EPA talk
Feb. 15, 2005

Evolutionary Genomics of Cytochrome P450 with emphasis on mouse, human and rat.

Title slide (1) Chimp on Darwin looking at a skull (to right) Title on left

My opening slide is a collage (2) containing 35 eukaryotic species that have a genome project, either complete or underway. Since I made this slide, more than 20 species have been added to the list. Advisors to the National Human Genome Research Institute have selected additional organisms of great evolutionary significance to be sequenced. These include the oppossum, an armadillo, the duckbilled platypus, an elephant, a sea lamprey, a snail, a beetle, a hydra, a sponge, a flatworm, an acorn worm, a moss and quite few more. The genomes we have now and the ones in the pipeline provide a treasure trove of data for comparative genomics. However, the genomes must be annotated and the information extracted. Major genome publications cannot accomplish this in detail. Since they generally get one or two paragraphs for significant gene families, they can only give a brief survey of the rich data. It has been my goal to cover the cytochrome P450 gene superfamily in great detail, for all these organisms and more, that do not have a genome project.

In this talk I will start with a history of P450 sequencing, move to a survey of the distribution of P450s in living species and then go on briefly to plants, finishing with human, mouse and rat P450s.

The information I gather is posted on my website and can be accessed by clicking on entries in the table shown in the next two slides (3). The top of the table has animals and plants, while the bottom (4) covers lower eukaryotes including fungi and protists. I am constantly adding new data to these pages. Notice that the dates on this slide are all from 2004 except for the slime mold Dictyostelium. There are over 1000 files on the website. The most recent additions are to the rat P450 collection and I will talk about that later. The P450 site receives between 1.2 and 1.4 million hits per year. These are largely due to indexers like Google constantly indexing the site, but there are a lot of individuals and companies visiting also.

This graph (5) shows the history of P450 sequencing, starting in 1982 with 3 sequences (2B1 and 2B2 rat, CYP101A1) and taking off in 1996 after the yeast genome was completed. The main eukaryotic genomes released are listed above the line. As with the opening slide, I cannot put all the genomes on this graph since there is not enough room. The total named P450 count is now greater than 4500 Since 1988, I have named all of these P450s and I continue that task today.

If we sort the sequences into 5 major groups, (6) we find about 40% are from animals, 30% are from plants. About 12 to 13% each are from fungi and bacteria

and only 5% are from protists, though the protist sequences may be highly interesting from an evolutionary point of view.

Another way to display this information is on a tree of life (7). Archaea have very few P450s so far and at least one of these (CYP119) seems to be a lateral gene transfer. Plants, animals and fungi at the tip of the eukaryotes are well represented. Even bacteria have many sequences, with many more waiting to be named. Note the 163 unnamed sequences from J. Craig Venter's sequencing of bacteria from the Sargasso Sea. His yacht Sorcerer II is now circumnavigating the globe, sampling seawater about every 200 miles. This has not been popular everywhere. Venter has been accused of biopiracy in taking samples off the Galapagos Islands.

These 80 protist sequences are mostly from two Phytophthora species and a diatom shown here. (8) Phytophthora ramorum is the Sudden Oak Death organism that is killing oaks in California. Phytophthora infestans caused the Irish Potato famine in 1846-1850 which led to mass emigration of the Irish to the US. Phytophthora and diatoms belong to the stramenopiles, which is a major branch on the eukaryotic crown group of organisms. Another related branch are the Alveolates that include ciliates and the Apicomplexan parasites like malaria. (9) These are the ciliates Tetrahymena with 48 P450s and Paramecium with 22.

If we look in more detail at the eukaryotic P450s, (10) we see that they are found in almost every branch of the eukaryotic tree. In this figure, a red X indicates the presence of P450s. The Archaezoa are anaerobic and are not expected to have any P450s. The two other branches without P450s just have not been sampled yet. Red species names indicate completed or draft quality genome sequences. Blue means a sequence project is underway. When the malaria genome was sequenced, I was surprised to find no P450s in the genome. This is probably the consequence of a parasitic lifestyle, allowing Plasmodium to streamline and dump uneeded genes. Another alveolate parasite, Toxoplasma gondii does have at least one P450 so this branch has an X based only on one known sequence.

Cytochrome P450s perform many biosynthetic and degradative reactions (11) as shown here for human steroid metabolism, arthropod ecdysone metabolism (12) and plant brassinosteroid metabolism (13). Their substrates are often lipophilic. These enzymes are critical in biosynthesis of many antibiotics, with the Streptomyces species containing large numbers of P450s. Streptomyces avermitilis has 33 P450s and S. coelicolor has 18. Currently, 45 Streptomyces species make at least 174 P450s. The Streptomyces species are responsible for synthesis of about 2/3 of our antibiotics. Other bacterial P450s are involved in synthesis of erythromycin and vancomycin. These drugs are made with P450 dependent hydroxylations and ring closures.

About 30% of all known P450s are from plants. (14, avocado)

Even though plant P450s were relatively late comers with the CYP71A1 sequence from avocado published in 1989, we now have over 1600 named plant P450s. Most of these were from just two species, Arabidopsis and rice (15). On Sept. 21, the poplar genome was released. This is the first tree genome. I just analyzed the cottonwood P450s and named 566 sequences with 293 full length genes. Plant P450s make up about 1% of all the genes in plant genomes. In rice alone there are an estimated 356 functional P450 genes and about 100 pseudogenes. Since there are nearly 275,000 named angiosperm species, There are well over 50 million plant P450s.

That sounds hopelessly complicated, but this large number of individual sequences sorts into 62 CYP families. These can be even further ranked into just 10 plant P450 clans. (16) This slide shows the plant P450 families sorted into clans. These are named for the lowest numbered family in the set. Note that 6 of the 10 clans have only one family. The other four are larger. As an example I show you the CYP86 clan, (17) showing 5 families in different colored blocks. CYP86A1, 94A1, 94A2 and 94A5 are fatty acid hydroxylases most similar to CYP4 in animals and CYP52 in fungi, which are also known fatty acid or alkane hydroxylases.

Work is being done all around the world on characterizing the function of these CYP families. One way to get a handle on this great variety is to compare families across plant taxonomic groups. This slide (18) shows a map of plant P450 sequence space. The 62 plant families are listed across the top sorted by clan and the verticle axis shows plant phylogeny. This only includes angiosperms and one line for gymnosperms at the top. It does not include ferns, mosses, liverworts or green algae. Each dot represents the presence of that CYP family in that plant taxon. The three intense lines are rice (a monocot) at the top, cottonwood at Malpghiales, and Arabidopsis (in the dicots) at Brassicales. These genomes are complete. The others are only partial samplings by ESTs or genome survey sequencing. Fabales includes soybeans with 350,000 ESTs and Solanales includes tomato and potato each with more than 150,000 ESTs.

By comparing the dots between taxa, we find 39 of 59 families are shared between rice and Arabidopsis. 43 families are shared between rice and cottonwood. That is nearly 3/4 of CYP families existed in the common ancestor of monocots and dicots about 200 MYA. There is a lot of white space in this figure. We really need more genomes sequenced to sample the space more fully. As with bacteria, plant P450s are at the heart of human interest. Single P450 enzymes are responsible for saving and destroying human lives. I have marked two P450 families in red in this figure. The blue boxes show P450 families that are only found in one taxa so far. The CYP80 family is found only in Ranunculales. These P450s include an essential step in the synthesis of morphine in the opium poppy, a P450 responsible for the heroin trade. The other red P450 is from the yew plant. CYP725 is one enzyme required in the synthesis of taxol, the anti-cancer drug that has saved thousands of lives.

Gymnosperms (especially pine trees) have only been sampled by EST sequencing, but already 19/62 of the plant P450 families and all 10 of the plant P450 clans are found in these ESTs. The divergence date for gymnosperms and angiosperms is 360MYA, so a great deal of plant P450 biochemistry was already in existence that long ago. The green algae Chlamydomonas has only three of these 62 families and 12 new ones. CYP97 is a carotenoid hydroxylase, CYP51 is the sterol demythylase and CYP710 is of unknown function. Once more, we need more sequences like moss, liverworts and ferns to follow the transition from green algae to angiosperms.

You can read about this in more detail in a special Arabidopsis issue of Plant Physiology that came out in June. My collaborators on the rice and Arabidoipsis work were: (19)

CYP20

After naming thousands of P450 genes, a person develops some favorites. One of my favorites is CYP20. This is a gene found in sea squirts, fish and mammals (all chordates). Recently, a P450 sequence was found among Hydra ESTs (20) that had its best match in chordates to CYP20. This sequence matched another sequence in a leech (which is an annelid worm). The leech sequence also appeared to be a CYP20 ortholog. Halocynthia roretzi, which is another sea squirt, has a clear CYP20 ortholog with 51% sequence identity to Ciona. Surprisingly, a complete P450 from a sponge named CYP38A1, is also a best match to CYP20 sequences. (21) The presence in sponges, hydra, annelids and chordates means that CYP20 was in the common ancestor to all animals and it has been preserved over a billon years. A CYP20-like sequence is not detected in insects yet, though the function may be there, the sequence similarity is not.

I have asked several people to try experiments on CYP20 to see if it does anything interesting. I finally found someone who said yes. Todd Penberthy (a former student from this Depratment) was doing a postdoc in a zebrafish lab and was looking for some targets to knockdown with morpholino antisense probes. His lab was working on blood development, so they had a reporter strain of zebrafish with green fluorescent protein under the control of the gata1 promoter. Gata1 is a transcription factor important in erythropoesis. Normal fish would have typical levels of gata1 and they would express the GFP and make normal red cells. Fish with reduced gata1 would not make the normal number of red cells and they would not be green fluorescent.

In fact, a mutant in gata1 produces a bloodless phenotype in zebrafish called the vlad tepes mutant. Todd injected embryos with the morpholino antisense probe for zebrafish CYP20 or a sham injection with buffer and this is the result. (22) The antisense morpholinos greatly reduced the level of GFP in the treated embryos but not in the sham injected controls. Blood development was not turned off but it was reduced. The most dramatic effect was in the GFP expression in the eye and the lower part of the head.

This result suggests that CYP20 is an upstream regulator of Gata1, a critical transcription factor in erythropoesis and possibly in the development of the eye. We expect CYP20 is acting on a lipid substrate to make or degrade a signaling molecule. This molecule may then bind a receptor protein and alter gata1 expression.

Gata1 comes up again in an independent context. (23) To tell you about this I have to talk about mice. The mice offer a very unique method to look at P450s that is not available anywhere else yet.

We are going to talk about recombinant inbred strains of mice and quantitative trait mapping. Eldon Geisert last week presented the logic behind recombinant inbred mice. They are homozygous strains that scramble the parent strains genomic contribution. Traits that vary between the two parent strains can be mapped in some detail. This includes gene expression levels. Rob Williams here at UT Memphis has created a web site called WebQTL (24) that allows one to enter gene names for about 8500 different mouse genes represented on Affymetrix mouse chips. The software will analyze microarray data from about 30 different recombinant inbred mouse strains. Each line on this pull down menu is a separate experiment. The result is a QTL map showing the strength of a relationship between the expression levels of your gene and about 700 intervals defined by genotyped markers spread out over the whole genome. If a gene is significantly contributing to the variation in expression levels of your gene on the microarray chip, it will show up as a spike in the QTL interval map near the location of the controlling gene. The stronger the influence on expression, the higher the peak on the map.

Rob has done microarray analysis on these 32 inbred strains at 3 chips per mouse strain for dozens of separate experiments. That is about 100 Affymetrix chips per experiment. These data represent thousands of chips. Most of these are from brain, but there are three sets from hematopoietic stem cells, done by others.

I searched for all P450s on these chips by a text search for CYP*. This resulted in 74 hits. After subtracting out cyclophilins and duplicates, there were about 40 mouse P450s on the chips. I ran all the mouse P450s that were present against 24 microarray data sets on the server and found quite a few QTL peaks. The strongest result was for Cyp2b10, and it is shown here. (25). This data is from the three hematopoietic stem cells experiments. The numbers across the top are chromosome numbers. The dotted green line indicates the peak height for a suggestive QTL (p = .5), but it may be accidental. The dotted blue line is a significance threshold (p = .05). Peaks exceeding the blue line are probably real QTLs. These three peaks are all strong, but the third experiment is way above the significance threshold. The interpretation is that there is a gene in this region on the X chromosome that influences the level of Cyp2b10 (or a closely related Cyp2b) in hematopoietic stem cells. So now you must look at a list of the genes in this region.

(26) This list is a partial list of the genes on the mouse X chromosome at the region of the QTL. One gene stands out as a strong candidate gene. Gata1 is a transcription factor required for erythroid differentiation, so it will be active in hematopoietic stem cells. A search of the mouse Cyp2b10 upstream region identifies a gata1 binding site. In fact, there are two of these sites in the human CYP2B6 gene. This does not prove that GATA-1 is the gene responsible for the QTL, but it creates a testable hypothesis. This may provide clues to the endogenous function of Cyp2b, independent of drug metabolism, a possible role in differentiation of erythroid cells. Remeber the CYP20 gene seems to influence gata1 levels and this in turn may influence Cyp2b levels in a kind of P450 cascade.

(27) This is a list of additional QTLs that were significant or appeared multiple times in independent experiments. I have not tried to go in and look for candidate genes for these QTLs. The QTL peaks are mapped only to a 10-15 Mb region, that can cover 150-200 genes. One interesting point is the cis QTL at Cyp8b1. Cis QTLs may result from polymorphisms in the promoter for the gene that differ in the two parent strains of mice. I have also found strong cis QTLs in the mouse Cyp3a genes.

This QTL approach holds promise to identify upstream genes in P450 regulation. Most of the data so far is with brain, but liver is being done now, and that may be more useful for many P450s. The trick is going to be in identifying the best candidate genes and then testing them for effects on P450 gene expression.

The rat, mouse and human genomes.

Having more than just one mammalian genome is very useful for understanding how our genome has evolved. My colleagues Dan Nebert, Darryl Zeldin, Susan Hoffman and two genome nomenclature experts Hester Wain and Lois Maltais and I published a comprehensive comparison of mouse and human P450s one year ago in Pharmacogenetics. This is a tree of the mouse and humand P450s, with the ortholog pairs in red(28). These results showed an expansion of the mouse P450s in seven gene clusters. The mouse-human comparison is published, but I will now show these seven gene clusters including the rat genome. In general, the rat is intermediate in P450 cluster size and complexity. The mouse has gone to extremes in amplifying the P450 genes in these clusters. These figures emphasize the point that a rodent is not a human especially when it comes to drug metabolism.

We will begin with the most straight forward clusters and move to more and more complex clusters. The CYP4F cluster (29) has complete conservation of orthology between mouse and rat, not counting pseudogenes. In all of these figures, each dot equals one exon. In human, the only clear ortholog is CYP4F22 which is 86% identical to 4F39 in rat. This suggests that 4F39 was the ancestral sequence in this cluster and its function probably is the original function. Note that where names were not already assigned I have used the same names for ortholog pairs. (30) This is a diagram showing the CYP2, 3 and 4 clans in humans and fish. Note there is only one 4F sequence in fish, so it is possible that the ancestor of humans and

rodents had only one 4F that has diverged independently for 75 million years. Sequencing a marsupial like the oppossum may help clarify this issue.

(31) This is the Cyp2d locus. There is only one CYP2D6 functional gene in human, though a report has appeared showing a functional CYP2D7 in human brain, possibly due to a recombination event. This is a polymorphism in only a subset of humans. The mouse has 9 Cyp2d sequences, while the rat has 5. The red lines represent orthologous genes or groups of orthologous genes. This is a complication not observed in the 4F cluster. The common ancestor of mouse and rat apprently had 4 CYP2D genes. One duplicated in rats to give five genes. This same gene in mice expanded to 5 genes and several pseudogenes. Another gene in rats is related to two genes in mice separated by 6 pseudogenes. The pseudogenes are not clearly orthologs between mouse and rat except this outer one. Cyp2d22 and CYP2D4 are most similar to CYP2D6 and these may represent the common ancestral sequence. If this is correct, then there may be a pattern emerging that the outside member of a cluster is the ancestor of the cluster, like we saw in the 4F cluster.

(32) This is the Cyp2j locus. There is one human CYP2J2 and 8 mouse Cyp2js. The April 2004 Nature issue shows 6 in the rat, but I only found 5. Perhaps they are counting one of the pseudogenes. Once again there is strong orthology between mouse and rat, with some expansion in the mouse. Note that 2j5 in the mouse is a pseudogene in the rat. The six 2js in rodents dont have a clear ortholog to human. Another intermediate species might clarify which rodent sequence is the original 2J.

The next slide (33) shows the CYP2ABFGST cluster. This is the most complex P450 cluster in mammals with six subfamilies present. All six subfamilies were in the common ancestor, so this may be the oldest P450 gene cluster in mammals. All three species have CYP2S1 one one side and CYP2T on the opposite side. This continues the pattern that the edges of clusters are well conserved. The human cluster has undergone some rearrangements so it is not easily compared to the rodent clusters except on the CYP2T end. The mouse and rat have strong similarity with clear orthologs all across the cluster (shown by the red lines). Cyp2a4 is a duplication in mouse not seen in rat. The Cyp2b19 gene in mouse may be equivalent to 2B12/15 in rat. The human cluster seems to have a mirror symmetry indicating a possible inverted duplication event that makes it different from the rodent pattern.

The CYP2C cluster (34) is small in humans (only 4 genes) very large in mice (15 genes) and intermediate in rats (10-11 genes). The orthology across the cluster is broken by an inversion shown by the crossed red lines. Mouse has expanded several of the boxed gene subclusters. There is no obvious orthology to the human genes. Note that on the right side there is a 4.1Mb space between 2C22 and 2C23. This is conserved in mouse. Also interesting is the orthology between 2C11 and a pseudogene 2c52-ps in mouse, which has one stop codon, two frameshifts and a 4 aa deletion at the second frameshift. This must be a very recent pseudogene to retain 82% sequence identity. The genes flanking this CYP cluster are the same in all three species.

The 4ABXZ cluster (35) has some interesting features. There is no 4Z gene in the rodents. In fact, this gene is only seen in humans and chimpanzees, making it a rare hominid specific gene. I compared the human and chimp clusters and they are highly similar. Three of the four pseudogenes in the middle are 100% identical and the fourth has only 1 aa difference. This preservation has lasted 6 million years. By the time 17 million years have passed, very few pseudogenes are presereved in mouse and rat. There are some expansions in mouse and some in rat. Oddly, there are expansions in the human that are outside the cluster boundaries and we have not seen that before. The 4ABX cluster seems to arise from the fish 4T sequence(36). This is also shown on the fish human comparison diagram for CYP4 (37).

The CYP3A cluster (38) is the last of the seven CYP clusters. The CYP3A9 gene is the ortholog of Cyp3a13 in mouse. These genes are outside the main cluster by 1 million bp in mouse and 7.5 million bp in rat. 3A1 and 3A62 in the rat are not in the rat genome sequence yet, but 3A1 may be in a sequence gap found in the cluster. 3A62 is related to 3A9 and may be closer to it. There is expansion in the mouse and some duplications are moved to new locations as with the 3a57 gene that is clearly related to 3A18. The human genes do not have clear orthologs to the rodent genes, suggesting a single ancestral sequence evolved independently in human and rodent lines.

I end with my opening slide (39) which reminds us that these are living organisms, not just black boxes with so many Mb of DNA. It is life, represented by these beautiful creatures that we, as biologists, are attempting to understand.