Plant Biotechnology 31, 579–584 (2014)
DOI: 10.5511/plantbiotechnology.14.0901a

Note

Expansion of specialized metabolism-related superfamily genes via whole genome duplications during angiosperm evolution

Yosuke Kawai[1], Eiichiro Ono[2], Masaharu Mizutani[3,*]

[1]Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Miyagi 980-8573, Japan; [2]Research Institute, Suntory Global Innovation Center Ltd., Mishima, Osaka 618-8503, Japan; [3]Functional Phytochemistry, Graduate School of Agricultural Science, Kobe University, Kobe, Hyogo 657-8501, Japan
*E-mail: mizutani@gold.kobe-u.ac.jp Tel: +81-78-803-5885

Received July 22, 2014; accepted September 1, 2014 (Edited by A. Sugiyama)

Abstract Specialized metabolism in land plants produces the diverse array of compounds, which is important in interaction with the environments. Generally, specialized metabolism-related genes consist of large gene families (superfamily), including cytochrome P450 monooxygenases (CYPs), 2-oxoglutarate-dependent dioxygenases (DOXs), and family-1 UDP-sugar dependent glycosyltransferases (UGTs), especially in angiosperms and gymnosperms. We investigated the changes in the numbers of these superfamily genes during the evolution of angiosperms by inferring gain and loss events in ancestral lineages of 5 angiosperms and 1 lycophyte. We observed the clear difference in the changes in the gene number among ancestral lineages. Intriguingly, gene gain events were coordinately occurred among CYP, DOX and UGT in lineage-specific manner, and the gain events were in good accordance with ancient whole genome duplication (WGD) events. Thus, the WGD events in angiosperms would have an important role in the expansion and evolution of specialized metabolism by providing prerequisite genetic resources for subsequent lineage-specific local tandem duplication (LTD) of superfamily genes as well as functional differentiation of these superfamily genes.

Key words: CYP, DOX, specialized metabolism, UGT, whole genome duplication.

Metabolic pathways of plants are classified into primary metabolism and specialized metabolism (also referred to secondary metabolism) e.g., flavonoids, terpenoids and alkaloids. Specialized metabolites are originally derived from primary metabolites, and a specific metabolic branch is extended via sequential catalysis by newly emerged enzymes on lineage-specific manner. While primary metabolism produces compounds that are indispensable for survival, specialized metabolism produces a wide variety of metabolites which play crucial roles for increase of fitness of the producing plants in interactions between plants and environments (Bourgaud et al. 2001). It has been postulated and exemplified that the genes involved in primary metabolism are evolutionarily conserved and exhibit less variation in the number of genes across taxa (Clegg et al. 1997). Meanwhile, the number of specialized metabolism genes within a plant genome is very large and their compositions are different among taxa (Caputi et al. 2012; Kawai et al. 2014; Nelson and Werck-Reichhart 2011; Yonekura-Sakakibara and Hanada 2011). Important to note, even within cultivars of grapevine (*Vitis vinifera*) for wine, over thousand cultivar specific genes were discovered in an Uruguay cultivar, Tannat (UY11), compared to a French cultivar, Pinot Noir (PN40024) (Da Silva et al. 2013). Furthermore, cultivar-specific genes in Tannat were shown to contribute to the biosynthesis of phenolic and polyphenolic compounds (specialized metabolites) that contribute to the unique characteristics (Da Silva et al. 2013). This example implies the differences in the underlying evolutionary mechanisms between primary metabolism and specialized metabolism, which are expected to have high plasticity in the genomic structures and transcriptional regulation.

It has been shown that plant genomes contain substantial fraction of the genes associated with specialized metabolism, where numerous oxygenation and glycosylation reaction steps impact on structural diversity and solubility of specialized metabolites. Cytochrome P450 monooxygenases (CYPs) and 2-oxoglutarate-dependent dioxygenases (DOXs)

Abbreviations: CYP, cytochrome P450; DOX, 2-Oxoglutarate-dependent dioxygenase; LTD, local tandem duplication; UGT, Family-1 UDP glycosyltransferases; WGD, whole genome duplication.
This article can be found at http://www.jspcmb.jp/
Published online December 6, 2014

catalyze the oxygenation reactions via activation of molecular oxygen. Family-1 UDP-sugar dependent glycosyltransferases (UGTs) catalyze the transfer of a glycosyl moiety from UDP-activated sugars to a wide range of acceptor molecules. Sequential oxidation by CYP and DOX followed by glycosylation by UGT are often observed in various specialized metabolisms (Kawai et al. 2014). Surveys in plant genomes revealed that these three superfamily genes are highly multiple in range from 50 to 400 genes and are significantly diversified in terms of protein sequences and biochemical functions among seed plants (Caputi et al. 2012; Kawai et al. 2014; Nelson and Werck-Reichhart 2011; Yonekura-Sakakibara and Hanada 2011). It should be noted that lineage-specific gene clusters coordinately involved in certain specialized metabolisms were also found in several species (Chae et al. 2014; Fukushima et al. 2011; Nützmann and Osbourn 2014; Ono et al. 2010). In addition, a gene cluster in Solanaceae, which contains CYP, DOX, and UGT in that order, have been recently identified to be coordinately involved in the biosynthesis of steroidal glycoalkaloids (Itkin et al. 2013).

We investigated the number of specialized metabolism genes (CYP, DOX and UGT) from *Arabidopsis thaliana* (Swarbreck et al. 2008), potato (*Solanum tuberosum*) (The Potato Genome Sequencing Consortium 2011), grapevine (*Vitis vinifera*) (Jaillon et al. 2007), soybean (*Glycine max*) (Schmutz et al. 2010) and rice (*Oryza sativa*) (Ouyang et al. 2007), which are not only agronomically important crops but also represent the major taxonomic group of angiosperms. Lycophyte (*Selaginella moellendorffii*) (Banks et al. 2011) was included as an outgroup species. The amino acid sequences of CYP and UGT genes of *A. thaliana* are retrieved from Arabidopsis P450 database (http://www.p450.kvl.dk/index.shtml) (Paquette et al. 2000). We used the amino acid sequences of DOX from *A. thaliana*, *O. sativa* and *S. moellendorffii* identified in the previous study (Kawai et al. 2014), and we focus on the DOXC class in the DOX genes, which is associated with specialized metabolism but not the other classes (DOXA and DOXB), which are mainly associated with primary metabolism. In order to identify CYP, DOX and UGT genes of other species, we retrieved the amino acid sequences, which contain the sequence motif characterized by Pfam database (Punta et al., 2011) from whole amino acid sequences of each species. Pfam motifs used to detect CYP, DOX, and UGT proteins are p450 motif (Pfam ID: PF00067), 2OG-FeII_Oxy motif (PF03171) and DIOX_N motif (PF14226), and UDPGT motif (PF00201), respectively. The whole amino acid sequences are retrieved from PHYTOZOME9.0 (Goodstein et al. 2012) and the Pfam motif searches were carried out by using HMMER3.0 (Eddy 2011).

The total numbers of the superfamily genes were variable among CYP, DOX and UGT as well as among species (Table 1). However, the numbers of CYP genes were invariably higher than those of DOX and UGT genes for all species and the numbers of UGT genes were higher than DOX genes for all species except for *V. vinifera*. Comparing the number of these genes among species revealed that *G. max* and *S. tuberosum* were outstanding in the number of these genes. These results indicate that the lineage specific expansion of the specialized metabolism genes have occurred after divergence of angiosperm lineage from ancestral vascular plants.

To further investigate the evolutionary changes of the number of specialized metabolism genes in the ancestral lineages of angiosperm, we reconstructed phylogenetic tree of CYP, DOX and UGT genes. The gain and loss events in each lineage of these trees were inferred by the reconciled tree method using Notung 2.8 (Stolzer et al. 2012). Because the amino acid sequences of CYP, DOX and UGT are too diverse to obtain reliable sequence alignments, we first classified CYP, DOX and UGT sequences into orthologous groups by the OrthoMCL method (Fischer et al. 2011). This clustering procedure resulted in 173, 101 and 108 orthologous groups, each of which contains at least two sequences, for CYP, DOX and UGT genes, respectively. The genes that did not cluster with other genes were excluded from the further analysis. The multiple sequence alignment of each group was carried out by MAFFT (Katoh and Standley 2013) with default settings. The phylogenetic trees of the orthologous groups, which contain more than 4 sequences, were reconstructed by maximum likelihood method by RAxML version 7.2.6 (Stamatakis 2006). An LG amino acid substitution matrix with gamma model rate heterogeneity and empirical amino acid frequencies was used for the analysis. Statistical support for the nodes on the maximum likelihood tree was evaluated by

Table 1. The number of specialized metabolism genes.

Species	CYP	DOX	UGT	Total number of genes in a genome
S. moellendorffii	293	57	143	22,273
O. sativa	355	89	200	39,049
G. max	444	236	250	56,044
V. vinifera	323	115	102	26,346
S. tuberosum	412	141	253	35,119
A. thaliana	238	100	112	27,416

bootstrap analysis with 100 replicates. The phylogenetic trees of each orthologous group were subjected by reconciled method with the species tree of plants used in this study. The phylogenetic relationships represented by the species tree are followed by APG III classification system (Bremer et al. 2009). Each reconciled tree yields the number of loss and gain events in each lineage for respective orthologous group. Then, these numbers of gain and loss events observed in each branch were summed per gene.

Figure 1 shows the total numbers of the gain and loss events on every lineage of species tree. The gene numbers in the common ancestor between angiosperms and lycophyte (CA1 in Figure 1) were 28, 12, and 8 for CYP, DOX, and UGT, respectively. These are likely involved in core reactions/pathways conserved in vascular plants, such as the biosynthesis of biopolymers, defense chemicals, and phytohormones (Kawai et al. 2014; Mizutani and Ohta 2010; Yonekura-Sakakibara 2009). Clear distinction in the number of gain and loss events was observed among ancestral lineages. For example, the excesses in the gain events over loss events are significant in the lineage leading to *G. max* for CYP (267 gains and 49 losses), DOX (149 gains and 19 losses), and UGT (177 gains and 26 losses) genes. Conversely, the significant excesses in the loss events were inferred at the branch leading to common ancestor between *A. thaliana* and *G. max* (from CA4 to CA5). In this branch, 1 gain and 76 loss events, 1 gain and 25 loss events, 2 gain and 26 loss events were inferred for CYP, DOX and UGT genes, respectively. Intriguingly, the patterns of the changes in the gene number are similar among CYP, DOX and UGT genes on the same branch. In other words, the changes in the number of the superfamily genes are highly correlated among CYP, DOX and UGT genes. Whole genome duplication (WGD) rather than local tandem

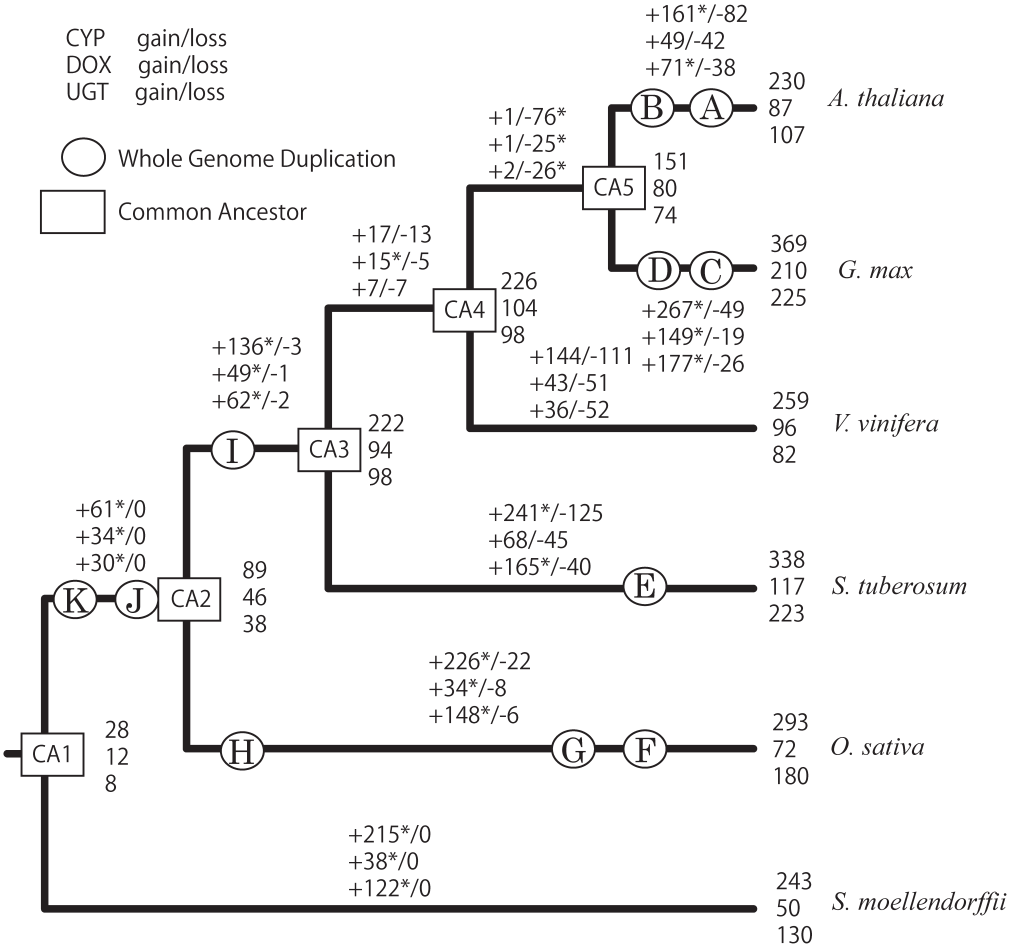

Figure 1. Ancestral gain and loss events of CYP, DOX and UGT genes in plants. The numbers of gain and loss events are indicated on braches of phylogenetic tree of plants for CYP (top), DOX (middle) and UGT (bottom). Numerator and denominator represent the numbers of gain and loss events, respectively. Asterisks indicate the 1% significance levels by binomial test. The whole genome duplication (WGD) events that have been reported in the literature are indicated by circles. The references of WGD events are following: A and B; (Bowers et al. 2003), C; (Doyle and Egan 2010), D; (Schlueter et al. 2008), E; (The Tomato Genome Consortium 2012), F; (Yu et al. 2005), G and H; (Tang et al. 2010), I; (Jaillon et al. 2007), J and K; (Jiao et al. 2011). The branch length of tree and the relative positions of WGD events on branch are not scaled. The common ancestors (CA1–CA5) are designated by rectangles. Because the genes that do not cluster with other genes were excluded, the gene numbers in extant species in this figure differ from those indicated in Table 1.

duplication (LTD), which is also known as small-scale duplication (SSD) (Tamate et al. 2014), is more plausible for explanation of this phenomenon because WGD significantly increases the gene number across a whole genome, whereas LTD partly affects the gene number in a genome for duplication of restricted region of a genome. Indeed, excessive gene gain events, concentrated on particular branches, were shown to be correlated with WGD events (Figure 1). For instance, the increases in the gene numbers are outstanding in the ancestral lineages of *A. thaliana* and *G. max*. The whole genome studies supported two successive WGD events in ancestral lineages of *A. thaliana* (Bowers et al. 2003) and *G. max* (Doyle and Egan 2010; Schlueter et al. 2008). These WGD events are indicated by A and B for *A. thaliana* and C and D for *G. max* in Figure 1. Rice has experienced a whole genome triplication (H in Figure 1) and two whole genome duplication (F and G) events in the ancestral lineage (Tang et al. 2010; Yu et al. 2005). There is a significant increase in the gene numbers (+226/−22, +34/−8, and +148/−6 for CYP, DOX, and UGT, respectively) in this lineage. The whole genome study of *V. vinifera* revealed the triplication event has occurred on the common ancestral lineage of eudicot (I in Figure 1) (Jaillon et al. 2007). Again, there are significant excess in the gain events (+136/−3, +49/−1, and +62/−2 for CYP, DOX, and UGT, respectively) in this lineage. Collectively, these results support the notion that gene resources for evolution of specialized metabolism had been enlarged via WGDs during evolution of angiosperm in lineage-specific manner. The ratios of the gene of three superfamilies were roughly kept across ancient lineages of angiosperm and lycophyte, also supporting this notion (Figure 2). Obviously, CYP is the most prevalent gene in both ancestral (CA1–CA5 in Figure 1) and extant species among three superfamilies. For instance, the number of CYP gene in the most common ancestor of angiosperm (CA2) is estimated to be 89 whereas DOX and UGT are 48 and 38, respectively. Thus, it is likely that CYP has been diversified earlier than DOX and UGT. Although we hypothesized that the increases in the number of the superfamily genes are mainly due to WGS events, LTD and pseudogenization are also responsible for the gene number variation of specialized metabolism genes (Figure 3). Indeed, small differences in the ratio observed among lineages would be reflected to lineage-specific LTD (Figure 2), mainly associated with specialized metabolisms, after WGD (Chae et al. 2014).

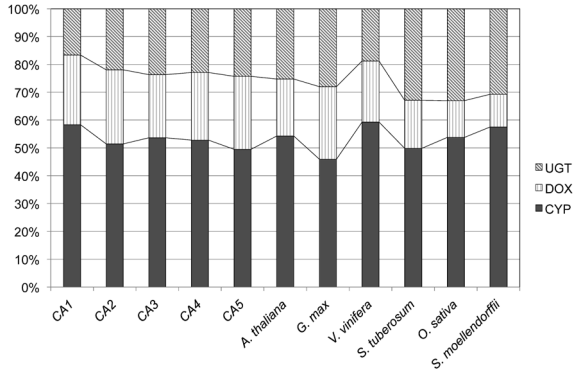

Figure 2. Ratios of the number of CYP, DOX and UGT genes within common ancestral and extant species. The ratios of CYP, DOX and UGT genes are presented by the common ancestral species (CA1–5) and six extant species. The phylogenetic positions of CA1-5 are indicated in Figure 1.

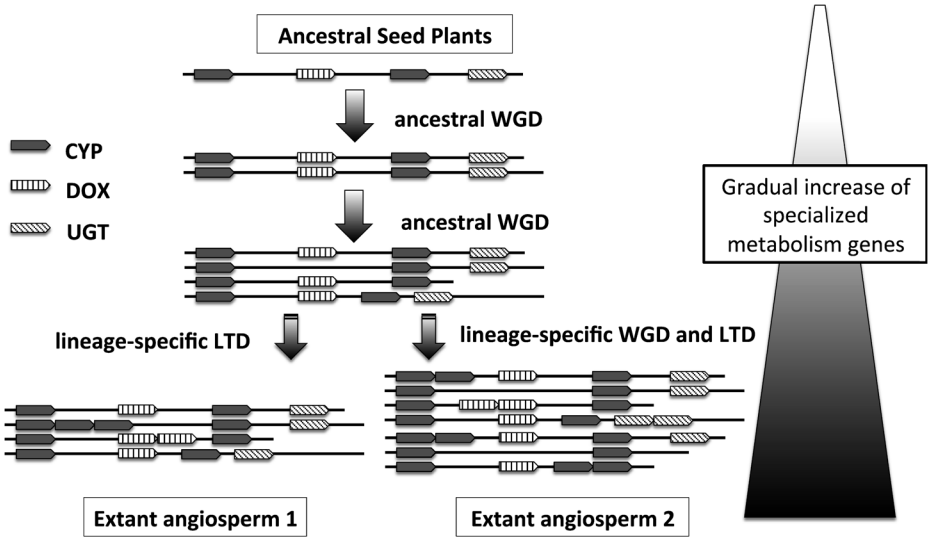

Figure 3. Schematic illustration of evolutionary perspective of specialized metabolisms via WGDs and LTDs in angiosperms. A possible evolutionary scenario of the diversification of specialized metabolism gene by the whole genome duplication (WGD) and local tandem duplication (LTD) is presented. Arrows indicate the hypothetical specialized metabolism genes on chromosomes of ancestral and extant plant. In this hypothetical scenario, two subsequent ancestral WGD followed by the recent lineage specific WGD and LTD events are represented. Note that some duplicated genes are also lost after WGD events.

We demonstrated the dynamics of the evolutionary changes in the numbers of specialized metabolism genes during the evolution of angiosperms. It is most likely that the dramatic expansion of specialized metabolism genes is due to an adaptation of land plants to a particular environment. Here we showed that the duplications of the specialized metabolism genes roughly coincide with WGD events, suggesting that the ancestral WGD events are important in terms of enlargement of genetic resources for neofunctionalization of novel specialized metabolism. We also found that the ratios of CYP, DOX and UGT genes are invariable among both extant and ancestral species; CYP is most prevalent followed by UGT then DOX. This observation suggests a notion that these genes are associated with each other in their enzymatic function and they have evolved in a coordinated manner. Sequential oxidation by CYP and DOX followed by glycosylation by UGT are often seen in pathways of specialized metabolisms (Kawai et al. 2014), also support this evolutionary notion on specialized metabolism.

Acknowledgement

We thank National Institute of Genetics (NIG) for providing super computer resource. This study was supported in part by the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry (BRAIN).

References

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M, DePamphilis C, Albert VA, Aono N, Aoyama T, Ambrose BA, et al. (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332: 960–963

Bourgaud F, Gravot A, Milesi S, Gontier E (2001) Production of plant secondary metabolites: A historical perspective. *Plant Sci* 161: 839–851

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433–438

Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Bailey LH, Soltis DE, Soltis PS, Stevens PF (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* 161: 105–121

Caputi L, Malnoy M, Goremykin V, Nikiforova S, Martens S (2012) A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. *Plant J* 69: 1030–1042

Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic signatures of specialized metabolism in plants. *Science* 344: 510–513

Clegg MT, Cummings MP, Durbin ML (1997) The evolution of plant nuclear genes. *Proc Natl Acad Sci USA* 94: 7791–7798

Da Silva C, Zamperin G, Ferrarini A, Minio A, Dal Molin A, Venturini L, Buson G, Tononi P, Avanzato C, Zago E, et al. (2013) The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* 25: 4777–4788

Doyle JJ, Egan AN (2010) Dating the origins of polyploidy events. *New Phytol* 186: 73–85

Eddy SR (2011) Accelerated Profile HMM Searches. *PLOS Comput Biol* 7: e1002195

Fischer S, Brunk BP, Chen F, Gao X, Harb OS, Iodice JB, Shanmugam D, Roos DS, Stoeckert CJ (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr Protoc Bioinformatics Chapter 6: Unit 6 12: 1–19

Fukushima EO, Seki H, Ohyama K, Ono E, Umemoto N, Mizutani M, Saito K, Muranaka T (2011) CYP716A subfamily members are multifunctional oxidases in triterpenoid biosynthesis. *Plant Cell Physiol* 52: 2050–2061

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. (2012) Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* 40(D1): D1178–D1186

Itkin M, Heinig U, Tzfadi O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, et al. (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* 341: 175–179

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–467

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30: 772–780

Kawai Y, Ono E, Mizutani M (2014) Evolution and diversity of the 2-oxoglutarate-dependent dioxygenase superfamily in plants. *Plant J* 78: 328–343

Mizutani M, Ohta D (2010) Diversification of P450 genes during land plant evolution. *Annu Rev Plant Biol* 61: 291–315

Nelson D, Werck-Reichhart D (2011) A P450-centric view of plant evolution. *Plant J* 66: 194–211

Nützmann H-W, Osbourn A (2014) Gene clustering in plant specialized metabolism. *Curr Opin Biotechnol* 26: 91–99

Ono E, Homma Y, Horikawa M, Kunikane-Doi S, Imai H, Takahashi S, Kawai Y, Ishiguro M, Fukui Y, Nakayama T (2010) Functional differentiation of the glycosyltransferases that contribute to the chemical diversity of bioactive flavonol glycosides in grapevines (*Vitis vinifera*). *Plant Cell* 22: 2856–2871

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al. (2007) The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res* 35(Database): D883–D887

Paquette SM, Bak S, Feyereisen R (2000) Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol* 19: 307–317

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al. (2011) The Pfam protein families database. *Nucleic Acids Res* 40(D1): D290–D301

Schlueter JA, Scheffler BE, Jackson S, Shoemaker RC (2008) Fractionation of synteny in a genomic region containing tandemly duplicated genes across Glycine max, Medicago truncatula, and *Arabidopsis thaliana*. *J Hered* 99: 390–395

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W,

Hyten DL, Song Q, Thelen JJ, Cheng J, et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183

Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690

Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28: i409–i415

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. (2008) The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res* 36(Database): D1009–D1014

Tamate SC, Kawata M, Makino T (2014) Contribution of nonohnologous duplicated genes to high habitat variability in mammals. *Mol Biol Evol* 31: 1779–1186

Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci USA* 107: 472–477

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475: 189–195

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–641

Yonekura-Sakakibara K (2009) Functional genomics of family 1 glycosyltransferases in Arabidopsis. *Plant Biotechnol* 26: 267–274

Yonekura-Sakakibara K, Hanada K (2011) An evolutionary view of functional diversity in family 1 glycosyltransferases. *Plant J* 66: 182–193

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, et al. (2005) The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol* 3: 266–281