# Putative cytochrome P450 genes in rice genome (*Oryza sativa* L. ssp. *indica*) and their EST evidence

ZHONG Lan (钟 兰)[1,2], WANG Kai (王 凯)[2], TAN Jun (谭 军)[2], LI Wei (李 蔚)[2,3] & LI Songgang (李松岗)[2,1]

1. College of Life Science, Peking University, Beijing 100871, China;
2. Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China;
3. Laboratory of Bioinformatics, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
Correspondence should be addressed to Li Songgang (email: lisg@genomics.org.cn)

Received July 26, 2002

**Abstract** We discovered 528 putative cytochrome P450s (P450s) in *Oryza sativa* L. ssp. *indica* using *Arabidopsis thaliana* P450s as database. Those putative rice P450s are thought to belong to 40 families classified in *Arabidopsis thaliana*. We compared distributions of *Arabidopsis thaliana* and *Oryza sativa* P450s and found the two species have similar distribution patterns. However, family distributions of two species also have some differences. For example, in rice, the gene number in families of CYP71, CYP72, CYP76, CYP89, CYP94 and CYP709 is more than twice that in *Arabidopsis thaliana*; and there are 33 CYP705 members in *Arabidopsis thaliana* but none in rice. We also found gene members in CYP71 and CYP81 are organized as tandem arrays repeated in the rice genome; maybe they are duplications in the evolutionary event. Furthermore, we accumulated expression sequence tag (EST) evidence for 263 putative rice P450s, which are expressed at transcriptional level and more likely to be true P450s.

**Keywords: cytochrome P450, EST, *Oryza sativa* L. ssp. *indica*, *Arabidopsis thaliana*.**

The term ‘cytochrome P450’ (P450s) was coined in 1962 as a temporary name for a colored substance in cell[1]. This pigment, when reduced and bound with carbon monoxide, produced an unusual absorption peak at a wavelength of 450 nm[2]. In fact, cytochrome P450s are not true cytochromes but enzymes involved in many basic metabolic pathways. However, the name ‘cytochrome P450’ has been widely accepted, any change will be inconvenient[2].

P450s are heme-thiolate proteins playing significant roles in biological systems. Their functions range from synthesis and degradation of endogenous steroid hormones, vitamins and fatty acid derivatives (‘endobiotics’) to the metabolism of foreign compounds such as drugs, environmental chemicals, and carcinogens (‘xenobiotics’)[3]. In plants they are involved in plant hormone synthesis, phytoalexin synthesis, flower petal pigment biosynthesis, and herbicide degradation. Although reactions they catalyze are extremely diverse, P450s usually work by activating molecular oxygen with inserting one of its atoms into the substrate and reducing the other to form water. $RH + O_2 + NADPH + H^+ = ROH + H_2O + NADP^+$. So they are also called monooxygenases[4].

P450s compose a very large gene super family. They have been isolated from bacteria, fungi, yeast, insects, plants, mammals and so on[5]. Thus, they are thought to present in all three kingdoms[6]. About 80 P450s in animal genomes and hundreds of P450s in plant genomes have been reported[7]. In prokaryotes, P450s are soluble proteins. While in eukaryotes, they are usually bound to the inner mitochondrial membranes or endoplasmic reticulum[8] through a short hydrophobic segment of their N-terminus, and possibly a hydrophobic loop of the protein[9].

Plant P450s are generally classified into two main clades: A-type and non-A type[7,10]. The A-type clade is specific to plants, some P450s involved in the biosynthesis of secondary metabolites or natural products are found in this group. In contrast, the non-A type clade is a much more divergent group of sequences consisting of several individual clades, which often show more similarity to non-plant P450s than to the other plant P450s[7]. It was initially proposed by Durst and Nelson in 1995 that the A-type P450s originated from a single common ancestral gene[10]. A previously published detailed phylogenetic analysis of 135 *Arabidopsis thaliana* sequences coupled to a comparison of the gene organization[7] confirmed this proposal.

Original nomenclature system of P450 genes from all organisms has been set up on the basis of protein sequence identity and phylogeny[3]. P450s in the same family usually share at least 40% identity. In the same subfamily they normally share at least 55% identity. However, this identity rule has some exceptions, especially in plants, where gene duplication and shuffling sometimes makes a straightforward nomenclature difficult. Sequence identity among P450s from *Arabidopsis thaliana* can be less than 20%. In this case, family assignment is based on phylogeny and gene organization. P450s are named in chronological order of submission to a nomenclature committee (David Nelson: DNELSON@utmem1.utmem.edu). And for plant, P450s are assigned names from CYP71A1 to CYP99XY, then from CYP701A1 and above[4].

## 1 Materials and methods

First, rice (*Oryza sativa* L. ssp. *indica*) scaffolds were retrieved at web site http://btn.genomics.org.cn/rice. We selected FgeneSH[11], which was found the best software available at present for finding genes of Monocot[12,13], to predict genes from those scaffolds. Meanwhile, with the coding sequences information reported in FgeneSH results, DNA sequences of those predicted genes were picked up from the corresponding scaffolds.

Second, amino acid sequences of the predicted rice genes were aligned to P450 database of *Arabidopsis thaliana*, which is released at web site http://drnelson.utmem.edu/Arabidopsis.Blast.file.html on August 17, 2001, by BLASTP (E-value: <=1e-7 and identity%: >=25%) search[14,15]. Three other databases of *Arabidopsis thaliana*, maize, and *Oryza sativa* were also downloaded from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

Third, we tried to explore EST evidence for those putative rice P450 genes. ESTs were generated in Beijing/Hangzhou Genomics Center. The ESTs were assembled using Phred[16,17], a base calling tool, and Phrap, a sequence alignment and contig assembly program (P. Green, http://

genome.washington.edu), and thus generating EST contigs. In order to confirm the BLASTP results, we aligned DNA sequences of the putative rice P450s genes to EST contigs by stringent BLASTN (E-value: <=1e-15; Identity%: >=80%) search[18].

## 2 Results

FgeneSH predicted 71993 putative genes from a total of 103044 rice scaffolds. We aligned those putative rice genes to 289 known *Arabidopsis thaliana* P450s and found 528 putative rice cytochrome P450s. According to the BLASTP results, those putative P450s are located in 433 scaffolds and belong to 40 families classified in *Arabidopsis thaliana*. Furthermore, 263 of the 528 putative P450s genes have been confirmed by EST evidence.

Table 1 lists A type and non-A type P450 genes according to their family information. We totally found 17 families of A type P450s and 23 families of non-A type P450s. Putative P450 number and EST confirmed P450 number of rice are listed for each family.

Table 1 528 putative rice P450s and EST evidence for 263 of them

| A type P450s | | | Non-A type P450s | | |
|---|---|---|---|---|---|
| Family | putative P450 number | EST confirmed P450 number | family | putative P450 number | ESTs confirmed P450 number |
| CYP71 | 128 | 65 | CYP51 | 16 | 4 |
| CYP73 | 4 | 2 | CYP72 | 29 | 12 |
| CYP75 | 18 | 15 | CYP74 | 5 | 3 |
| CYP76 | 54 | 29 | CYP85 | 1 | 0 |
| CYP77 | 5 | 5 | CYP86 | 14 | 4 |
| CYP78 | 9 | 3 | CYP87 | 17 | 10 |
| CYP79 | 4 | 4 | CYP88 | 6 | 2 |
| CYP81 | 16 | 8 | CYP90 | 11 | 2 |
| CYP82 | 1 | 0 | CYP94 | 38 | 9 |
| CYP83 | 2 | 1 | CYP96 | 16 | 7 |
| CYP84 | 5 | 2 | CYP97 | 3 | 3 |
| CYP89 | 33 | 20 | CYP702 | 2 | 0 |
| CYP93 | 3 | 2 | CYP704 | 9 | 6 |
| CYP98 | 7 | 7 | CYP707 | 3 | 3 |
| CYP701 | 5 | 5 | CYP708 | 0 | 0 |
| CYP703 | 2 | 1 | CYP709 | 20 | 8 |
| CYP705 | 0 | 0 | CYP710 | 5 | 5 |
| CYP706 | 4 | 3 | CYP711 | 6 | 1 |
| CYP712 | 0 | 0 | CYP714 | 9 | 4 |
| | | | CYP715 | 4 | 0 |
| | | | CYP716 | 9 | 6 |
| | | | CYP718 | 0 | 0 |
| | | | CYP720 | 0 | 0 |
| | | | CYP721 | 2 | 0 |
| | | | CYP722 | 2 | 1 |
| | | | CYP724 | 1 | 1 |
| Total | 300 | 172 | total | 228 | 91 |

In figs. 1 and 2, family distributions of *Arabidopsis thaliana* and rice P450s were compared for A type and non-A type respectively. The gene numbers of A type range more widely than that of non-A type. In fact, the non-A type P450 numbers range from 0 to 38, while the A type P450 numbers range from 0 to 128.

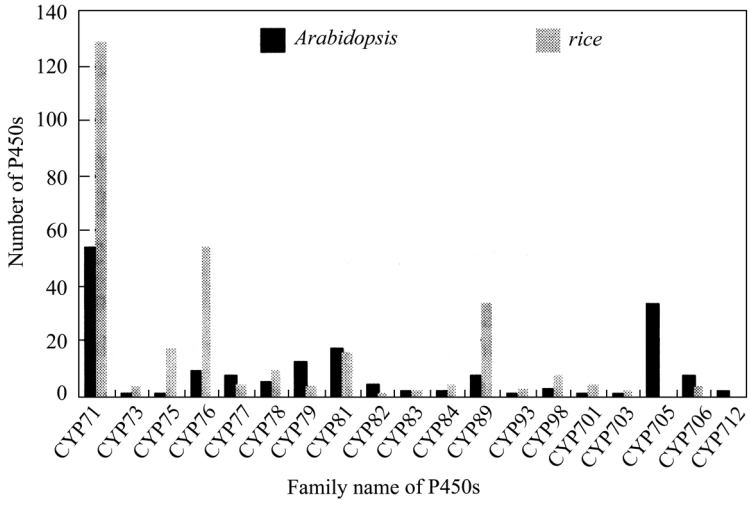


Fig. 1. Family distributions of *Arabidopsis thaliana* and rice A type P450s. No putative rice P450 was found in CYP705 and CYP712. CYP71 seems to be the largest P450 family in rice and has 128 gene members.

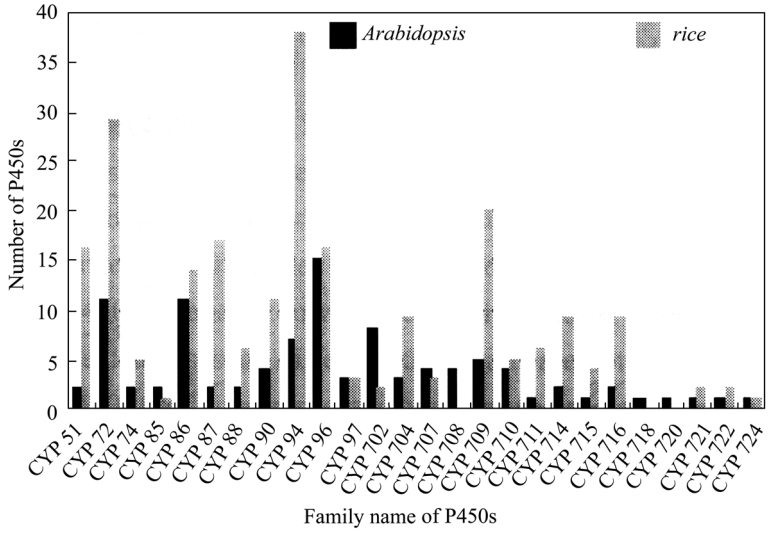


Fig. 2. Family distributions of *Arabidopsis thaliana* and rice non-A-type P450s in families. No putative rice P450 was found in CYP708, CYP718 and CYP720.

## 3 Discussions

First, in some families, P450 members neighbor to each other very closely in the genome. For example, Scaffold238_1, Scaffold238_4, Scaffold238_5, Scaffold238_7, Scaffold238_8, Scaffold238_9, Scaffold238_10 are all CYP71 family members and located in the very same scaf-

scaffold238; and Scaffold10047_2, Scaffold10047_3, Scaffold10047_4, Scaffold10047_5 are all CYP81 family members and organized as tandem arrays in the Scaffold10047. They might be duplications in the evolution event.

Second, no rice gene was found in families of CYP705, CYP708, CYP712, CYP718, and CYP720. But this does not mean there was no qualified BLASTP match to those 5 families. In fact, there were many qualified BLASTP matches (with E-value less than 1e-7; and identity% more than 30%) to the 5 families. However, the predicted rice genes matched better with other families than with the 5 mentioned above.

Third, in figs. 1 and 2, we compared the family distributions of *Arabidopsis thaliana* and rice P450s. The two plants have almost similar P450 distribution patterns but still have some differences. The gene number in some rice P450 families, such as CYP71, CYP72, CYP76, CYP89, CYP94, CYP709 is more than twice that in the corresponding *Arabidopsis thaliana* families. There is a puzzling phenomenon that *Arabidopsis thaliana* has 33 CYP705 members, whereas rice has none of them. Consequently, intron-exon organization and multiple alignment of the candidate P450s is further needed to verify the family assignment.

In order to have more confidence to say the candidate genes are true P450s in rice, we align them to rice EST contigs. As a result, 263 candidate genes have EST homologies and are more likely to be real cytochrome P450 genes than the other 265 candidates. As we know, rice is monocot, while *Arabidopsis thaliana* is dicot. They are quite different organisms in biological sense. Intrinsic differences of the two species, for example GC gradient in monocots but not in dicot[19], may affect their sequence arrange mode at protein level as well as at DNA and RNA level. Consequently, BLAST search among different species is just a fast but not an accurate way to detect homologies between the queries and the subjects. Unknown P450s in rice cannot be found for the limit of database search method. When we use the 289 known *Arabidopsis thaliana* P450s retrieved from the University of Tennessee as database, unknown P450s or the P450s unique in monocot but not in dicot cannot align well with *Arabidopsis thaliana* sequences, therefore they could not be detected. Further prudent analysis is needed to confirm or revise this preliminary blast result.

In addition, we aligned rice genes with three other P450 databases got from National Center for Biotechnology Information (NCBI). Because the databases were different, the results were not the same. 638, 521 and 473 putative P450s were found respectively when we used databases of *Arabidopsis thaliana*, maize, and *Oryza sativa*. After de-redundancy and excluding the forgoing mentioned 528 putative P450s, 125 new putative P450s were found. So a total of 653 P450s were found in *Oryza sativa* L. ssp. *indica* by BLASTP (E-value: <=1e-7; Identity%: >=25%) search with the entire four public P450 databases.

We made a rude survey of rice P450s from our own data. In order to study the exact functions of individual P450, further work, such as cloning and sequencing P450s remains to be done. We hope molecular biologists can find some clues in identifying individual P450 by reading this

paper. By the way, there is detailed information of P450s from tens of different species at web site http://drnelson.utmem.edu/CytochromeP450.html, which also released rice P450s data got from public database. Our work in this paper is a complementary to the experts' efforts.

**Acknowledgements** We thank all the people who contributed to sequencing the rice working-draft and ESTs in Beijing/Hangzhou Genomics Center. They enable us to have this chance to study one of the most important super families of cytochrome P450.

## References

1. Oruma, T., Sato, R., A new cytochrome in liver microsome, J.Biol.Chem., 1962, 237: 1375—1376.
2. McKinnon, R. A., Cytochrome P450 (I)——Multiplicity and function, Aust. J. Hosp. Pharm., 2000, 30: 54—56.
3. Nelson, D. R., Koymans, L., Kamataki, T. et al., P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature, Pharmacogenetics, 1996, 6(1): 1—42.
4. Werck-Reichhart, D., Bak, S., Paquette, S., Cytochromes P450, in The *Arabidopsis* Book (eds. Somerville, C., Meyerowitz, E.), American Society of Plant Biologists, 2002, 1—29.
5. http://www.csupomona.edu/~jis/1997/Bozak.pdf.
6. Xu, W., Bak, S., Decker, A. et al., Microarray-based analysis of gene expression in very large gene families: The cytochrome P450 gene superfamily of *Arabidopsis thaliana*, Gene, 2001, 272(1-2): 61—74.
7. Paquette, S. M., Bak, S., Feyereisen, R., Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*, DNA Cell Biol., 2000, 19(5): 307—317.
8. Werck-Reichhart, D., Feyereisen, R., Cytochromes P450: a success story, Genome Biol., 2000, 1(6): 3003.1—3003.9.
9. Williams, P. A., Cosme, J., Sridhar, V. et al., Mammalian microsomal cytochrome P450 monooxygenase: Structural adaptations for membrane binding and functional diversity, Mol. Cell, 2000, 5: 121—131.
10. Durst, F., Nelson, D. R., Diversity and evolution of plant P450 and P450-reductases, Drug Metabol. Drug Interact., 1995, 12(3-4): 189—206.
11. Salamov, A. A., Solovyev, V. V., *Ab initio* gene finding in Drosophila genomic DNA, Genome Res., 2000, 10(4): 516—522.
12. Yu, J., Hu, S. N., Wang, J. et al., A draft sequence of the rice (*Oryza sativa* ssp. *indica*) genome, Chinese Science Bulletin, 2001, 46(23): 1937—1941.
13. Yu, J., Hu, S., Wang, J. et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science, 2002, 296(5565): 79—92.
14. Altschul, S. F., Gish, W., Miller, W. et al., Basic local alignment search tool, J. Mol. Biol., 1990, 215(3): 403—410.
15. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res., 1997, 25(17): 3389—3402.
16. Ewing, B., Green, P., Base-calling of automated sequencer traces using phred (II)——Error probabilities, Genome Res., 1998, 8(3): 186—194.
17. Ewing, B., Hillier, L., Wendl, M. C. et al., Base-calling of automated sequencer traces using phred (I)——Accuracy assessment, Genome Res., 1998, 8(3): 175—185.
18. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. et al., Identification of novel genes coding for small expressed RNAs, Science, 2001, 294(5543): 853—858.
19. Wong, G. K., Wang, J., Tao, L. et al., Compositional gradients in gramineae genes, Genome Res., 2002, 12(6): 851—856.