Dallas Talk June 3, 2005

[slide 1]
The evolution of animal cytochrome P450s from sponges to mammals.

As one reviews the history of life, it is marked by innovations and increases in complexity. This has been studied for nearly 150 years at the level of anatomy and morphology and depicted in trees like this one of Ernst Haeckel from 1891 [slide 2]. Almost 80 years later in 1969, Pauling and Zuckerkandel applied the idea of a molecular clock to cytochrome c and ushered in the era of protein sequence comparison. The evolution of life is now viewed from the molecular perspective. Many individual genes have been used to construct trees of life like Haeckel's, though they have a slightly more modern look now [slide 3]. What is new in the past few years is the availability of whole eukaryotic genomes. With these vast data sets, it is possible to follow the evolution of whole gene families and not just single genes. This brings us back to the point of innovations and increases in complexity. Ultimately, these changes will have a genetic basis. We should be able to look in these genomes and find the changes in molecules that result in changes in form, biochemistry and physiology. Today we will focus on just one branch, the animals.

The world's sequencing capacity is on the order of several hundred million bases per day. Among the mammals, human, mouse and rat have been completed and the chimp, dog and cattle genomes have been assembled based on the human sequence. One does not expect large sequence changes in the CYP genes of such close neighbors as mouse and rat, but there are substantial changes in gene cluster organization. [slide 4] This is a comparison of the mouse and rat Cyp4abx gene cluster. Red lines indicate orthology. The gene order is conserved but even over 20 million years there is much gene duplication , some in each species. Overall, there is greater expansion in the mouse. This whole cluster in mammals appears to be derived from the single CYP4T gene of fish. The opossum genome may catch this gene cluster in an intermediate stage of evolution.

[slide 5] Here we see the Cyp2d gene cluster. Once again there is preservation of gene order and expansion exclusively in the mouse. Remember that humans have only one CYP2D6 gene.

Aside from the well known genome projects like human, mouse and rat, many other genomes are being sampled. The trace archive at NCBI has the raw trace files for over 300 species. This graph [slide 6] shows the growth of the trace archive since 2001. There are 705 million traces or about 700 billion letters in the trace files. You might be surprised at what is already sequenced. This slide [slide 7] shows the top 36 animals with more than one million sequence reads each. Notice that second on the list is the opossum. This first genome of a marsupial has just been assembled as a preliminary genome assembly (version 0.5) at ENSEMBL. Also of interest for evolutionists is the duck-billed platypus with over 27 million reads. Some other mammals include an elephant, a hedgehog, and the nine-banded armadillo, a favorite here in Texas. The amphioxus and hydra are shown in blue because I will talk about them in more detail today.

To follow the evolution of a gene family like cytochrome P450, one needs a diverse collection of genomes to sample [slide 8]. The set shown here would provide wide but not deep coverage of animal evolution. In the upper left is a sponge currently being sequenced at the Joint Genome Institute in California, but the data is not public yet. Today, the only known P450 from sponge is CYP38, which is a CYP20 ortholog. This is juvenile *Ciona intestinalis*, a sea squirt whose genome has already been published. Hydra, amphioxus and the opossum are in the trace archive.

The main problem with the trace archive is the lack of good tools to search the data. At NCBI you can only use Megablast, which is optimized for high percent identity matches at the nucleotide level. There is no TBLASTN search of protein against DNA at high expect values. This makes finding distant matches impossible.

To get around this limitation, I downloaded 2 million sequences via ftp from the hydra and amphioxus genomes. I placed these on my Mac and set up a standalone blast server to search the data with a protein query sequence at an expect value of 10. This permitted finding all P450 sequences in the data.

I will show you what I found in a minute, but first I would like to talk about time scales and evolution. All animals share a common ancestor about 700 million to one billion years ago. But no animal today is the common ancestor to any other. They have all been on separate tracks since they diverged from their last common ancestor. So amphioxus is not the common

ancestor to all vertebrates, it is the descendant of that ancestor. Therefore, the P450 sequences from amphioxus are not ancestral sequences. Similarly, the hydra is not the ancestor to bilateral animals. Please keep that in mind as we continue to look for features of CYP evolution in these sequences.

Ciona is a key organism since it branched off early in chordate evolution, before steroid and vitamin D receptors evolved. Therefore, Ciona should be lacking the P450s required for steroid and vitamin D metabolism. This slide [slide 9] shows a tree of Ciona intestinalis CYPs along with some from Ciona savigny. Most are in the CYP2 clan. There are 7 clans present. The CYP19 Clan, the CYP39 clan and CYP51 are missing.

Amphioxus is the next organism after Ciona with a genome sequence available. Biochemical assays show that amphioxus has estrogen, therefore, it must have this pathway [slide 10] to make estrogen. That includes CYP11A, CYP17 and CYP19. This tree of amphioxus CYPs [slide 11] shows that CYP17 and CYP19 are clearly present, but these sequences are not very sequence similar to their vertebrate orthologs. CYP19 is 37-38% identical to zebrafish CYP19s. CYP17 is 35% identical to Xenopus CYP17 and 36% identical to CYP1A6 and CYP1A7 from Xenopus. In contrast, CYP19 of zebrafish is 51% identical to human CYP19. CYP17 from fish is 48% identical to human CYP17, so these early CYP17 and CYP19 sequences are barely recognizable as belonging in these families.

Also present in amphioxus but missing in Ciona are CYP51, CYP39 and CYP46, both involved in cholesterol metabolism in vertebrate brain. There are two CYP7 sequences but no CYP8, so prostacyclin synthase may be absent in amphioxus. CYP20 is here and there are two, but not three CYP26 sequences. There is no obvious CYP1 sequence in amphioxus. Instead the CYP17 sequence is the best match to CYP1. This suggests that the true CYP1 we know from vertebrates evolved after amphioxus developed the first CYP17 sequence and steroid biosynthesis. We may be able to see CYP1 in the lamprey genome that has been marked for sequencing. The CYP1 related sequences in Ciona are probably the precursors to CYP17, but they must have been doing something different in Ciona than vertebrate CYP1s.

The other CYP2 clan members in amphioxus are most similar to CYP2U1. This may indicate an ancient and important role for CYP2U1. A paper last year from the company Cytochroma identified CYP2U1 as a novel human

thymus- and brain-specific cytochrome P450, catalyzing omega- and (omega-1)-hydroxylation of arachidonic acid. The authors suggested it had an important physiological role in fatty acid signaling.

This tree does not include the mitochondrial clan which includes CYP11, CYP24 and CYP27 sequences. These enzymes are essential for the first step in steroid biosynthesis and for vitamin D metabolism. [slide 12]. A vitamin D receptor has been cloned from lamprey, but it was not found in Ciona. Amphioxus may be the point of origin of these critical signaling pathways. Four genes of the amphioxus mitochondrial clan are shown here [slide 13] with representative sequences of vertebrates. Two sequences cluster with CYP24s. One sorts with the CYP11s, and one with CYP27s. There are three more mitochondrial clan CYPs in amphioxus, but placing them all in the same tree causes them to cluster together and this obscures their relationships to the vertebrate sequences. There is one more CYP24-like sequence, and two more CYP11-like sequences. One of these is shown on a separate tree. [slide 14] These CYP11s probably include the first side chain cleavage enzyme, though it is not very clear which of the three this is. The others are probable early vitamin D metabolizing enzymes.

What is in common between deuterostomes like Ciona and other chordates and protostomes like Drosophila and C. elegans? We can see [slide 15] here that four CYP clans are present in all the bilateral animal genomes. These are the CYP2, 3, 4 and mitochondrial clans. We will soon see that CYP20 was in the common ancestor and CYP51 also had to be there. That makes six P450 clans. The protostomes are simpler because they have lost CYP51 and CYP20. [slide 16] this tree of all the Drosophila CYPs from two species shows only the CYP2, CYP3, CYP4 and mito clans.

The question then becomes, when did these four clans arise. We can address that question by examining the hydra genome. Hydra is a radial animal or Cnidarian, a group that includes coral, jellyfish and sea anemones. I have analyzed the CYPs in hydra and found a very clear CYP20 gene. This makes CYP20 one of the oldest animal CYPs. Remember, earlier I said CYP38 from a sponge is also a CYP20 ortholog, so this gene predated all known animal species and we will have to look for it in the choanoflagellate genome of *Monosiga ovata* that is targeted for sequencing by the National Human Genome Research Institute. Choanoflagellates are thought to be the precursors of all animals.

The hydra genome is available as about 9 million sequence reads with mate pairs taken from inserts about 3.5 kb in size. It is rare to find two exons on a single contig, so the introns in hydra genes seem to be larger than 3.5 kb. This makes assembly difficult for genes with multiple exons. However, it is possible to find nearly every exon since there is about 5-6 fold coverage of the genome. This slide [slide 17] shows a summary of the genes and unlinked exons for CYP2 clan members in hydra. There are 13 CYPs predicted, six complete, the rest as single exons that I have not been able to join as mate pairs or by walking the chromosomes. Comparison to CYP2 genes shows that no intron boundaries are shared with vertebrates and there are fewer introns in the hydra genes.

This slide [slide 18] shows a similar exon diagram for the CYP4 clan members. There are 17 genes predicted with 9 complete. Here we see that two of the five intron boundaries are conserved with CYP4T, CYP4F or CYP4V. The highly conserved ETAM exon boundary is not present in the Hydra CYP4 clan genes.

[slide 19] This tree shows all the complete hydra genes and some large C-terminal exons from the CYP2 clan genes. Some fish and insect CYPs are included. The CYP3 clan sequences are shown to emphasize that there are no CYP3 clan genes in hydra. There are also no mitochondrial clan genes. These two clans arose in the bilateral animal lineage after it separated from the radial animals. The mitochondrial clan did not come with the mitochondria but derived from a mistargeting event sending a microsomal P450 into the mitochondria where it adopted an existing electron transport system.

The tree shows that the CYP4 clan sequences are most similar to CYP4V and CYP4C. The CYP2 clan sequences are most like CYP17, suggesting that CYP17 has retained more of the ancestral CYP2 clan sequence character than other P450s.

[slide 20] On a phylogeny of animals, we can review some of CYP evolutionary history. The CYP51, CYP2, CYP4 and CYP20 clans are very ancient and were in the common ancestor to radial and bilateral animals. The CYP3 and mitochondrial clans arose in the bilateral animals alone. Many organisms have lost CYP51, including hydra, probably all protostomes, and Ciona. These organisms must eat their sterols. Blocking sterol uptake would be toxic to these organisms. Protostomes also lost

CYP20. Since CYP20 is present in all other animals, the loss of such a conserved enzyme may be linked to some of the differences in development between protostomes and deuterostomes. Early in the chordate line CYP7 arose. Between Ciona and amphioxus, vitamin D receptors and vitamin D metabolism evolved and steroid biosynthesis evolved. It is still a mystery from which clan the CYP19 gene derived, though it shares a similar three step oxidation mechanism with CYP17, which precedes it in the pathway to estrogen. CYP19 may have evolved very rapidly from a CYP17 gene duplication.

[slide 21] We will have to wait a few more months to find out what CYPs are present in sponges. At the opposite end of the time scale, the opossum genome should help to clarify the evolution of the seven mammalian CYP gene clusters. These studies begin to outline the framework of animal CYP evolution, but there is much still to uncover and functional studies are needed to tie the innovations in sequences to innovations in biochemistry and physiology.

Graph of seq traces
http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=graph&f=1&m=stat&s=graph

MUS MUSCULUS	mouse	98,207,074
MONODELPHIS DOMESTICA	opossum	41,719,216
HOMO SAPIENS	human	40,603,046
RATTUS NORVEGICUS	rat	40,170,471
PAN TROGLODYTES	chimp	37,748,666
CANIS FAMILIARIS	dog	36,348,041
ORNITHORHYNCHUS ANATINUS	duckbilled platypus	27,418,699
BOS TAURUS	cattle	25,912,263
MACACA MULATTA	rhesus monkey	23,694,106
DANIO RERIO	zebrafish	23,253,722
XENOPUS TROPICALIS	frog	22,665,141
AEDES AEGYPTI	mosquito	15,481,684
GALLUS GALLUS	chicken	14,735,863
ECHINOPS TELFAIRI	Madagascar hedgehog	10,338,518
SCHMIDTEA MEDITERRANEA	flatworm	10,227,522
LOXODONTA AFRICANA	elephant	10,108,334
STRONGYLOCENTROTUS PURPURATUS	purple sea urchin	9,377,212
DASYPUS NOVEMCINCTUS	nine-banded armadillo	9,234,001
ORYCTOLAGUS CUNICULUS	rabbit	9,173,426
BRANCHIOSTOMA FLORIDAE	amphioxus	8,929,324
FELIS CATUS	cat	8,096,929
CAVIA PORCELLUS	guinea pig	7,435,247
NEMATOSTELLA VECTENSIS	starlet sea anemone	6,476,762
CIONA INTESTINALIS	sea squirt	5,815,065
ANOPHELES GAMBIAE	mosquito	5,396,497
CIONA SAVIGNYI	sea squirt	4,745,578
HYDRA MAGNIPAPILLATA	hydra	4,060,348
APIS MELLIFERA	honeybee	3,827,673
TAKIFUGU RUBRIPES	fugu	3,639,800
TETRAODON NIGROVIRIDIS	freshwater pufferfish	2,974,016
LOTTIA GIGANTEA	owl limpet (gastropod)	2,598,996
CAENORHABDITIS BRIGGSAE	nematode	2,357,666
TRIBOLIUM CASTANEUM	red flour beetle	1,864,564
DAPHNIA PULEX	freah water crustacean	1,752,306
CAENORHABDITIS REMANEI	nematode	1,705,545
OTOLEMUR GARNETTII	small-eared galago (primate)	1,078,349
Plants		
ORYZA SATIVA (JAPONICA)	rice	7,798,053
POPULUS BALSAMIFERA	poplar tree	7,598,270
ORYZA SATIVA (INDICA)	rice	6,423,393
ZEA MAYS	maize	3,684,496
VOLVOX CARTERI	colonial algae	2,929,783
PHYSCOMITRELLA PATENS	moss	1,684,248
MEDICAGO TRUNCATULA	barrel medic	1,114,626
Other		
EMILIANIA HUXLEYI CCMP1516	haptophyte protist	3,868,934
PHYTOPHTHORA SOJAE	stramenopile	1,533,511
TRICHOMONAS VAGINALIS	Parabasalid	1,403,503
PUCCINIA GRAMINIS	fungi	1,043,499
THALASSIOSIRA PSEUDONANA	diatom	1,200,119
TETRAHYMENA THERMOPHILA	ciliate	1,128,627

http://www.mbl.edu/publications/books/Bigelow/Medusa/medusa_plt8.jpg