The Cytochrome P450 Homepage

*David R. Nelson**

Department of Molecular Sciences, University of Tennessee, 858 Madison Avenue, Memphis, TN 38163, USA
**Correspondence to*: Tel: +1 901 448 8303; Fax: +1 901 448 7360; E-mail: dnelson@uthsc.edu

Date received (in revised form): 21st August 2009

Abstract

The Cytochrome P450 Homepage is a universal resource for nomenclature and sequence information on cytochrome P450 (*CYP*) genes. The site has been in continuous operation since February 1995. Currently, naming information for 11,512 *CYPs* are available on the web pages. The P450 sequences are manually curated by David Nelson, and the nomenclature system conforms to an evolutionary scheme such that members of *CYP* families and subfamilies share common ancestors. The organisation and content of the Homepage are described.

Keywords: *cytochrome P450, annotation, hand curation, protein family databases*

History of the Cytochrome P450 Homepage

The Homepage (http://drnelson.utmem.edu/CytochromeP450.html) grew out of a need for unlimited space to present nomenclature and annotation information on cytochrome P450 sequences. The 1993 cytochrome P450 nomenclature paper[1] was 51 pages long and was the last dedicated P450 nomenclature publication before the website was opened in February 1995. That paper had 12 authors, all well known in the field. Multi-author agreement was the strategy of Daniel W. Nebert, who with Frank J. Gonzalez launched the standardised cytochrome P450 nomenclature system in 1987,[2] with follow-ups in 1989[3] and 1991.[4] The P450 nomenclature prior to this system was fragmented and difficult, with many laboratories having their own shorthand notation for P450s, often based on molecular weight or migration position on gels. For example, *CYP2A1* had six different names in the literature. The new *CYP* nomenclature was one of the first systematic nomenclatures for a protein superfamily and it became adopted for use by many other groups struggling with the rapid expansion of sequence data in the 1980s.

The Cytochrome P450 Homepage was started on a desktop Macintosh Quadra 650, running WebStar as the server software. I believe MOSAIC was the web browser at that time. The 1993 paper had 221 P450 genes and 12 pseudogenes listed.[1] On 10th October, 1995 I gave a talk at the Third International Symposium on Cytochrome P450 Biodiversity in Woods Hole, Massachusetts, titled: '450 Cytochrome P450s', so the number of *CYPs* had doubled in less than two years. I posted a sequence alignment on a wall at that meeting with all the non-confidential P450s. This occupied about a 2×3 metre squared space, and everyone wanted to come to see their own sequences.

Clearly, publications could not include such large alignments, and there was no space to discuss individual sequences. Even 51-page papers could only include long tables to list the genes. A website was the solution. The initial web pages were very simple lists of html code showing entries by *CYP* name in alpha-numeric order. The purpose was to include the 200+ new sequences that were not in the 1993 paper. This was stated at the top of the nomenclature pages as: 'P450s that have appeared since the 1993 P450 nomenclature update'. At that time, entering 200 sequence names, accessions, references and percentage identity information

seemed like a large task. Today, a single plant genome may have over 300 *CYP* genes, and if you include pseudogenes, over 400 *CYPs*. The soybean genome has 332 full-length *CYP* genes and 368 pseudogenes for a total of 700 *CYP* sequences.

Organisation of the Homepage

The Homepage has grown over time. A format was chosen to allow easy expansion and rapid access to the parts you most wanted to visit. The first page of the website is the main access point (Figure 1) and it serves as a visual sitemap. The coloured box at the top marked P450 WEB MATRIX is a link to another page with P450-related sites. Some of these sites are dedicated to specific groups of species, like insects, plants or fungi, and they may include expression data, phylogenetic trees and other features. Below the WEB MATRIX box is a link to a description of how to search the site using Google. This may allow the discovery of things that are not obvious from the table layout. The next section is the main table. At the top there are four rows with general links. The most important of these is the nomenclature link, which takes you to a page with links to five main bibliographic files that include the *CYP* nomenclature entries. These are broken into sections to keep the file sizes smaller. Even so, the files are getting to be fairly large. Part C is 685 html text pages. The nomenclature has *CYP* numbers reserved, as shown in Table 1. Original plans were for 100 *CYP* families only, but this was a gross underestimate. We had to go to three-digit *CYP* names and now we are using four-digit *CYP* names for bacteria and lower eukaryotes. This required the first reserved blocks of numbers to be multiplied by ten and then 100 to continue the numbering. Four of the five files are specific for animals (Part A, Part B), plants (Part D) and bacteria (Part E). Part C covers families — *CYP10–69*, *CYP301–699* and *CYP3001–6999* — which include both animals and lower eukaryotes. This may need separation in the future.

There are three families that occur in more than one group. *CYP51* is found in all kingdoms: plants, animals, lower eukaryotes and bacteria, although there are species that do not have *CYP51. CYP51* was the first fungal P450 sequence identified from *Saccharomyces cerevisiae.* Therefore, it has a lower eukaryote number. *CYP97* is found in plants and some lower eukaryotes like diatoms. It was first described in *Arabidopsis*, so it has a plant *CYP* number. *CYP701* is also found in plants and lower eukaryotes, including choanoflagellates, the ancestors of animals. This family is orthologous to fungal *CYP61. CYP51*, *CYP61* and *CYP701* are found in sterol biosynthesis pathways.

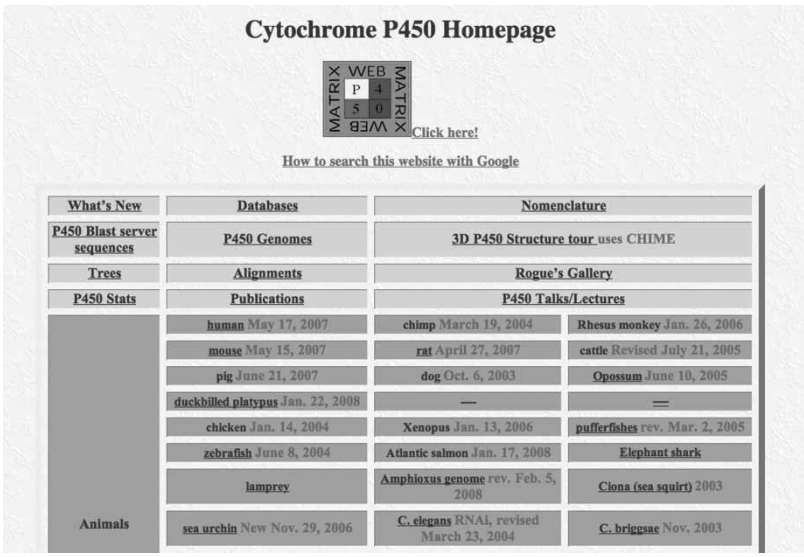

Figure 1. The top of Cytochrome P450 Homepage.

Table 1. CYP numbering scheme

Animals	Plants	Lower eukaryotes
1–49 (*CYP49A2*)	71–99 (*CYP99A12*)	51–69 (*CYP69A1*)
300–499 (*CYP379A3*)	701–999 (*CYP804A1*)	501–699 (*CYP699A2*)
3001–4999 (reserved)	7001–9999 (reserved)	5001–6999 (*CYP5314A1**)
Bacteria		
101–299 (*CYP*299A1)		
1001–2999 (*CYP*1015A1)		

Numbers in parentheses are the last assigned *CYP* name in that block.
*An unusual group of hybrid fungal P450s are given the block 6001–6099 (6004A2) so that they can be kept together.

Navigation inside the nomenclature files can be accomplished by bookmarks to specific locations. These are links found at the top of the file. These links will jump you to the family or subfamily indicated. You can use a browser find command at any time. This will take you to the next occurrence of a word or name in the file. You can use the back button to return to the nomenclature index page or the main page.

The link marked 'databases' in row one is an effort to make the data more accessible. The Cytochrome P450 Homepage is not a relational database and it suffers from the drawbacks of a long text file that contains the information. These database files were intended to bring information on individual *CYP* subfamilies into a format where they could be easily accessed, with links to the sequence data, phylogenetic trees and the nomenclature pages all in one table each for plants, animals, lower eukaryotes and bacteria/archaea. The idea is good but little data entry has occurred in these files, as the pressure to name more and more genes has pushed these other aspects of the Homepage into lower priority status.

On the second row of the main table there is a link to P450 BLAST server sequences. These are files of *CYP* sequences that are searchable by BLAST on the P450 BLAST server (http://blast.uthsc.edu). These sequences are presented by *CYP* name for the given species or group. The files on the server have not been updated recently, but more current sequences have been added to this page. The files: 'All bacterial P450s' and 'All fungal P450s' are new. Plans are in progress to revamp the P450 Blast server to include all 11,512 named P450 sequences. Currently, there are no single files that contain all the animal (3,282 sequences), plant (4,266 sequences) or protist (247 sequences) P450s; however, many of these sequences are available on the appropriate species pages.

The three-dimensional P450 structure tour was constructed using Protein Explorer, which depends on a CHIME plug-in. The newer Netscape and Firefox browsers do not support the CHIME plug-in, so this page is difficult to view as it was intended to be. It does run perfectly well on a Mac G4 OS 9.1 with Netscape Communicator 4.7 as the browser, the CHIME plug-in installed and the mime-type chemical/x-pdb added to the accepted mime types in the browser preferences. There are still places to download old versions of Netscape (http://browser.netscape.com/releases) for the determined P450 afficionado. The structure tour has 37 views of *CYP101A1* and *CYP2C5* P450 structure, emphasising various features around the molecule, including salt bridges, interactions of side chains with the haem propionates, the PERF motif, etc. I am not currently planning to convert this tour to a more user-friendly version.

The P450 genomes page (row two) lists completed genomes and the number of P450s found in each. This page has not been updated and will need considerable revision. The fungal genomes page (http://drnelson.utmem.edu/fungal.genomes.html) now includes 55 complete fungal genomes, with links to the named P450 sequences for each species.

In 1995, when the Homepage was started, I created a directory of P450 researchers, with e-mail addresses and pictures if I could find them. This is called the Rogue's Gallery. There is a picture of the very young author in these pages. Although many of the links are no longer functional, these pages are being updated as time permits.

The P450 statistics page on row four has information on how many P450s are named in the

various sections: plant, animal, fungi, protist and bacteria. The older versions from previous years are included. This page has links to some useful spreadsheets that contain all the known *CYP* names. Most of these are in *CYP* name order, but a few new ones have been sorted by species. There are August 2009 files for all P450s from plants, animals, insects, fungi, protists and bacteria. These files were used to count the P450s after subtracting the variants (usually indicated by v1, v2 etc after the *CYP* name). The P450 counts do include pseudogenes. Table 2 shows the current P450 statistics.

Several links, such as 'What's New', 'Trees' and 'Alignments', are self explanatory. The 'What's New' page is not updated very often, even though many additions and revisions to the Homepage are being made frequently. The 'Trees' and 'Alignments' pages are also pretty old files. More comprehensive sets of alignments and trees can be found at the Cytochrome P450 Engineering Database[5] (http://www.cyped.uni-stuttgart.de/). More trees showing fungal P450 clans can be found on the fungal P450 page. Additional trees are linked from inside the database's pages.

The publications page has been updated to include papers from 2009. Many of the older papers have links to pdf files. The newer papers have links to the MedLine abstract. Often the high-profile genome papers in *Science* and *Nature* are free of charge at the publisher's site. Next to the publications page is the P450 Talks/Lectures page. Here are found the transcripts and PowerPoint presentations of 14 lectures I have given since 2000. The text has indications of slide numbers embedded, so one can easily follow along with the PowerPoint presentation. Some of the older lectures have the slides inserted in the text. There is also a 2009 version of a medical school introductory lecture on human P450s.

Table 2. Named P450 sequences (19th August, 2009)

Animals	3,282
Insects	1,675 (part of the animal total)
Animals (not insects)	1,607
Plants	4,266
Fungi	2,784
Protists	247
Bacteria	905
Archaea	26
Viruses	2 (Mimivirus)

Total 11,512

The species pages

The rest of the main table is divided into four sections: animals, plants, lower eukaryotes and bacteria. The lower eukaryotes section is subdivided into additional sections and fungi have been given a whole page of their own, due to the recent sequencing of nearly 100 species of fungi. Clicking on a species link will take you to a species page, where sequence data annotation information and more links will be available. The best way to explore the Homepage is to click on links and read the annotations at the tops of the files.

Animals

Human and mouse are the most extensively annotated of the animal *CYPs*.[6] I have made master tables with links to a wide variety of data for these two species. Links to the master tables are found on the mouse and human species pages. The human master table contains more than 700 links to human P450 data. There are ten columns of data in the human table (Figure 2). These include links to protein, mRNA, genomic DNA, Online Mendelian Inheritance in Man (OMIM), HUGO Gene Nomenclature Committee (HGNC), Unigene and Entrez Gene entries, plus three-dimensional structures, including the recent aromatase CYP19 structure. Column one also has links to seven manually created gene cluster maps. This table is at the mercy of the other databases that are linked. In 2007, all the HGNC data links looked

like http://www.gene.ucl.uk/nomenclature/data/get_data.php?hgnc_id=nnnn, but in 2009 they all look like http://www.genenames.org/data/hgnc_data.php?hgnc_id=nnnn. Therefore, 141 links to the HGNC had to be changed in the table. The mouse master table is very similar but it does not contain OMIM or HGNC data columns. The mouse table has links to the UCSC browser (column 6), which shows the vicinity of the *CYP* genes. From the mouse species page, a tree is available that includes 103 mouse and 60 human sequences. Orthologue pairs are shown in red.

Each species page was created separately and contains different information in addition to sequence data. *Drosophila melanogaster* is another highly studied animal. There is a dedicated P450 site that includes several insect species, including *Drosophila*. (http://p450.sophia.inra.fr/), which is linked at the top of the *D. melanogaster* species page. I highly recommend it. There are now 12 species of *Drosophila* that have been sequenced.[7] Our site has only analysed two: *D. melanogaster* and *D. pseudoobscura*. The *D. pseudoobscura* page has both species' P450s side by side for easy comparison, with *D. melanogaster* in blue and *D. pseudoobscura* in red. A tree with 169 *Drosophila* sequences included is linked from the *D. pseudoobscura* page. Unique information on our *D. pseudoobscura* page includes a sequence alignment with all the intron locations colour coded for phase. There is also a discussion of intron gain and loss. The Fungal Cytochrome P450 Database[8] in Korea is beginning to include animal data. The sequences from the 12 *Drosophila* species are included on the website (http://p450.riceblast.snu.ac.kr/index.php?a=view).

Species pages include the dates of the last major revision. The animal pages have dates from 2003 to 2008. Pages may be updated at any time. For example, the zebrafish page will soon be updated, since a paper on the *Danio rerio* genome *CYP*s is in preparation. Some pages are placeholders for data that have not yet been processed. Elephant shark, lamprey and Acropora (a coral) are examples. Some of these, such as sponge, are awaiting genome papers to be released before the sequence data can be made public.

Plants

Plants have many more P450s than animals. The soybean has about six times more *CYP* genes than the human — and that only counts full genes. Because of the larger number of genes, and the

CYP	Genbank	RefSeq mRNA	Genomic Assemblies	RefSeq Protein	OMIM	HGNC	Unigene	EntrezGene	comments
1A1	see related sequences under the Entrez Gene link	NM_000499	see reference assembly under the Entrez Gene link	NP_000490	108330	2595	Hs.72912 (70 ESTs)	CYP1A1	15q24.1
1A2	see related sequences under the Entrez Gene link	NM_000761	see reference assembly under the Entrez Gene link	NP_000752	124060	2596	Hs.1361 (26 ESTs)	CYP1A2	15q24.1
1D1P/1A8P	AL359997.8	no entry	NT_023935.17	no entry	no entry	hold	no entry	no entry	9q21.12 43% identical to 1A2, ortholog to Danio 1D1
1B1	see related sequences under the Entrez Gene link	NM_000104	see reference assembly under the Entrez Gene link	NP_000095	601771 see additional entries under phenotypes in the Entrez Gene link	2597	Hs.154654 (584 ESTs)	CYP1B1	2p22.2
2A 2ABFGST gene cluster map	see related sequences under the Entrez Gene link	no entry	NG_000008	no entry	123960	2607	no entry	CYP2ABFGST cluster	19q13.2
2A6	see related sequences under the Entrez Gene link	NM_000762	see reference assembly under the Entrez Gene link	NP_000753	122720	2610	Hs.439056 (217 ESTs)	CYP2A6	19q13.2 Structures 2FDY 2FDW 2FDV 2FDU 1Z11 1Z10
2A6v2	U22027	no entry	no entry	no entry	122720	hold	no entry	CYP2A6	2A6*3 allele
2A7v1	see related sequences under the Entrez Gene link	no entry	see reference assembly under the Entrez Gene link	no entry	608054	2611	Hs.250615 (34 ESTs)	CYP2A7	19q13.2 wt seq

Figure 2. The top part of the Human P450 Master Table.

emphasis on animals, not many plant genomes have yet been sequenced. Other plant genomes, such as pine and ferns, have been deferred because the genomes are huge and too expensive to sequence at the moment. Papers have been published comparing *CYP*s from *Arabidopsis* and rice;[9] these two plus *Chlamydomonas*, poplar and moss;[10] and comparing *CYPs* from six completed plant genomes, including papaya and grape.[11] *Arabidopsis* has other sites dedicated to *CYPs* in Denmark (http://www.p450.kvl.dk/) and The University of Illinois at Urbana-Champaign (http://arabidopsis-p450.biotec.uiuc.edu/). The Arabidopsis Information Resource site (TAIR) also contains P450 data (http://www.arabidopsis.org/).

Genome sequencing centres such as the Broad Institute and the Joint Genome Institute (JGI) have their own genome browsers to access data from their projects. The cottonwood *Populus trichocarpa* was sequenced at JGI. My *P. trichocarpa* page has a detailed tutorial on how to use the JGI browser. This tutorial is applicable to any of the genomes at JGI. Recently, a plant genome browser called Phytozome 4.0 was released (http://www.phytozome.net/). This is like the University of California, Santa Cruz (UCSC) browser for animals except that it has 14 complete plant genomes included. One can search for P450 as a keyword against 20 nodes on a schematic phylogeny of the plant genomes. As an example, a search of the rice/*Brachypodium distachyon* clade identified 309 P450 gene clusters that presumably contain orthologues or clusters of closely related genes. Our rice page has a file showing the comparison of rice and *Arabidopsis CYP*s with gene names in a side-by-side table format (http://drnelson.utmem.edu/rice.arab.list.htm).

A more comprehensive table with six plant species is included as Supplementary Table 2 to Nelson[10] (http://drnelson.utmem.edu/Nelson.SuppFigsandTables.pdf).

Fungi

Fungal genome sequences have been accumulating rapidly. The dedicated Korean Fungal Cytochrome P450 Database[8] has P450s from 87 species of fungi, five oomyctes (stramenopiles) and *Leishmania infantum* (Euglenozoa) (http://p450.riceblast.snu.ac.kr/index.php?a=view). These data are obtained from an automated pipeline designed to find P450s from genomic data. The method is rapid and very good at finding P450s. Many of the gene assemblies found by this automated method are very short fragments, however, often missing exons and/or splicing the exons together at incorrect locations. The authors have attempted to assign P450 families based on a comparison with the manually curated sequences at the Cytochrome P450 Homepage. The process is very useful for surveying a genome for P450s, but manual curation is needed in the end to achieve a more accurate gene assembly. So far, I have processed 55 fungal genomes, three oomycete genomes and six Euglenozoa genomes. These P450s are on my website, with *CYP* names assigned after manual annotation of the sequences. The Korean Fungal Cytochrome P450 Database contains 32 fungal genomes that had not been entered into the Cytochrome P450 Homepage as of 16th August, 2009. These genomes have 3,471 predicted P450s in them, so the Homepage site, with 2,784 named fungal P450s, is not up to date, but, ultimately, hand curation produces more accurate gene assemblies. Other sites, such as the Fungal Cytochrome P450 Database[8] and the Cytochrome P450 Engineering Database,[5] depend on the systematic nomenclature found on the Homepage to predict family or subfamily membership for new sequences.

Other lower eukaryotes and Mimivirus

Below 'fungi' in the main table is a collection of protist P450s. These organisms are eukaryotes that are not plants, animals or fungi, so they are very diverse. Aside from *CYP51* and *CYP710*, these organisms do not share *CYP* families with other eukaryotes. A complete list of the 247 *CYP*s in a spreadsheet format is available under the 'P450 Stats' link. The sequences are accessible from the species pages. The large virus called Mimivirus has two *CYP*s, *CYP5253A1* and *CYP5254A1*.[12] The origin of these *CYP*s is unclear, but they may have been

picked up from an amoeba host, so they are included in the *CYP* numbers reserved for lower eukaryotes.

Bacteria

The last link in the main table is to bacteria. This also includes the 26 named 'Archaeal *CYP*s'. Some bacteria in the Actinobacteria have 18 to over 30 *CYP*s (*Streptomyces* species and *Mycobacterium* species), while most other bacteria tend to have only a few or none. Because bacteria generally have few P450s, there are no bacterial species pages. The *CYP*s are presented in name order. P450s in bacteria are often associated with antibiotic biosynthesis. As an interesting aside, the first analysis of the Sorcerer II Global Ocean Sampling expedition[13] identified 3,305 *CYP*s from marine bacteria. These have not been named.

Links below the main table

The Cytochrome P450 Homepage is visited by many P450 researchers, so links have been provided to P450-related meetings. These appear below the main table. Many of the old meeting links still work. There are also some links to other P450 sites, although these have mostly been moved to the P450 WEB MATRIX table. The old links have not been removed, so this is more or less a historical list.

Future plans

The Homepage is primarily a nomenclature site. It has run on a desktop computer in my office for 14 years, but it needs a long-term permanent home, possibly with the National Center for Biotechnology Information (NCBI) or some other established sequence repository. Once the nomenclature has saturated a phylum such that no new families are being discovered, it may not be necessary to continue to name every sequence from every species. The recent sequencing of numerous mammalian genomes has shown that all mammalian, and possibly all vertebrate, *CYP* families have been found. Orthologue identification, and naming orthologues with the same name, is already beginning to simplify the nomenclature. There is a need for a deeper time nomenclature that recognises very deep time clades of P450s. This is the clan nomenclature, and there will be more about this nomenclature on the Homepage in the future, as it becomes more fully developed. Of great significance for this Homepage is a mechanism for funding the work. Even though this nomenclature system is 22 years old, and this website is 14 years old, no grant has ever been given specifically for this annotation effort. This is probably not sustainable, yet annotation is fundamental to progress in genomics. Funding agencies need to realise the essential nature of gene annotation by funding nomenclature efforts.

References

1. Nelson, D.R., Kamataki, T., Waxman, D.J., Guengerich, F.P. *et al.* (1993), 'The P450 superfamily: Update on new sequences, gene mapping, accession numbers, early trivial names of enzymes and nomenclature', *DNA Cell Biol.* Vol. 12, pp. 1–51.
2. Nebert, D.W. and Gonzalez, F.J. (1987), 'P450 genes: Structure, evolution, and regulation', *Annu. Rev. Biochem.* Vol. 56, pp. 945–993.
3. Nebert, D.W., Nelson, D.R., Adesnik, M.A., Coon, M.J. *et al.* (1989), 'The P450 superfamily: Update on listing of all genes and recommended nomenclature of the chromosomal loci. *DNA* Vol. 8, pp. 1–13.
4. Nebert, D.W., Nelson, D.R., Coon, M.J., Estabrook, R.W. *et al.* (1991), 'The P450 superfamily: Update on new sequences, gene mapping and recommended nomenclature', *DNA Cell Biol.* Vol. 10, pp. 1–14.
5. Fischer, M., Knoll, M., Sirim, D., Wagner, F. *et al.* (2007), 'The Cytochrome P450 Engineering Database: A navigation and prediction tool for the cytochrome P450 protein family', *Bioinformatics* Vol. 23, pp. 2015–2017.
6. Nelson, D.R., Zeldin, D.C., Hoffman, S.M.G., Maltais, L. *et al.* (2004), 'Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes including nomenclature recommendations for genes, pseudogenes, and alternative-splice variants', *Pharmacogenetics* Vol. 14, pp. 1–18.
7. Drosophila 12 Genomes Consortium (2007), 'Evolution of genes and genomes on the Drosophila phylogeny', *Nature* Vol. 450, pp. 203–218.
8. Park, J., Lee, S., Choi, J., Ahn, K. *et al.* (2008), 'Fungal cytochrome P450 database', *BMC Genomics* Vol. 9, p. 402.
9. Nelson, D.R., Schuler, M.A., Paquette, S.M., Werck-Reichhart, D. *et al.* (2004), 'Comparative genomics of *Oryza sativa* and *Arabidopsis thaliana*. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot', *Plant Physiol.* Vol. 135, pp. 756–772.
10. Nelson, D.R. (2006), 'Plant cytochrome P450s from moss to poplar', *Phytochem. Rev.* Vol. 5, pp. 193–204.
11. Nelson, D.R., Ming, R., Alam, M. and Schuler, M.A. (2008), 'Comparison of cytochrome P450 genes from six plant genomes', *Tropical Plant Biology* Vol. 1, pp. 216–235.
12. Lamb, D.C., Lei, L., Warrilow, A.G., Lepesheva, G.I. *et al.* (2009), 'The first virally encoded cytochrome p450', *J. Virol.* Vol. 83, pp. 8266–8269.
13. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L. *et al.* (2007), 'The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families', *PLoS Biol.* 5, p. e16.