Life through a cytochrome P450 lens.
David Nelson June 8, 2008 Nice, France

Once again we are gathered to spend five days talking about our favorite protein family on earth: cytochrome P450. In July 2006 we met in Wales where the P450 gene count was 6051. [slide 2] I am happy to report that the named P450 count is now 8812 with over 2700 new P450 sequences added in the past two years. At this pace we should hit 10,000 named P450s sometime in 2009. [slide 3] Plants experienced the largest increase with the addition of genomes from [slide 4] *Medicago truncatula*, papaya, *Selaginella mollendorffii*, a lycopod moss and grapevine. [slide 5] Animals saw the first crustacean genome in *Daphnia pulex* and the jewel wasp genome *Nasonia vitripennis* was sequenced. Fungi added 152 CYPs from *Aspergillus niger* and 149 P450s from *Phanerochaete chrysosporium*, the white rot fungi, were all finally named.

For this introductory talk I have chosen three areas as a sample of P450 biodiversity. These are bacteria, plants and fungi. There are many animal genomes now sequenced with hundreds of P450s that are waiting to be named and we will hear more about those tomorrow morning from Jed Goldstone, so I won't talk much about animals today.

Bacteria

[slide 6] 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. [slide 7] The current set of named bacterial P450s sort into 205 families and we can see that 61% of these families have only 1 or 2 sequences. About every third P450 discovered is in a new CYP family, so we are not saturating the sequence space of bacterial P450s yet. Note the CYP105 and CYP107 families, the two mot abundant bacterial families. We will see them again. Sampling biodiversity at the gene level has been a byproduct of genome sequencing. Whatever genomes were chosen to sequence was usually driven by medical, agricultural or evolutionary importance. However, sampling of biodiversity just for the sake of biodiversity has become popular with many metagenome projects underway. [slide 8] 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. He 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) released 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 [slide 9] and even though millions of new sequences were found there is a linear relationship of sequence families vs. number of sequences. It is not saturating.

I contacted The JCVI and [slide 10] 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 11]. 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. The 105/107 blue cluster includes CYP105 and CYP107 sequences and closely related families. These are very abundant among the named sequences. The Blue E labels the eukaryotic-like branch of bacterial P450s that includes the CYP51, CYP102 and CYP110 families. LB1 and LB2 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?

Plants

The situation is better in plants [slide 12]. This figure 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. This figure was made before the grape and Selaginella 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 families. Papaya had 142 CYP genes and they also fit in existing families. So we are seeing saturation among the seed plants.

This is not true in Selaginella, moss or Chlamydomonas. *Selaginella mollendorffii* has been sequenced to 20X coverage and both haplotypes of the heterozygous genome have been assembled. I just finished naming all the CYPs and 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 13].

The CYP families are grouped according to clan and then by occurrence in the 8 species. This arrangement highlights groups of P450s that are common to different taxa. The orange box includes 12 CYP families in 6 clans that are found in all land 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. When this information is

distilled down we get the following [slide 14]. I am sure you recognize your favorite families here. I have listed some of the pathways below each group, but this is not a comprehensive listing.

To close our brief section on plants we should mention the grape genome since were are in France and we will be sampling some grapes soon. The number of CYP genes in grape at 315 is very similar to Populus with 312 and a little less than rice at 336. It is 70 genes more than Arabidopsis and 173 more than papaya, the only other woody fruiting plant we have to compare it against. Grape has expanded some of its CYP families compared to the other known genomes. I will show you three examples. [slide 15]. The CYP75 family that include flavonoid hydroxylases have expanded from a single gene in Arabidopsis and two in papaya to 10 in grape. Flavonoids are involved in the color and flavor of the grape berries. [slide 16] Flavonoid phenolic compounds are the main components of red wine flavor. The CYP76 family has expanded from 3 genes in papaya to 24 in grape. Functions of CYP76 proteins include lauric acid omega hydroxylation and geraniol 10 hydroxylation. [slide 17] The most dramatic increase is in the CYP82 family, which has jumped 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 18].

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 19] 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.

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. In fact, 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 20] Here is a list of the usual suspects in fungal secondary metabolite clusters. A search of A. nidulans for P450s less than 7 genes apart found 13 P450 gene clusters. 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. 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 21] 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.

Animals

This concludes my brief survey of life through a cytochrome P450 lens. [slide 22] My last slide is intended to set the scene for Jed Goldstone's lecture tomorrow on animal P450 evolution. The current set of about 3000 named animal P450s fall into 110 CYP families. The diversity is in between plants and fungi with 38% of the families having one or two sequences. There are multiple as yet un-annotated animal genomes that will hopefully reach the naming stage this summer. Then this figure will completely change. I am looking forward to the meeting and all the new stories. I'll be happy to answer your questions.