Life through a cytochrome P450 lens.
David Nelson Nov. 20, 2008 Rhodes College, Memphis

This seminar resulted from a chance meeting in Edinburgh this past April. I was there for a workshop on *Aspergillus nidulans* genome annotation. The Aspergillus meeting followed that and it was there that I was introduced to Terry, Darlene and Sarah by a friend since he knew we were all from Memphis. Science is like that. You can travel 3000 miles to meet scientists who work only 5 miles from you.

Today I will tell you about the cytochrome P450 superfamily. I am in a special and privileged position in P450. I am the curator of the nomenclature for this gene family. This came about because the first paper I published as a post-doc was on cytochrome P450 evolution based on 34 sequences that were known in 1987. That same year another group proposed a nomenclature system for cytochrome P450 based on evolution using the same sequences. My paper attracted their attention and I was asked to join the nomenclature committee. Over the years I have become the nomenclature committee. In 1995 I created a website for the P450 nomenclature and this includes species pages for all the P450s from a given species. As the genome projects began to roll in during the late 1990s this nomenclature service began to grow, until now there are over 9000 P450 sequences that I have named. [slide 2] Plants have the largest number with over 3500. Animals are next with about 3000. Bacteria and fungi are in the 1000 range. Notice that Archaea have very few, probably because most of these are anerobes and P450s are mostly oxygen utilizing enzymes. I anticipate we should reach 10,000 named P450s sometime in 2009.

There were 481 P450 sequences known in the fall of 1995 [slide 3]. This was the year of the first genome sequence of the bacterium *Haemophilus influenza* and this was followed the next year by the first eukaryote genome from baker's yeast. Genomics began with these two genomes and the discovery of P450 gene sequences accelerated with the rapid increase in sequencing capacity worldwide. You can see the progress as the genomes became public. We are now off the top of the chart. Currently there are over 2400 genome sequencing projects underway with 371 of those in eukaryotes[slide 4]. I have prepared a two minute slideshow featuring 53 of the 192 eukaryotic genomes that have been sequenced or are in draft assembly, with emphasis on animals and plants [slides 5-44]. So lets just take a couple of minutes and enjoy this brief parade of organisms. Some are model organisms like the mouse (4) and the rat (5) or they are valuable economically like cattle (8) or chickens (10) Here we have several fish: pufferfish (11), zebrafish (12) and stickleback (13). Some are evolutionary mileposts in the history of life, like amphioxus (15) and choanoflagellates (18). Some are significant pests of humans like these mosquitos (20) or they are pests of crops like the red flour beetle (23). Remember your high school biology as we see Daphnia (26) and the hydra (28). Then we have our crop plants: rice (32), grapes (34) papaya (35) and maize (36). And now we finish with some filamentous fungi (38) and a few protists (39, 40).

There is not enough time to discuss 9000 genes, but I would like to organize my talk around three major groups: bacteria, plants and fungi. There is not enough time today to talk about animal P450s.

First I will give a brief introduction to cytochrome P450. All cytochrome P450 crystal structures look pretty much like this one from rabbit [slide 45]. They are triangular in shape with a helix rich region and a beta sheet rich region. They all have a very long helix that spans the whole length of the molecule called the I-helix. In the middle is the heme. The substrate binding site is above the heme as it is shown here and below the I-helix. Below the heme is a cysteine that is negatively charged. This thiolate anion interacts directly with the heme iron and it is this thiolate group that gives cytochrome P450 its strong absorption peak at 450 nm. This is where the name comes from: pigment at 450 nm. The first P450 spectrum was published in 1958 so this is the 50th anniversary year for P450.

Eukaryotic P450s are membrane bound with a few exceptions. The exceptions appear to be lateral transfers from bacteria, which are soluble proteins. The membrane bound forms have nearly the same shape as the soluble forms. The main difference is a membrane anchoring peptide at the N-terminal [slide 46] and some hydrophobic patches on the surface of the molecule. In animals most P450s are in the ER membrane where they receive electrons from NADPH through a special reductase molecule called NADPH cytochrome P450 reductase. A smaller group of P450s is in the mitochondria where the electron transfer chain is different [slide 47]. All the mitochondrial P450s form a single clade on phylogenetic trees so they all arose from a single ancestor that was accidentally mistargeted to the mitochondrion. This only occurred once in the history of animals.

P450s do oxidation-reduction reactions. The most common is hydroxylation, but they can do many other types of reactions like demethylations and desaturations. These reactions require oxygen, but some substrates such as peroxides bring oxygen with them, so these unusual substrates do not need molecular oxygen in the reaction. P450s are named according to their clustering on phylogenetic trees. [slide 48] The names all begin with CYP for cytochrome P450 and this is followed by a number that is the family number. Members of the same family are generally 40% identical or higher in amino acid sequence. A letter follows the family number and this is the subfamily (>55% identical). The final number is a sequential sequence number given as the sequences are discovered. Blocks of family numbers are reserved for different taxa such as plants animals, prokaryotes and fungi and other lower eukaryotes.

Bacteria

With that as background, let's look at some bacterial P450s. [slide 49] Bacterial P450s are often used for metabolism of unique carbon sources, or to make antibiotics and other secondary metabolites. This is just a sample of the diversity of molecules bacteria make or modify by P450s. CYP family names begin at CYP101. The current set of 914 named bacterial P450s sort into 207 families and 61% of these families have only 1 or 2 sequences. About every third bacterial P450 discovered is in a new CYP family, so we are not saturating the sequence space of bacterial P450s yet. Bacterial P450 sequencing has been driven by the quest for understanding how bacteria metabolize complex molecules like camphor or the explosive RDX. P450s are also crucial to making many antibiotics or pharmaceuticals like erythromycin, vancomycin and daunorubicin. These are molecule or pathway driven research projects, however, sampling of biodiversity just for the sake of biodiversity has recently become popular with many metagenome projects underway. [slide 50] The largest is J. Craig Venter's Global Ocean Sampling project. Venter sent his yacht around the world to collect seawater samples at the locations indicated on this map. The chosen path was inspired by the journey of the H.M.S. Beagle that Darwin traveled on around the world. Venter's crew filtered the water to select eukaryote sized, prokaryote sized or virus-sized particles and cloned and sequenced the DNA from these samples. The first leg of this trip (orange dots plus Sargasso Sea) netted 7.7 million sequences mostly from bacteria and doubled the number of novel protein sequences in Genbank. The red dots in the Indian Ocean represent another 2.4 million sequences released in March 2008.

One result of this analysis was a view that protein biodiversity is very great and even though millions of new sequences were found there is a linear relationship of sequence families vs. number of sequences [slide 51]. It is not saturating. This slide shows 45,000 protein families with 10 members or more.

I contacted The JCVI and Shibu Yooseph sent me the 3305 P450s from the first leg of the project. I sent him 519 named public bacterial P450s that were not pseudogenes and he added them to the seawater sequences and ran them through a clustering algorithm called CD-Hit. These sequences were clustered at the 90% identity level. At this high percentage identity the named P450s did not cluster with the seawater sequences. A representative sequence from each of 2211 clusters was used in another round clustered at 60% identity. This produced 1143 clusters, which included 326 named P450s. Shibu used these to make a sequence alignment using the program Muscle, then he removed sequences that were less than 60% of the alignment length. The last step was to make a NJ tree with 566 remaining sequences, which is shown here [slide 52]. The red branches are seawater sequences and the blue branches are named P450s. The red Ss label seawater specific clades and there are only three of these. The named blue P450s intermingle with the seawater sequences over most of the tree. This means that the ocean sequences are not in a separate universe from the named sequences, even though most of the named sequences come from non-marine sources. The Blue E labels the eukaryotic-like branch of bacterial P450s that cluster with eukaryotic P450s on mixed trees. LB1 and LB2 (for long branch) are from partial N-terminal sequences that should have been deleted from the data set.

The question now is what happens when these red sequences are named and new sequences are found, will there be some sign of saturation or will one third of the new sequences still fall into new families? We need to wait for the sequences from the next leg of the voyage before we can know that.

Plants

Now lets move on to plants [slide 53]. The first P450 to be sequenced from plants was a fruit ripening mRNA from avocado named CYP71A1. Plants have reserved CYP family names from CYP71 to CYP99 and then three digit names from CYP701 to CYP999. The sampling of P450s is much better in plants [slide 54]. This figure that was made in April shows the P450 abundance by family among the 95 plant CYP families. The brown bar represents seed plants and the blue bar represents moss and green algae combined. If we exclude algae and moss, there are 61 seed plant families with from 7 to 470 members each. This figure was made before the grape and Selaginella lycopod moss genomes were analyzed. Nevertheless, all of the seed plant families have multiple sequences. This spring I named all the grape genome CYPs. There were 315 genes and 201 pseudogenes for more then 500 P450 sequences. They all fit in existing CYP families. Papaya had 142 CYP genes and they also fit in existing families. No new angiosperm plant CYP families have been discovered in the past four years, so we are seeing saturation among flowering plants.

This is not true in the lycopod Selaginella, moss or the green algae Chlamydomonas. *Selaginella mollendorffii* has been sequenced to 20X coverage and both haplotypes of the heterozygous genome have been assembled. There are 225 P450s not counting the pseudogenes. There are 24 new CYP families in Selaginella. The moss *Physcomitrella patens* only has 71 P450s, but it has 15 novel CYP families. Chlamydomonas has 39 total CYPs but 18 novel families. The family distribution among 8 completed plant genomes is shown here [slide 55].

The CYP families are grouped according to clan and then by occurrence in the 8 species. A P450 clan is a clade of P450 families that always group together. Seed plants have 61 families in 10 clans. This arrangement highlights groups of P450s that are common to different taxa. The left edge of the figure shows three families present in all plants. These are the only three that are so conserved and they are doing essential plant functions of making sterols and carotenoids. The orange box includes 12 CYP families in 6 clans that are found in all land plants. These were crucial genes for adaptation to life on land. These gene products make a waxy water barrier to prevent water loss. They make flavonoids for protection from UV light, and they make basic plant hormones like gibberellins and jasmonates. One of them makes a tough polymer called sporopollenin to protect pollen grains. As we move up on this graph, we find more recent P450 families that evolved in more limited groups of plants. The yellow boxes cover 43 families in 10 clans that are common to angiosperms. The white boxes enclose 51 CYP families in common among eudicots. You will notice that Arabidopsis has some holes where some P450 families are missing. Arabidopsis was partly chosen as the first plant genome to be sequenced because it has a small genome. This has clear advantages, but it also meant that Arabidopsis is perhaps not entirely a representative plant genome.

This data was generated from a much more comprehensive analysis of all the plant P450s shown in the next two slides [slide 56]. I won't go through this in detail, but I will point out that the arrangement of the CYP families against plant phylogeny allows a stairstep pattern to emerge at the top of the figure. This shows the first appearance of each family in the phylogeny on the left and this is related to evolutionary time. The boxes represent taxon specific families. For example, this one is only found in Euphorbia and it makes an epoxy fatty acid called vernolic acid that is commercially important as a high temperature lubricant. [slide 57] This figure shows the remaining CYP families. Coverage is pretty good for the angiosperms, but we really need more data for gymnosperms, and higher. It would be very nice to have a fern genome but they have very large genomes and it costs too much right now to do a fern, but that will come down soon. A liverwort is being done now and that will be quite useful for defining early plant P450 evolution.

The number of CYP genes is different in each plant [slide 58]. Grape is a fruiting woody plant like papaya, but grape has 173 more CYPs than papaya. Arabidopsis and papaya are both in the order Brassicales, but Arabidopsis has 103 more CYPs than papaya, even though it is missing some families that papaya has retained. Each genome undergoes rounds of gene expansion and gene loss. I will show you [slide 59] a dramatic increase in the CYP82 family, which has jumped from two genes in papaya to 34 genes in the grape. It is tempting to speculate that these gene family expansions have been brought about by human selective pressure over several thousand years and these genes are critical to the properties that humans have desired in wine.

The last segment of my talk is on fungi. [slide 60].

Fungi

There have been many fungal genomes sequenced recently, especially filamentous fungi like Aspergillus. These are important to human health and agriculture. Fungal P450 diversity is very great, similar to bacterial CYP diversity. [slide 61] This slide shows the abundance of CYPs in 329 fungal CYP families. Note as in bacteria that 60% of the families have only one or two sequences. We have a long way to go to reach saturation of P450 families in the fungi.

Fungal P450 diversity is quite high, but if one limits the discussion to a subset of filamentous ascomycetes including Aspergillus and Fusarium the diversity should drop. [slide 62 fungal diversity 17 species A.] This chart shows the P450 family distribution among 17 species, including *Aspergillus niger* that I recently annotated. A. niger has 152 P450s with 19 new families. The set of 8 rows beginning with N. crassa are from complete genomes. The upper 9 rows are from incomplete genomes.

The sequences are displayed in CYP name order. The older sequences have the lower CYP numbers and they tend to be found in more species. One peculiar feature of this layout is the blocks of solid red in some rows. This occurs because a whole genome's collection of P450s are usually named together. The block of sequence families without any gaps represents the historical naming order and these were all in new CYP families at the time they were named. In the case of N. crassa, 31/39 P450 families were new when the genome was annotated. Magnaporthe grisea was next with 46/74 families new. This is followed by Fus. Gram. [slide 63 fungal diversity 17 species B.] The second part of this chart continues with four more genomes. As we move to higher CYP family numbers the sequences become less abundant in other species, since the most prolific P450 families have already been found. You would think the number of new families would decrease over time, but *Aspergillus niger*, the fourth Aspergillus genome still had 19/87 new families. [slide 64 new vs. old families] This is summarized in this graph with the blue indicating previously named P450 families and red are the new families in each genome. Asp. fumigatus looks like an exception, but this is due to sharing 47 CYP families with Asp. nidulans.

The next [slide 65 shows the top 52 P450 families] from this data set. The families have been sorted by number of species present per family. Note that only 8 of 252 families are found in all 8 species. Two of these are the ergosterol biosynthesis P450s CYP51 and CYP61. They are required by all fungi. The fission yeast S. pombe has only these two CYPs. The other abundant P450s are carbon source utilization enzymes or general toxin biosynthesis P450s.

About half of the CYP families in these 8 filamentous fungi have only one member. This is very different from what we saw in plants or especially in vertebrates, where every vertebrate has the same set of P450 families and only the subfamilies differ.

I illustrate this by a car analogy. [slide 66 vert. cars] Vertebrates are very similar while [slide 67 fungal cars] fungi are quite novel except for the basic CYP51 and CYP61. Everything else can vary.

One of the most interesting discoveries to come from genome sequencing in fungi has been the observation that fungal secondary metabolites are often made by gene clusters. These clusters have all or most of the genes required to synthesize a toxin or pigment. There is a tendency for these clusters to occur near the telomeres. I got involved in looking at these clusters because they often include cytochrome P450 genes.

As an example I show you the gliotoxin gene cluster [slide 68]. This cluster is related to the sirodesmin gene cluster in *Leptosphaeria maculans* as shown in this figure from a 2005 paper by Gardiner, Waring and Howlett. P450s are labelled. These genes are color-coded to show orthologs. Note that CYP613B1 gliC is a presumed ortholog to CYP5082B1 sirC but they are only 34% identical. This is an example of a place where the nomenclature for P450s has failed to identify the ortholog when the genes were named. When examining many fungal metabolite gene clusters there is a fairly diagnostic set of gene types that are associated with these clusters and by looking for their occurrence in a group you can find most of these clusters easily.

[slide 69] Here is a list of the usual suspects in fungal secondary metabolite clusters. We found that Aspergillus nidulans has 111 CYP genes plus 8 pseudogenes [slide 70].
A search of A. nidulans for P450s 7 or fewer genes apart found 13 P450 gene clusters [slide 71]. Nine of these were near a PKS or NRPS gene. If the search is expanded to find a single P450 near a PKS or NRPS gene then 15 more candidate clusters are found. There are more gene clusters that do not contain any P450s. Further searching near PKS

or NRPS genes identifies 11 more potential secondary metabolite clusters without CYP genes. This is a total of 39 secondary metabolite clusters in A. nidulans. A similar search was conducted by Nancy Keller's lab in A. fumigatus and 22 potential gene clusters were found. In A. nidulans 28% of the P450s are in these clusters and in A. fumigatus it is 24%, so understanding the functions of these clusters will contribute to understanding the functions of the P450s. A bioinformatics search for these associations should be feasible on any fungal genome. In fact I am mentoring a student from the University of Memphis who is writing an algorithm to do this. Once these clusters are identified then it becomes a question of matching them to the known secondary metabolites made by that organism. There is also great potential for comparative genomics.

I have one example to show [slide 72] comparing a cluster from A. nidulans with six other fungal genomes. The P450s are shown as filled arrows. A. nidulans and A. clavatus both have four CYPs and they are in the same subfamilies and in the same gene order and orientation. Looking at the other genes in these two clusters it is clear they are orthologous, with some variation at the left-hand side. Note that the PKS genes are not the same size and the MFS gene has moved.

The feature that unites all these clusters is the terpene cyclase gene. This gene is aristolochene synthase in A. terreus, but the rest of the cluster is not sequenced in A. terreus. These genes are farnesylpyrophosphate cyclases that form sesquiterpene rings. The A. nidulans cluster also has a PKS gene. This cluster is probably joining a sesquiterpene ring to a PKS product and decorating the whole by four P450 oxidations.

The A. oryzae and A flavus clusters are nearly identical to each other and they have the terpene cyclase gene and three P450s. Two of the P450s are orthologs from the upper two clusters, one is new. These clusters seem to be doing something new but still based on an oxidized sesquiterpene ring.

The last two clusters have the terpene cyclase and some of the same P450s as the upper clusters, but the structure of the cluster is pretty different and these are probably making different products, variations on a theme.

Analysis of the genes in the cluster may give a pretty solid clue about the product. [slide 73] This is part of the A. nidulans P450 list in blocks of chromosome order. The predicted gene clusters are shaded yellow or green. Notice the large yellow cluster near the bottom. There are five P450s in a span of 16 genes. Normally we would not have a clue what these P450s were doing, but by looking at the other genes in the cluster we see a PKS gene, an NRPS gene, a flavin monooxygenase, an efflux pump and the best clue is a dimethylallytryptophan synthase-like gene. These DMAT synthase genes catalyze the first step in ergot alkaloid biosynthesis by prenylating tryptophan with dimethylallypyrophosphate from the mevalonic acid pathway. It is highly probable that this 17-20 gene cluster is making an alkaloid and the P450s are probably decorating the structure with hydroxyl groups. A brief search of the literature turned up this compound called terrequinone A an alkaloid made by A. nidulans and this structure has a prenylated tryptophan ring. This gene cluster may be responsible for making this compound. A knockout experiment could fairly quickly verify or rule out this prediction by looking for loss of the alkaloid in the knockout strain.

That concludes our survey of cytochrome P450 in bacteria, plants and fungi. We did not have time to talk about animal P450s today [slide 74], perhaps in another visit.