify these interesting lineage-specific differences. For example, the fish genome has new CYP1C, CYP3B, and CYP7C subfamilies not seen in mammals. The CYP1C genes were first observed in the scup by Godard, Said, and Stegeman (GenBank Accession Nos. AF131885 and AF235138). Gene defects or knockouts in these fish-

specific P450s might have unique phenotypes. Such mutations might be detected in the large-scale zebrafish mutagenesis screening that is under way [12,13]. Preliminary analysis of *Ciona intestinalis* and *Ciona savignyi* (tunicates or urochordates) P450s shows that there is not the strong correspondence between P450 families as seen here with fugu. This next level on the time scale (about the time of the Burgess Shale and the Cambrian explosion 540 mya) may have more to tell about the origins of P450 families during deuterostome evolution.

References

[1] J.J. Brocks, G.A. Logan, R. Buick, R.E. Summons, Science 285 (1999) 1025–1027.
[2] M. Nei, P. Xu, G. Glazko, Proc. Natl. Acad. Sci. USA 98 (2001) 2497–2502.
[3] F.J. Ayala, A. Rzhetsky, F.J. Ayala, Proc. Natl. Acad. Sci. USA 95 (1998) 606–611.
[4] P. Ahlberg, A. Milner, Nature 368 (1994) 507–514.
[5] J. Li-Hawkins, E.G. Lund, A.D. Bronson, D.W. Russell, J. Biol. Chem. 275 (2000) 16543–16549.
[6] S.F. Smith, P. Snell, F. Gruetzner, A.J. Bench, T. Haaf, J.A. Metcalfe, A.R. Green, G. Elgar, Genome Res. 12 (2002) 776–784.
[7] J.S. Taylor, Y. Van de Peer, I. Braasch, A. Meyer, Philos. Trans. R. Soc. London B Biol. Sci. 356 (2001) 1661–1679.
[8] I. Choi, D.L. Troyer, D.L. Cornwell, K.R. Kirby-Dobbels, W.R. Collante, F.A. Simmen, DNA Cell Biol. 16 (1997) 769–777.
[9] L.G. Graddy, A.A. Kowalski, F.A. Simmen, S.L. Davis, W.W. Baumgartner, R.C. Simmen, J. Steroid Biochem. Mol. Biol. 73 (2000) 49–57.
[10] E.F. Chiang, Y.L. Yan, Y. Guiguen, J. Postlethwait, B.c. Chung, Mol. Biol. Evol. 18 (2001) 542–550.
[11] G.V. Callard, A. Tchoudakova, J. Steroid Biochem. Mol. Biol. 61 (1997) 387–392.
[12] A. Amsterdam, S. Burgess, G. Golling, W. Chen, Z. Sun, K. Townsend, S. Farrington, M. Haldi, N. Hopkins, Genes Dev. 13 (1999) 2713–2724.
[13] G. Golling, A. Amsterdam, Z. Sun, M. Antonelli, E. Maldonado, W. Chen, S. Burgess, M. Haldi, K. Artzt, S. Farrington, S.Y. Lin, R.M. Nissen, N. Hopkins, Nat. Genet. 31 (2002) 135–140.