Construction of a Rice Glycosyltransferase Phylogenomic Database and Identification of Rice-Diverged Glycosyltransferases

Pei-Jian Cao¹, Laura E. Bartley¹, Ki-Hong Jung¹ and Pamela C. Ronald¹

a Department of Plant Pathology, University of California, Davis, CA 95616, and the Joint BioEnergy Institute, Emeryville, CA 94710, USA
b Institute of Bioinformatics, Zhejiang University, Hangzhou, Zhejiang 310029, China

ABSTRACT Glycosyltransferases (GTs; EC 2.4.x.y) constitute a large group of enzymes that form glycosidic bonds through transfer of sugars from activated donor molecules to acceptor molecules. GTs are critical to the biosynthesis of plant cell walls, among other diverse functions. Based on the Carbohydrate-Active enZymes (CAZy) database and sequence similarity searches, we have identified 609 potential GT genes (loci) corresponding to 769 transcripts (gene models) in rice (Oryza sativa), the reference monocotyledonous species. Using domain composition and sequence similarity, these rice GTs were classified into 40 CAZy families plus an additional unknown class. We found that two Pfam domains of unknown function, PF04577 and PF04646, are associated with GT families GT61 and GT31, respectively. To facilitate functional analysis of this important and large gene family, we created a phylogenomic Rice GT Database (http://ricephylogenomics.ucdavis.edu/cellwalls/gt/). Through the database, several classes of functional genomic data, including mutant lines and gene expression data, can be displayed for each rice GT in the context of a phylogenetic tree, allowing for comparative analysis both within and between GT families. Comprehensive digital expression analysis of public gene expression data revealed that most (~80%) rice GTs are expressed. Based on analysis with Inparanoid, we identified 282 ‘rice-diverged’ GTs that lack orthologs in sequenced dicots (Arabidopsis thaliana, Populus trichocarpa, Medicago truncatula, and Ricinus communis). Combining these analyses, we identified 33 rice-diverged GT genes (45 gene models) that are highly expressed in above-ground, vegetative tissues. From the literature and this analysis, 21 of these loci are excellent targets for functional examination toward understanding and manipulating grass cell wall qualities. Study of the remainder may reveal aspects of hormone and protein metabolism that are critical for rice biology. This list of 33 genes and the Rice GT Database will facilitate the study of GTs and cell wall synthesis in rice and other plants.

Key words: glycosyltransferases; GT; phylogenetic tree; orthologs; expression pattern; mutant lines; phylogenomic database; rice; cell wall.

INTRODUCTION

Glycosyltransferases (GTs; EC 2.4.x.y), which add sugars onto acceptor molecules, are involved in numerous biological processes. Some GT functions are conserved across eukaryotes, such as protein sorting in the secretory pathway and metabolic regulation (Gomez et al., 2006; Henquet et al., 2008). Other functions have diverged to fill plant-specific roles, including synthesis of diverse secondary metabolites, modification of hormones, and cell wall synthesis (Bowles et al., 2005; Farrokh et al., 2006). Because of the potential of biofuels made from the lignocellulose of plant cell walls to serve as an alternative energy source, the mechanism of cell wall synthesis is a crucial research topic. A promising source of sustainable biofuel production is conversion to sugars of the large, untapped resource (several billion tons per year) of plant lignocellulosic biomass for use in fermentation of liquid fuels. Due to its abundance and high rate of production, lignocellulosic biomass has advantages as a biofuel feedstock compared with currently used starch and cane sugar; however, these advantages are overcome by the current expense of extracting sugars from cell walls. Understanding and manipulating cell wall synthesis may facilitate the use of lignocellulose for biofuel production. In
addition, elucidation of the functions of non-cell wall synthesizing GTs may illuminate other essential and practically useful roles in plant biology for this large class of enzymes.

Plant cell walls are a complex and dynamic extracellular matrix that regulates cell growth, provides plants with mechanical support and protects against pathogens (Carpita, 1996). Primary cell walls are composed of cellulose microfibrils embedded in a semi-structured matrix of non-cellulosic polysaccharides (Carpita et al., 2001; Somerville et al., 2004). As plant cells age and cease to grow, secondary cell walls, in which the cellulose matrix becomes denser and typically cross-linked by a phenyl propanoid-derived lignin meshwork form. There are two major classes of cell walls in plants—type I and type II, which differ in architecture, chemical composition, and their associated biosynthetic processes (Carpita, 1996). Type I walls are found in dicotyledonous plants. In type I primary walls, cellulose microfibrils are interwoven with xyloglucans and embedded in a matrix of pectin polysaccharides and glycoproteins. As type I primary walls transition to secondary walls, glucuronoxylan and mannan also accumulate (Pauly and Keegstra, 2008). Type II walls are characteristic of Commelinoid monocots, including grasses such as rice and switchgrass. In such walls, glucuronoxylans and mixed linkage glucan form the meshwork surrounding cellulose microfibrils. Though present, type II walls possess a lower abundance of pectin polysaccharides, xyloglucan, and structural proteins relative to type I walls (Carpita, 1996). In type II walls, the transition to secondary wall is marked by lignin accumulation.

GTs known or hypothesized to be involved in cell wall synthesis and other diverse biological processes have been hierarchically classified based on the following criteria: three-dimensional structure, catalytic reaction mechanism, and their donor and acceptor substrates (Coutinho et al., 2003). At the tertiary structure level, GTs adopt one of the following two major folds: the GT-A (SpsA and SpsA-like) fold or the GT-B (B-GT and B-GT-like) fold (Bourne and Henrissat, 2001; Hu and Walker, 2002; Wimmerova et al., 2003; Breton et al., 2006). Recently, researchers have identified a new GT fold, GT-C, adopted by CstII, a Campylobacter jejuni sialyltransferase (Chiu et al., 2004) and Pyrococcus furiosus oligosaccharide/transferase (OST) (Igura et al., 2008). The fold class of a large number of GTs has not been determined experimentally and is designated ‘GT-U’ for unknown fold. Nevertheless, modeling suggests that most of these GTs are in fact adopting either the GT-A or GT-B fold (Breton et al., 2006). At the catalytic reaction level, glycosylation proceeds via two reaction mechanisms—inversion or retention of stereochemistry at the C1 position of the donor sugar. Beyond these general criteria, GTs have traditionally been grouped into families based on their activated sugar substrate (e.g. galactosyltransferases, sialyltransferases, etc.) and, in many cases, the acceptor group (e.g. protein, lipid, glycolgen, etc.). However, sequence data have far outpaced our ability to identify the biochemical activity of enzymes. This led to the creation of the Carbohydrate-Active enZymes (CAZy; www.cazy.org/) database to build on the biochemical data by developing a hierarchical family classification scheme for grouping GTs at the primary sequence level (Campbell et al., 1997; Coutinho et al., 2003). As of February 2008, CAZy contained 33 359 GTs from organisms across all the kingdoms of life classified into 90 different GT families, primarily based on amino acid sequence similarity.

While much remains to be learned, the last decade has seen a great expansion in our understanding of the GTs that synthesize the major constituents of type I primary walls. Use of diverse plant species and the reference dicot, Arabidopsis (Arabidopsis thaliana), has led to the identification of many genes involved in synthesis of xyloglucan, mannan, and pectin (Farrokhi et al., 2006). In contrast, our depth of knowledge regarding synthesis of type II wall-enriched polysaccharides components has lagged behind. The Cellulose synthase-like F (CslF) gene family has been shown to have a role in synthesis of mixed linkage glucan (Burton et al., 2006), but the synthesis of glucuronoxarabinoylan in grasses remains obscure. Progress may be possible based on emerging studies of glucuronoxylan in Arabidopsis secondary cell walls (Lee et al., 2007a, 2007b; Pena et al., 2007; York and O’Neill, 2008). However, the surprisingly complex view that those studies provide remains to be tested for grass primary cell walls.

There is now an opportunity to apply the genomic resources that have accumulated for grasses toward understanding the synthesis of type II cell walls. Rice serves as a reference monocot species because of its small, sequenced genome (~389 Mb) and the availability of genetic and molecular resources, including indexed insertion mutants (IRGSP, 2005; Jung et al., 2008a). Rice itself is also a potentially attractive biofuel feedstock source because it comprises a large portion of the world’s agricultural residue (Sticklen, 2008). Due to a high level of genomic colinearity among grass species (Devis and Gale, 2000), information learned regarding rice cell walls is likely to apply to the cell walls of the other major cereal crops, such as maize (Zea mays) and wheat (Triticum aestivum), and potential dedicated energy crops, such as switchgrass (Panicum virgatum) and Miscanthus. Conversely, extensive cell wall biochemical and physiological studies on maize, wheat, and barley are also likely to apply to rice.

One of the challenges to discovery of cell wall gene function in Arabidopsis has been genetic redundancy, such that single gene mutants have no measurable phenotype. For example, Richmond and Somerville (2001) reported that a number of single CslI gene mutants provide no phenotype. This challenge is likely to be exacerbated in grasses, which possess a larger gene complement compared with Arabidopsis. Estimates for the percent of the rice genome that consists of segmental duplication vary from 27 to 66%, depending on the method of detection used; however, consensus leans towards a higher value of ~50% (Ouyang et al., 2007; Yu et al., 2005). Thus, the rice genome encodes a large number of genes likely to possess redundant functions, creating a considerable challenge to the functional analysis of individual genes (Jung et al., 2008a, 2008b). This is also the case for GTs, for which we anticipate...
that the large number of members in some families, especially those associated with cell wall synthesis, will create difficulties for functional analysis. Incorporating diverse systems biological databases, including bioinformatic, genomic, gene expression, and proteomic data, can inform rational strategies for gene function discovery, even in large gene families. However, these approaches are hampered by current database formats that display only a single gene or field at a time, preventing simultaneous comparisons of multiple datasets and multi-gene families (Jung et al., 2008a). Scattering of genomic data across multiple databases, exacerbated by different gene nomenclatures and data formats, creates additional challenges to integration. The field of phylogenomics, which merges phylogenetics and genomics and puts genomic data in a phylogenetic context, helps us to resolve these limitations. A successful application of phylogenomics for a family of genes for which redundancy poses enormous challenges is the Rice Kinase Database (http://rkd.ucdavis.edu/), which provides a template for the design of new phylogenomic databases (Dardick et al., 2007).

Genomic and transcriptomic analyses of glycosyltransferase genes have been conducted for two dicot species, Arabidopsis and poplar (Populus trichocarpa), toward identifying the functions of GTs throughout development (Geisler-Lee et al., 2006; Henrisatt et al., 2001). For rice, until now, the available genomic data have not been extensively utilized. The Carbohydrate Active enZymes (CAZy) database contains only GT family classification and sequence information for the genes of rice and other species. Yokoyama and Nishitani (2004) compared the numbers and phylogenetic relationships of known cell wall-related genes between rice and Arabidopsis soon after the publication of the rice genome. However, that analysis focused only on genes known at the time to be involved in cell wall synthesis, including six GT families, and did not include other types of analyses. Mitchell et al. (2007) conducted a much more extensive comparison between grass and dicot GTs and other gene families. This effort examined expressed sequence tag (EST) abundance for orthologous gene groups from Arabidopsis and rice, leading to the identification of several gene families that are more abundantly expressed in grasses than dicots. These genes, including members of the GT47 and GT61 families, are good candidates for involvement in synthesis of glucuronoarabinoxylan and other type II wall-specific or enriched components (Mitchell et al., 2007). The work here provides a complementary approach by examining all rice GTs, not only those with Arabidopsis orthologs, by considering more classes of data, and by developing a database for public use.

With completion of rice (Oryza sativa ssp. japonica) genome sequencing and deposition of a large number of GTs in the CAZy database, we now have the opportunity to identify all the rice GTs and analyze them on a whole-genome scale (IRGSP, 2005). In this study, we identified 609 rice GT loci (769 gene models) and executed a set of genome-scale analyses on these GTs. We used the data to identify proteins that have diverged significantly compared with dicot GTs and that may contribute to the synthesis of type II-specific cell wall components or be responsible for more subtle divergences in grass versus dicot structure and function (e.g. different functions throughout development in type I versus type II walls). We also report construction of a phylogenomic database for rice GTs (http://rice-phylogenomics.ucdavis.edu/cellwalls/gt/), which provides a logical format to integrate, host, and display diverse sets of functional genomic information in a phylogenetic context, thereby facilitating plant GT research. Using the database, we identified 33 rice-diverged GT genes (45 gene models) that are highly expressed in vegetative, above-ground tissues. Twenty-one of these (25 gene models) are strong candidates for further functional analysis toward understanding and manipulating grass cell walls for biofuel production.

**RESULTS AND DISCUSSION**

**Identification of Rice GTs**

Glycosyltransferases from across the kingdoms of life have previously been identified based upon domain compositions, sequence similarity and function. The CAZy database is a comprehensive database for carbohydrate enzymes that degrade, modify, or create glycosidic bonds. CAZy classifies GTs into different families primarily based on amino acid sequence similarities (Campbell et al., 1997; Coutinho et al., 2003). As of February 2008, there were 90 GT families and 33,359 entries in the CAZy database. We identified a total of 548 rice GT genes (loci) from this database. We then converted the Rice Annotation Project and other various identifiers associated with the rice GTs from CAZy into The Institute of Genomic Research (TIGR) Version 5 Locus Identifiers (IDs) for convenience in the further analysis. Not all GTs are included in the CAZy database (Egelund et al., 2004). Thus, we took advantage of the availability of the complete rice genome sequence to identify GT genes not included in the CAZy database. Searching of the rice genome with the known GTs from rice and Arabidopsis led to the identification of an additional 34 GTs by homolog search, 12 GTs by domain search and 15 GTs by paralog search (see Methods).

In total, 609 rice GT genes were identified in our analysis and classified into 40 CAZy families and an additional unknown class. One hundred and seven of the rice GT genes are predicted to code for 160 additional alternative splicing isoforms, resulting in a total of 769 GT transcripts (gene models) encoded in the rice genome. The TIGR Version 5 ID, CAZy family assignment, protein length, chromosome position, EST or full-length cDNA (FL-cDNA) support, identification method, GT-related domain and TIGR Version 5 annotation for each of these rice GT gene models are listed in Supplemental Table 1. The 609 rice GT loci are distributed across all 12 rice chromosomes in proportion to the size of each chromosome. The maximum number are present on the largest chromosome, Chr 1 (85), and the minimum on chromosome 11 (16) (Supplemental Figure 1). BLASTP searching with these 769 GT proteins in the
domains according to this observation. Pfam database, they should be considered as GT-related (PF04577 and PF04646) is unknown according to the current PF01762. Although the function of these two domains other GT31 members contain a galactosyl transferase domain, of 58 GT31 members contain PF04646 domain, whereas the family members contain Pfam domain PF04577. Fourteen out associated with these rice GT families, respectively. All 39 GT61 domains of unknown function, PF04577 and PF04646, are as- are annotated as GT-related. Rather, we found that two Pfam members of the GT31 family also do not possess domains that included rice GT77 proteins in the list. GT61 proteins and some lacturonan II (Egelund et al., 2004, 2006). We therefore fold recognition and then experimentally validated to be xylo- syltransferases involved in synthesis of the pectin, rhamnogalacturonan II (Egelund et al., 2004, 2006). We therefore included rice GT77 proteins in the list. GT61 proteins and some members of the GT31 family also do not possess domains that are annotated as GT-related. Rather, we found that two Pfam domains of unknown function, PF04577 and PF04646, are associated with these rice GT families, respectively. All 39 GT61 family members contain Pfam domain PF04577. Fourteen out of 58 GT31 members contain PF04646 domain, whereas the other GT31 members contain a galactosyl transferase domain, PF01762. Although the function of these two domains (PF04577 and PF04646) is unknown according to the current Pfam database, they should be considered as GT-related domains according to this observation.

**Database Construction and Navigation**

Though the GT section of the CAZy database is reasonably comprehensive in scope, it lacks depth of information on each GT, limiting further functional and reverse genetic analyses of this large gene family. Several kinds of functional genomic data are now available, such as expression data from expressed sequence tags (ESTs), massively parallel signature sequencing (MPSS), and oligonucleotide microarrays. However, these data are scattered in different databases and are not easily inte- grated for comparison between and within different rice GT families. To resolve this problem, we created a publicly accessible,phylogenomic database—the Rice GT Database (http://ricephylogenomics.ucdavis.edu/cellwalls/gt)—to integrate, host, and display functional genomic data for rice GTs in a phylo- genetic context. As listed in Table 1, we gathered eight types of functional genomic data for rice GTs for inclusion in the database, including sequence and ortholog information, mutant availability, protein topology predictions, and gene expression data. Further information about the development and content of the Rice GT Database is provided in subsequent sections.

To assist in use of the Rice GT Database, links to the database information, database search, chromosome distribution map, and phylogenetic tree viewer are provided on the home page. Most data are available in the context of a phylogenetic tree of rice GTs to aid comparisons within and between GT clades. In the tree viewer page, functional genomic fields can be selected by checking each box (Figure 1). Pressing the submit button will display the selected data adjacent to the GT phylogenetic tree (Figure 2). The spreadsheet format allows data to be readily transferred into any database or software, such as Excel, for further analysis. Once displayed, the spreadsheet can be searched for a particular locus ID or other field with the browser’s search function. Clicking on a gene model ID (12XXX.mXXXX) link brings up a summary webpage for that gene model, showing all available data except for the micro- array data, but including histogram representations of expres- sion patterns of EST and MPSS counts (Figure 3). Links to the TIGR rice database, Rice Annotation Project Database (RAP- DB), CAZy database, and NCBI BLAST search are given for easy navigation. These links allow for simple navigation between all data display formats as well as complementary databases. Mutant line identification numbers are given as hyperlinks to the corresponding library when phenotypic information is

---

**Table 1. Data Available in the Rice GT Database.**

<table>
<thead>
<tr>
<th>Data type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence information</td>
<td>TIGR and RAP annotations, CAZy families, GT domains, and NCBI BLAST links</td>
</tr>
<tr>
<td>Sequence quality</td>
<td>FL-cDNA/EST evidence, BAC/PAC, and PASA status</td>
</tr>
<tr>
<td>Orthologs in dicots</td>
<td>Orthologs identified in four selected dicot species using Inparanoid2</td>
</tr>
<tr>
<td>Mutants</td>
<td>Knockout and activation mutant lines from several mutant libraries</td>
</tr>
<tr>
<td>Topology</td>
<td>Predicted protein topology (i.e. transmembrane domains) and sub-cellular localization</td>
</tr>
<tr>
<td>MPSS data</td>
<td>MPSS data determining the representation of transcripts within mRNA and regulatory small RNA</td>
</tr>
<tr>
<td>Digital northern data</td>
<td>Number of EST within different rice tissues/organs from TIGR database</td>
</tr>
<tr>
<td>Microarray data</td>
<td>There are currently three kinds of microarray platforms available in the database: Affymetrix, NSF 20K, and BGI/Yale. Several hundred hybridizations are presented and heatmaps are also provided for easy visualization.</td>
</tr>
</tbody>
</table>
available for that mutant. For display of microarray data, users may toggle between displaying numerical values for each replicate or averages for each sample. In separate links, we also provide red–green heatmaps for the easy examination of each microarray dataset. The chromosome distribution map is color-coded according to the different CAZy families and rice GT loci are represented as colored boxes (Supplemental Figure 1). Mousing over each box generates a pop-up showing the ID of each rice GT locus. Clicking on the box directs the user back to the tree viewer page, with the selected rice GT at the top of the view window. A search function is also available, enabling users to search the database with a locus ID or protein sequence.

Phylogenetic Analysis

Phylogenetic trees display genes in groups based on sequence similarity and are particularly valuable when studying large gene families (Jung et al., 2008a). Sensitive sequence-similarity detection methods such as hydrophobic cluster analysis or PSI-BLAST have revealed only very low sequence similarities between some GT families (Campbell et al., 1997; Wrabl and Grishin, 2001). These distant similarities, presumably a result of evolutionary divergence, make it difficult to construct a single phylogenetic tree using proteins from all the different GT families. Rather, we adopted the hierarchical classification approach presented by Coutinho et al. (2003) to build an assembled, whole phylogenetic tree.

The 40 GT families are hierarchically classified based on their GT fold type, reaction mechanism, and known activities (Figure 4). There are two different GT domains in the GT2, GT28, and GT31 families, so these families were divided into two subfamilies according to GT domain. Using the neighbor-joining method, we then constructed unrooted phylogenetic trees based on GT domain sequences for each GT family or subfamily with more than three members. For GT families with fewer than three members, the phylogenetic relationships among their members were determined manually. Among the families with more than three members, GT77 has no associated Pfam domain; thus, the entire protein sequences were used for phylogenetic analysis. Finally, all the family trees were assembled into a single phylogenetic tree according to the family hierarchical classification. This assembled tree is shown in Supplemental Figure 2 and used for data display in the Rice GT Database.

Interspecies Comparison Identifying Rice-Diverged GTs

In principle, the difference between type I and type II cell wall polysaccharide content might be reflected by qualitative difference in the GT content among reference plant species. To test this hypothesis, we compared the distribution of GT gene models between rice and the two dicots for which GTs have been comprehensively annotated, Arabidopsis and poplar (Populus trichocarpa) (Figure 5). In Arabidopsis, there are 452 GT genes (507 gene models) based on the content of the CAZy database and fold recognition (Egelund et al., 2004). Poplar contains approximately 840 GT gene models, the largest number of genes encoding glycosyltransferases observed among fully sequenced genomes (Geisler-Lee et al.,
Figure 2. Screen Shot of the Topmost Portion of the Rice GT Database Phylogenetic Tree.
A subset of database content, including the TIGR gene model ID, CAZy family, corresponding RAP2 ID and a hyperlink to the corresponding NCBI BLAST search are listed in spreadsheet format adjacent to the tree. This format allows easy and flexible visualization of the data within the context of the tree. Data obtained in the spreadsheet can be searched using the browser’s search function.
The poplar genome annotation is not yet complete on the gene loci level, so we conducted this analysis on the level of gene models, which includes different splice forms from single loci. We found the same GT families in rice, *Arabidopsis*, and poplar (Figure 5). This result is consistent with a previous analysis that found that known cell wall-related GT families are present in both rice and *Arabidopsis* (Yokoyama and Nishitani, 2004). The one apparent exception to representation in all three species is the GT76 family, which is absent from the poplar genome annotation. However, we detected a GT76 member in the poplar genome (E < e^{-100}) with a BLASTP search with the single members of the GT76 family from *Arabidopsis* and rice. Thus, the absence in poplar is likely due to the incomplete poplar genome annotation at the time of GT identification.

Seven GT families appear to have significantly greater representation in the rice genome compared to *Arabidopsis* and poplar. GT families 5, 28, 30, 33, 37, 43, and 61 contain at least two-fold the number of genes in rice versus the two dicots (Figure 5). Three of these seven families are not expected to be involved in cell wall synthesis. GT5s are starch synthases (James et al., 2003); GT28s are lipid (i.e. diacylglycerol) galactosyltransferases (Jarvis et al., 2000); and GT33s participate in transfer of mannose residues to endoplasmic reticulum-associated proteins (O’Reilly et al., 2006). Four families on this list are known or hypothesized to catalyze synthesis of cell wall polysaccharides. GT30s are 3-deoxy-D-manno-2-octulosonic acid transferases, hypothesized to be required for the incorporation of this sugar into plant cell wall pectin (Royo et al., 2000). GT37s include orthologs of the *Arabidopsis* FUT proteins, which possess α-fucosyltransferase activity involved in xyloglucan synthesis (Perrin et al., 1999). The increase in number of GT8, GT31, and GT47, which is also the case for *Arabidopsis* and poplar.
genes in this family in rice compared to Arabidopsis has been noted previously (Yokoyama and Nishitani, 2004). Whether the family possess the same activity in grasses, which possess far lower quantities of xyloglucan, remains to be seen. In the GT43 family, the Arabidopsis gene, IRX9, has recently been implicated in synthesis of xylan in secondary cell walls (Lee et al., 2007a; Pena et al., 2007). Mitchell et al. (2007) also identified genes in the GT43 and GT61 families as having higher expression in cereals than in dicots. This coarse analysis of the number of genes in various GT families begins to suggest gene families for further exploration toward understanding the synthesis of grass cell walls. Subsequent analyses provide further information toward choosing specific genes for reverse genetic analysis.

Although the same GT families are present in rice, Arabidopsis, and poplar, we hypothesized that, within each GT family, ‘rice-diverged’ GTs with significantly different primary sequences compared to dicots may have evolved since the last common ancestor between rice and dicots. Orthology detection (and conversely detection of genes that lack orthologs) is critically important for accurate functional annotation, and has been widely used to facilitate studies on comparative and evolutionary genomics (Chen et al., 2007a). We hypothesize that differences in primary sequence might be a proxy for functional divergence in some cases. The ongoing individual gene duplication events in the rice genome provide redundant genes that can serve as raw materials for genesis of new gene functions (Yu et al., 2005). A large part (about one-third, data not shown) of rice GTs are involved in tandem and segmental duplication events, and substantial clustering of rice GTs is evident on different chromosomes (Supplemental Figure 1). Thus, some rice-diverged GTs may have evolved after duplications, through a process known as neo-functionalization in which duplicated genes obtain novel functions compared to ancestral genes (Zhang, 2003). In support of this approach with respect to cell wall-related genes, phylogenetic analysis of the CSL genes within the GT2 family of rice and Arabidopsis led to the identification of rice-diverged CSL gene families CSLF...
GT families are listed along the x-axis. Number of GT gene models in each species is shown on the y-axis. GT families are listed along the x-axis.

and CslH (Hazen et al., 2002). Subsequent heterologous expression studies have demonstrated that the CslF gene family is involved in synthesis of the type II wall-specific mixed linkage glucan polysaccharide (Burton et al., 2006).

We computationally identified rice-diverged GTs by detecting which rice GTs lack orthologs in sequenced dicots. Several methods can be used for detecting orthologs are now available, including methods such as reciprocal smallest distance (Wall et al., 2003), Inparanoid (Remm et al., 2001), and BLASTP (Altschul et al., 1990), among others. Inparanoid exhibited the best overall performance, with both low false-negative and false-positive rates, in an orthology detection assessment experiment on divergent eukaryotic genomes (Chen et al., 2007a). Inparanoid is an automated method for finding orthologs and ‘in-paralogs’ from two species. It functions by detecting ortholog clusters with two-way best pair-wise matches then it adds related in-paralogs that are predicted to have diverged since speciation (Remm et al., 2001).

We used Inparanoid Version 2.0 to identify rice GT orthologs in the sequenced dicots, Arabidopsis (family: Brassicaceae), poplar (Salicaceae), medick (Medicago truncatula, Fabaceae), and castor bean (Ricinus communis, Euphorbiaceae). The orthologs of rice GTs identified in these selected dicots are listed in Supplemental Table 3. Based on orthology search in these selected dicots, 282 rice GTs (36.7%) lacked orthologs and were therefore considered to be rice-diverged GTs. One hundred and ninety-seven (70%) of these are expressed, and were therefore considered to be rice-diverged GTs. One hundred and ninety-seven (70%) of these are expressed, providing strong indication that most rice GTs are expressed. In contrast, just less than 60% of all TIGR Version 5 gene models have EST evidence (Jung et al., 2008c). Gene expression patterns can inform hypothesis regarding which gene family members are expected to perform distinct or similar roles. Predominance, or higher expression, of one or more gene family member under a particular set of conditions may indicate a role for the predominantly expressed gene(s) in the process under examination. For example, we recently found evidence that gene family members predominantly expressed in the light were more likely to have a role in light responses compared with genes in the same family that were lowly expressed in the light (Jung et al., 2008b). Thus, we sought to further refine the list of rice-diverged genes for reverse genetic analysis using three classes of publicly available transcriptome information: EST, MPSS, and microarray data.

**Digital Expression Analysis**

Phylogenetic trees provide a context to compare the properties of gene family members and identify similarities and differences (Dardick et al., 2007; Jung et al., 2008a). Gene expression patterns can inform hypothesis regarding which gene family members are expected to perform distinct or similar roles. Predominance, or higher expression, of one or more gene family member under a particular set of conditions may indicate a role for the predominantly expressed gene(s) in the process under examination. For example, we recently found evidence that gene family members predominantly expressed in the light were more likely to have a role in light responses compared with genes in the same family that were lowly expressed in the light (Jung et al., 2008b). Thus, we sought to further refine the list of rice-diverged genes for reverse genetic analysis using three classes of publicly available transcriptome information: EST, MPSS, and microarray data.

**EST Analysis**

We analyzed rice EST data using the Rice Gene Expression Anatomy Viewer (http://rice.plantbiology.msu.edu/dnav.shtml), which provides the number of ESTs from different rice tissues mapped onto the TIGR gene models to estimate gene expression levels (Jung et al., 2008c). This analysis revealed that one or more EST has been recorded for 628/769 (81.7%) rice GT gene models (Supplemental Table 4), providing strong indication that most rice GTs are expressed. In contrast, just less than 60% of all TIGR Version 5 gene models have EST evidence (Jung et al., 2008c). The frequency of total ESTs for rice GT gene models varied greatly from 1 to 770, suggesting that the expression levels among rice GTs vary dramatically.

The Gene Expression Anatomy Viewer data cover ESTs isolated from 20 rice tissue sources (anther, callus, endosperm,
flower, immature seed, leaf, mixed tissues, panicle, phloem, pistil, root, root tip, seed, seedling, sheath, shoot, stem, suspension cells, unknown samples, and whole plant), but rice GTs were found to only be expressed in 12 tissues (Table 2). The absence of expression evidence in other tissues may be due to the small number of ESTs sequenced in the libraries from those plant parts. For example, the phloem tissue library only contains eight ESTs. Among the plant materials that show evidence of GT expression, callus has the largest number of expressed GTs (408, 53.1%), followed by shoot (395, 51.4%). In leaf tissue, only 188 (24.4%) rice GTs have EST evidence, although the leaf library has the largest number of ESTs (204 353). Low representation of diverse GTs in leaf tissue may be due to relative cell wall homogeneity in leaves compared with other tissues or temporal regulation of GT expression such that many GTs involved cell wall synthesis during early developmental stages are no longer expressed in more mature tissues. Alternatively, the difference may be due to different coverage of the actual total of expressed genes from different libraries, if leaf libraries are less diverse due to very high expression of a few genes, such as those involved in photosynthesis, for example. Among the 282 rice-diverged GTs, 46 (16.3%) and 104 (36.9%) are expressed in leaf and shoot, respectively.

**MPSS Analysis**

We also extracted information from the Rice MPSS Project (http://mpss.udel.edu/rice/) for each GT gene model (Nobuta et al., 2007). Massively parallel signature sequencing consists of deep, high-throughput sequencing of short segments of expressed transcripts and can provide a sensitive, quantitative measure of gene expression for most genes in the genome (Brenner et al., 2000). Data from 18 MPSS libraries representing 12 different tissues/organs of rice were extracted for 17 base pair (bp)-tag signature libraries. MPSS tags (1 or more) were available for 628 (81.7%) GT gene models, providing further evidence that most rice GTs are expressed. As in the EST data, substantial differences were found in abundance of different rice GT gene models in tags per million (TPM), with expression varying from marginal (1–3 TPM) to strong (>250 TPM) (Supplemental Table 5). The distribution of expressed GT gene models among different MPSS libraries at different expression levels is shown in Supplemental Figure 3. A large percentage (30–50%, depending on the source tissue) of rice GTs are moderately expressed (26–250 TPM), while only a few (1–13%) are highly expressed (>250 TPM).

**Microarray Analysis and Expression Platform Comparison**

In addition to sequence-based expression analysis methods, we also used publicly available data from rice microarrays, which rely on hybridization of transcript-derived sequences to arrayed DNA oligonucleotides. Microarrays allow biologists to measure the transcript amounts for tens of thousands of genes simultaneously, thus providing a high-throughput tool for analyzing gene expression at the whole-genome level.

Four rice whole-genome oligonucleotide arrays have been developed and several hundred datasets from them are available in the public microarray database, NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/). As of July 2008, microarray data from 359 hybridizations, including those from the rice Affymetrix (148), NSF 20K (114), and BGI/Yale platforms (97), are available in a phylogenetic context in the Rice GT Database.

Here, for comparison with the EST and MPSS data, we use the rice Affymetrix microarray dataset of Jain et al. (2007), which profiles expression patterns for different rice tissues and developmental stages. The rice tissues/stages in these data include seedling, root, mature leaf, young leaf, shoot apical meristem (SAM), and various stages of panicle (P1–P6) and seed (S1–S5) development (Jain et al., 2007). Following whole-chip data processing, we extracted and averaged the log2 signal values for the 634 rice GTs represented on the array (Supplemental Table 6). These data are represented as a heatmap in the context of the rice GT phylogenetic tree in Supplemental Figure 4.

Compiling the EST, MPSS, and microarray data for rice GTs provides an opportunity to contrast and compare the data from these diverse platforms. While analyses have been published that compare two expression platforms (Chen et al., 2007b; Fernandes et al., 2002; Liu et al., 2007), rarely have comparisons been made among these three types of expression evidence (Meyers et al., 2004). As is discussed in the previous sections, all of the platforms provide evidence that GTs show diverse expression, from low to moderately high in comparison with the average level of gene expression. Thus, GTs as a group show typical expression, and we expect that analysis for this gene superfamily might be generalized to most of the rice gene complement.

Figure 6 shows the percent of rice GTs that give an expression signal in each tissue with data from more than one

---

**Table 2. Distribution of Expressed GT Gene Models among Different EST Libraries.**

<table>
<thead>
<tr>
<th>EST library source</th>
<th>No. of ESTs</th>
<th>No. of expressed gene models</th>
<th>No. of expressed GT gene models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total ESTs</td>
<td>33 807</td>
<td>628</td>
<td></td>
</tr>
<tr>
<td>Callus</td>
<td>184 189</td>
<td>408</td>
<td></td>
</tr>
<tr>
<td>Shoot</td>
<td>139 157</td>
<td>395</td>
<td></td>
</tr>
<tr>
<td>Mixed tissues</td>
<td>99 921</td>
<td>371</td>
<td></td>
</tr>
<tr>
<td>Panicle</td>
<td>150 845</td>
<td>307</td>
<td></td>
</tr>
<tr>
<td>Flower</td>
<td>51 582</td>
<td>277</td>
<td></td>
</tr>
<tr>
<td>Root</td>
<td>79 340</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>Seed</td>
<td>26 407</td>
<td>203</td>
<td></td>
</tr>
<tr>
<td>Unknown samples</td>
<td>53 978</td>
<td>199</td>
<td></td>
</tr>
<tr>
<td>Pistil</td>
<td>77 110</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>Leaf</td>
<td>204 353</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>Anther</td>
<td>14 156</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Whole plant</td>
<td>64 601</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>
expression platform. The only common tissues reported on by all three platforms are leaf and root. However, with the exception of callus tissue, the trends for each pair-wise comparison between the platforms are the same as for these two tissues. The observed trends meet expectations based on the technical capabilities and limitations of each platform, with EST evidence showing expression of the lowest number of GTs and the Affymetrix microarray data giving signals above background for the highest number of GTs (Figure 6). The long sequence reads of ESTs provide fairly unequivocal evidence of gene expression; however, the cost of this method prevents deep and broad application, namely sequencing of very large numbers of clones from diverse tissue types under diverse conditions. Thus, the expectation is that percent expression based on ESTs represents a lower limit. In contrast, MPSS can be conducted with much greater depth and breadth, but the short tags can be generated through artifacts, reducing confidence in tags with very low frequency (Nobuta et al., 2007). Rice Affymetrix microarray data from Jain et al. (2007) are also divided into two classes: log2 signal > 7 (128 units) and log2 signal < 7.

Microarray expression data are known to be inaccurate for lowly expressed genes (Meyers et al., 2004). However, compared with sequencing-based methods, lowly expressed genes are equally likely as abundant transcripts to be detected given sufficient signal-to-noise. Therefore, microarray analysis provides the greatest opportunity for detection of any transcript uniquely represented on the array. Based on this rationale, we selected a log2 signal-value of >7 (corresponding to a spot intensity of 128) that yields ‘expression evidence’ for a similar, but larger number of GTs compared with the MPSS and EST data. At this log2 signal-value, 618 of the 634 (86%) rice GTs represented on the array are expressed in at least one of the rice tissues and developmental stages analyzed. For reference, a log2 signal-value of >6 yields expression evidence for 93% of rice GTs. In addition, the quantitative differences in percent of GTs expressed between the different methods, especially EST-leaf versus microarray-leaf compared with EST-shoot versus microarray-shoot, suggests that other differences between the datasets may interfere with this comparison. One source of the varying differences may be that the definitions of the sampled tissues vary between experiments. The EST data may be especially non-uniform, as they were gathered from multiple libraries. Despite technical challenges, this coarse analysis reveals that the different platforms consistently indicate that a large portion of GTs are expressed, rather than being pseudogenes. Further, the analysis reveals reasonable metrics of data confidence for the MPSS and rice Affymetrix data.

Having been roughly calibrated to the other expression platforms, the microarray data provide a broad picture of the expression differences among GT families. For example, most GT20 (trehalose synthases) and GT75 (UDP-arabinofuranose mutase) members are highly expressed in most tissues and developmental stages, while most GT1 (small molecule synthase) family as likely candidates for involvement in glucuronoarabinoylxylan synthesis. In the section that follows, we also distinguish genes in the GT47 family for which expression of different subclades differs drastically, is shown in Figure 7A. Mitchell et al. (2007) identified genes in the GT47 (β-glucuronotransferase and heparan synthase) family as likely candidates for involvement in glucuronarabinoylxylan synthesis. In the section that follows, we also distinguish genes from these families as rice-diverged and highly expressed in above-ground tissues. From Figure 7A, it is clear that most GT47 members are lowly expressed, with only a cluster of nine gene models (six loci) with high expression. Mitchell et al. (2007) identified these six loci as having high expression in monocots relative to dicots.
Figure 7. Affymetrix Rice Microarray Expression Profiles of Rice (A) GT47 and (B) GT61 Family Members in Different Tissues/Organs and during Different Developmental Stages.

The average log2 signal values of rice GTs in various tissues/organs and developmental stages (listed at the top of heatmap) are presented with the same gene order in the phylogenetic tree. The color scale (representing log2 signal values) is shown at the top. Data are those of Jain et al. (2007).
GT47 genes with low expression in rice (Figure 7A) are those that were found with similar expression levels between grasses and dicots (Mitchell et al., 2007). Furthermore, among the six highly expressed GT47 loci, we identified five to be rice-diverged GTs, while 25 out of the other 42 lowly expressed members have orthologs in dicots. The data in Figure 7A support the potential importance of the highly expressed subclade of GT47 genes in type II cell wall synthesis and the complementary methods used to identify these genes.

Gene expression for the GT61 (xylosyltransferase) family is shown in Figure 7B. GT61s are another family with a predicted role in glucuronorhabinoxylan synthesis (Mitchell et al., 2007). In contrast to those in the GT47 family, most GT61 transcripts are highly expressed in at least one tissue or developmental stage. Examination of the compiled list of rice GTs reveals that all GT61 members were found to have higher expression in grasses compared to dicots (Mitchell et al., 2007). These observations corroborate that GT61 members are candidates for glucuronorhabinoxylan biosynthesis. In both the GT61 and GT47 families, genes with similar expression patterns cluster together within the phylogenetic tree, suggesting that gene redundancy may be a barrier to functional studies. Loss-of-function analyses in these families may require simultaneous disruption or silencing of multiple genes (Miki et al., 2005). In conclusion, the availability of copious microarray data and other gene expression data in the Rice GT Database, combined with the phylogenetic tree, provides a powerful tool to study rice GT expression patterns toward developing hypotheses regarding function.

Identification of Rice-Diverged GTs with High Expression in Above-Ground Tissues

Most plant biomass under consideration for lignocellulosic biofuel production consists of vegetative, above-ground tissues, such as leaves, stems, shoots, and the progenitor of these tissues, the shoot apical meristem. Thus, identification of rice-diverged GTs with high expression in vegetative, above-ground tissues and elucidation of their function is likely to assist efforts to alter the composition of lignocellulosic biomass from grasses. To identify potential grass-diverged genes that show consistent expression in above-ground tissues, we identified rice-diverged genes that show moderate to high gene expression in at least two of the three rice gene expression datasets previously described, namely EST, MPSS, and microarrays. The datasets examined consisted of leaf and shoot EST libraries; young leaf, leaf, shoot, and meristem MPSS libraries; and young leaf, mature leaf, seedling, shoot, and meristem hybridizations from the Affymetrix microarray data. For each type of evidence, we selected the rice-diverged GTs that ranked in the top quartile of most highly expressed genes in at least one vegetative, above-ground tissue. We judge that the genes in the top quartile represent moderately to highly expressed genes based on the observation that almost half of all annotated rice gene models are not represented by ESTs (Jung et al., 2008c). The GT genes identified with these criteria for at least two expression analysis platforms are all highly expressed (Supplemental Table 7). Due to the arbitrary nature of the chosen gene expression criteria, if the list we have generated proves valuable for identifying genes centrally involved in above-ground cell wall synthesis, relaxing the criteria may allow us to identify additional high-quality targets for study. As listed in Table 3, 33 GT loci, representing 45 gene models, are rice-diverged and highly expressed in above-ground tissues. Detailed information for these GTs is shown in Supplemental Table 7. GTs from 14 families are represented. Thirty-six of the gene models (80%) have FL-cDNA support and the rest have EST support.

Based on analysis of the literature, we do not expect 12 of the 33 rice-diverged, highly expressed GT loci (20 gene models) to have direct roles in cell wall synthesis, including the GT1, GT4, and GT20 genes. GT1s glycosylate small molecule acceptors (Bowles et al., 2005; Vogt and Jones, 2000). Among those on the list, anthocyanidin and other flavanol GT1s assist in the stabilization and modification of these molecules (Vogt and Jones, 2000), which, in grass leaves, may have roles in both photoprotection and defense (Lo and Nicholson, 1998). In contrast, the GT1-mediated glycosylation of hormones, such as cytokinin and auxin, inactivates these molecules for storage, detoxification, and regulatory purposes (Hou et al., 2004; Jackson et al., 2002). Although almost all GT1 genes show low expression in some or all tissues by the Affymetrix microarray data, the GT1s on the list show high expression in certain tissues. For example, Os01g53350 has a high expression in shoot tissue (Affymetrix log2 intensity 12.2) and 14-day-old leaf (log2 signal 10.5), but low expression in other above-ground tissues. Moreover, nine ESTs were identified for this gene in the EST shoot library, indicating high expression in above-ground tissues. The GT20s on the list are involved in protein glycosylation in the endoplasmic reticulum and Golgi apparatus, which occurs via a diacylglycerol intermediate (Dormann et al., 1999; Silverstone et al., 2007). GT20s synthesize trehalose and have been found to have a role in regulating sugar metabolism (Blazquez et al., 1998; Gomez et al., 2006), which may explain the high levels of expression in rice photosynthetic tissues.

All of the remaining 21 GTs (25 gene models) are from families that have either been shown to be involved in cell wall synthesis, including the members of the GT2, GT8, GT37, GT43, GT47, GT48, and GT77 families, or have been implicated or hypothesized to be involved in cell wall synthesis, including the members of the GT29, GT31, GT61, and GT75 families. References that elucidate the connections or putative connections of these GT families with cell wall synthesis are provided. Of particular relevance to type II cell wall synthesis, the list includes two members of the CsIF gene family (GT2), which is involved in mixed linkage glucan synthesis (Burton et al., 2006). Furthermore, the families of a number of listed genes have been implicated in xylan synthesis, including the members of the GT8, GT43, GT47, and GT61 families (see Table 3 for references). Members of the GT77 family act as
xylosyltransferases in pectin synthesis and are involved with accumulation of arabinose-containing polymers in type I walls (Egelund et al., 2007), but may also be relevant to glucuronorabinoxyylan synthesis in type II cell walls. In summary, we expect a number of the rice-diverged, highly expressed GTs to play important roles in the synthesis of cell walls in vegetative, above-ground tissues of rice, distinguishing these genes among the hundreds of rice GTs as prime targets for functional studies in grasses.

### Table 3. Rice-Diverged GTs with High Expression in Vegetative, Above-Ground Tissues.

<table>
<thead>
<tr>
<th>TIGR ID</th>
<th>CAZy family</th>
<th>Acceptor class</th>
<th>Putative or known function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Os01g53350</td>
<td>GT1</td>
<td>Small molecule</td>
<td>Anthocyanidin 5,3-O-glucosyltransferase, putative</td>
<td>Lo and Nicholson, 1998</td>
</tr>
<tr>
<td>Os02g11110</td>
<td>GT1</td>
<td>Small molecule</td>
<td>Flavonol-3-O-glycoside-7-O-glucosyltransferase 1, putative</td>
<td>Ko et al., 2008</td>
</tr>
<tr>
<td>Os02g11640</td>
<td>GT1</td>
<td>Small molecule</td>
<td>Flavonol-3-O-glycoside-7-O-glucosyltransferase 1, putative</td>
<td>Ko et al., 2008</td>
</tr>
<tr>
<td>Os02g28900</td>
<td>GT1</td>
<td>Small molecule</td>
<td>Cytokinin-O-glucosyltransferase 2, putative</td>
<td>Hou et al., 2004</td>
</tr>
<tr>
<td>Os04g25440</td>
<td>GT1</td>
<td>Small molecule</td>
<td>Cytokinin-O-glucosyltransferase 2, putative</td>
<td>Hou et al., 2004</td>
</tr>
<tr>
<td>Os11g04860</td>
<td>GT1</td>
<td>Small molecule</td>
<td>Indole-3-acetate beta-glucosyltransferase, putative</td>
<td>Jackson et al., 2002</td>
</tr>
<tr>
<td>Os02g49332</td>
<td>GT2</td>
<td>Cell wall PS</td>
<td>CsfE2—cellulose synthase-like family E</td>
<td>Hazen et al., 2002</td>
</tr>
<tr>
<td>Os07g36630</td>
<td>GT2</td>
<td>Cell wall PS</td>
<td>CsfE8—beta1,3,1,4 glucan synthase</td>
<td>Burton et al., 2006</td>
</tr>
<tr>
<td>Os08g06380</td>
<td>GT2</td>
<td>Cell wall PS</td>
<td>CsfF6—beta1,3,1,4 glucan synthase</td>
<td>Burton et al., 2006</td>
</tr>
<tr>
<td>Os02g51060</td>
<td>GT2</td>
<td>Cell wall PS</td>
<td>CsfA6—mannan synthase, putative</td>
<td>Liepman et al., 2007</td>
</tr>
<tr>
<td>Os03g15840</td>
<td>GT4</td>
<td>Protein</td>
<td>N-acetyl glucosamine transferase, putative</td>
<td>Silverstone et al., 2007</td>
</tr>
<tr>
<td>Os03g16140</td>
<td>GT4</td>
<td>Lipid</td>
<td>Digalactosyldiacylglycerol synthase 2, putative</td>
<td>Dormann et al., 1999</td>
</tr>
<tr>
<td>Os11g05990</td>
<td>GT4</td>
<td>Lipid</td>
<td>Digalactosyldiacylglycerol synthase 1, putative</td>
<td>Dormann et al., 1999</td>
</tr>
<tr>
<td>Os03g11330</td>
<td>GT8</td>
<td>Cell wall PS</td>
<td>Transferase, transferring glycosyl groups, putative</td>
<td>Lee et al., 2007b</td>
</tr>
<tr>
<td>Os05g35200</td>
<td>GT8</td>
<td>Cell wall PS</td>
<td>Secondary cell wall-related glycosyltransferase family 8, putative</td>
<td>Lee et al., 2007b</td>
</tr>
<tr>
<td>Os06g12280</td>
<td>GT8</td>
<td>Cell wall PS</td>
<td>Glycosyl transferase family 8 protein</td>
<td>Lee et al., 2007b</td>
</tr>
<tr>
<td>Os02g54820</td>
<td>GT20</td>
<td>Small molecule</td>
<td>Trehalose-6-phosphate synthase, putative</td>
<td>Blazquez et al., 1998</td>
</tr>
<tr>
<td>Os02g44210</td>
<td>GT20</td>
<td>Small molecule</td>
<td>Alpha, alpha-trehalose-phosphate synthase, putative</td>
<td>Blazquez et al., 1998</td>
</tr>
<tr>
<td>Os08g34580</td>
<td>GT20</td>
<td>Small molecule</td>
<td>Trehalose-6-phosphate synthase, putative</td>
<td>Blazquez et al., 1998</td>
</tr>
<tr>
<td>Os12g05550</td>
<td>GT29</td>
<td>Unknown</td>
<td>Sialyltransferase-like protein, putative</td>
<td>—</td>
</tr>
<tr>
<td>Os08g02370</td>
<td>GT31</td>
<td>Cell wall Prot</td>
<td>N-glycan galactosyltransferase, putative</td>
<td>Qu et al., 2008; Strasser et al., 2007</td>
</tr>
<tr>
<td>Os02g52560</td>
<td>GT37</td>
<td>Cell wall PS</td>
<td>Galactoside 2-alpha-L-fucosyltransferase, putative</td>
<td>Perrin et al., 1999</td>
</tr>
<tr>
<td>Os07g49370</td>
<td>GT43</td>
<td>Cell wall PS</td>
<td>Beta3-glucuronyltransferase, putative</td>
<td>Lee et al., 2007a; Pena et al., 2007</td>
</tr>
<tr>
<td>Os01g70190</td>
<td>GT47</td>
<td>Cell wall PS</td>
<td>Secondary cell wall-related glycosyltransferase family 47, putative</td>
<td>Mitchell et al., 2007</td>
</tr>
<tr>
<td>Os04g57510</td>
<td>GT47</td>
<td>Cell wall PS</td>
<td>Exostosin-like, putative</td>
<td>Mitchell et al., 2007</td>
</tr>
<tr>
<td>Os02g58560</td>
<td>GT48</td>
<td>Cell wall PS</td>
<td>CALS1, putative</td>
<td>Hong et al., 2001</td>
</tr>
<tr>
<td>Os06g02260</td>
<td>GT48</td>
<td>Cell wall PS</td>
<td>Callose synthase catalytic subunit, putative</td>
<td>Hong et al., 2001</td>
</tr>
<tr>
<td>Os02g22380</td>
<td>GT61</td>
<td>Cell wall PS</td>
<td>Glycosyltransferase, putative</td>
<td>Mitchell et al., 2007</td>
</tr>
<tr>
<td>Os06g27560</td>
<td>GT61</td>
<td>Cell wall PS</td>
<td>HGA4, putative</td>
<td>Mitchell et al., 2007</td>
</tr>
<tr>
<td>Os06g28124</td>
<td>GT61</td>
<td>Cell wall PS</td>
<td>Glycosyltransferase, putative</td>
<td>Mitchell et al., 2007</td>
</tr>
<tr>
<td>Os03g40270</td>
<td>GT75</td>
<td>UDP-sugar</td>
<td>UDP-arabino pyranose mutase</td>
<td>Konishi et al., 2007</td>
</tr>
<tr>
<td>Os03g63270</td>
<td>GT77</td>
<td>Cell wall PS</td>
<td>Glycosyltransferase, putative</td>
<td>Egelund et al., 2006, 2007</td>
</tr>
<tr>
<td>Os07g19444</td>
<td>GT77</td>
<td>Cell wall PS</td>
<td>Glycosyltransferase, putative</td>
<td>Egelund et al., 2006, 2007</td>
</tr>
</tbody>
</table>

a TIGR annotation or function based on provided references.
b ‘Cell wall PS’ indicates a cell wall polysaccharide as a precursor.
c Function not known in plants, which lack sialic acid in N-linked glycans. Family members are similar to a Gossypium transcript associated with cell wall synthesis.
d ‘Cell wall Prot’ indicates a cell wall-associated protein.

Protein Topology

Determining the topology and sub-cellular localization of a protein is an important step toward understanding its function, especially for glycosyltransferases with diverse families and functions. In eukaryotes, most GTs are resident membrane proteins of the endoplasmic reticulum (ER) and the Golgi apparatus (Breton et al., 2001). Some plant GTs involved in starch biosynthesis, such as GT5 and GT35, are expected to be located...
in the chloroplast, the site of photosynthesis (Geisler-Lee et al., 2006). Though major questions still remain concerning the subcellular organization of many plant GTs, most of those involved in synthesis of cell wall components besides cellulose and callose have been found in the Golgi (Lerouxel et al., 2006). To help distinguish GTs putatively involved in cell wall synthesis from others, we adopted computational methods to predict the protein topology and sub-cellular localization of rice GTs. The program TMHMM predicts a transmembrane domain for 351 of the 769 putative rice GT proteins (45.6%) (Supplemental Table 8). TMHMM implements a circular Hidden Markov Model and is considered to be the best-performing transmembrane prediction program (Moller et al., 2001).

The GT1 family, which modifies small molecules in the cytoplasm (Vogt and Jones, 2000), has 224 members, but only 17 were predicted to contain a transmembrane domain. Excluding GT1s, the percentage of GTs with a transmembrane domain is predicted to be 61.3% (334/545). The others should be soluble proteins, but some may also be involved in the cell wall synthesis via protein–protein interactions with membrane-bound proteins.

In addition, we used other software to analyze the predicted sub-cellular localization based on the presence of sorting signals and other sequence properties. SignalP and ChloroP can detect the presence of N-terminal sequence motifs directing proteins to the secretory pathway and chloroplasts, then TargetP will predict sub-cellular localization based on the results of SignalP and ChloroP (Emanuelsson et al., 2007). Using SignalP, ChloroP, and TargetP to predict the sub-cellular localization of rice GTs, 140 (18.2%) were predicted to localize to the secretory pathway. The others were predicted to localize to the chloroplast (121), mitochondrion (204), and any other location (304). For the glycosyl synthases of the GT5 family, which are expected to localize to the chloroplast, 20 out of 26 members were predicted to localize in this organelle, two to the mitochondrion, and the others to any other location. One protein in the glycogen and starch synthase GT35 family was predicted to localize to the chloroplast and the other two to other locations. None was predicted to locate in the secretory pathway. This analysis suggests that localization prediction methods may, in some cases, particularly with respect to the chloroplast, be reliable. On the other hand, for the 41 Csl genes that are expected to localize to the Golgi (Richmond and Somerville, 2001), only three are predicted to localize to the secretory system. Thus, we are skeptical of the results of the localization prediction methods with respect to the secretory system and conclude that the predictions may not be helpful for studying cell wall synthesis (Egelund et al., 2006).

**Mutant Line Resources to Study Functions of GTs**

Gene-indexed mutant rice plants interrupting or activating expression of GTs may, in many cases, serve as useful resources for determining gene function. Several approaches have been undertaken to develop rice mutant lines in which genes are randomly tagged by DNA insertion elements, such as the Tos17 retrotransposon and Agrobacterium-derived T-DNA (An et al., 2003; Miyao et al., 2003). For rice GTs, we gathered mutant line information from available libraries (Table 4). Mutant lines corresponding to each rice GT are enumerated in Supplemental Table 9. Among the mutant libraries, NIAS Tos17, OTL Tos17 and T-DNA, and RMD T-DNA mutant lines have phenotype information available through their database website, and we have made hyperlinks to the phenotypic data in the Rice GT Database. The available phenotypic information implicates some GTs in processes associated with rice cell wall biosynthesis. For example, the rice GT2 gene, Os01g54620.1 (CESA4, an expressed cellulose synthase gene), has a Tos17 knockout line, NE1042, in the NIAS library that shows brittle, withering, and dwarf phenotypes. The homozygous deletion mutant for AT4G18780.1 (CESA8), the Arabidopsis ortholog of this rice GT, is severely dwarfed, sterile, and possesses dark green leaves that indicate an increase in chloroplasts per leaf area of the mutants, which may be due to reduced cell size (Persson et al., 2007). These phenotypes suggest a role for the rice GT, Os01g54620.1, in cell wall biosynthesis. Thus, the availability of mutant lines and any corresponding phenotypic information as linked through the Rice GT Database will be helpful for the further functional analysis of rice GTs.

In this study, we identified 609 rice GT genes, representing 769 gene models, and created the Rice GT Database to provide a logical and functional format to host and analyze diverse sets of information in a phylogenetic context. Rather than analyzing rice GTs one by one, this database allows simultaneous visualization of all the rice GT families and subfamilies. This format allows comparison of the features of rice GTs between and within different families. Using this database, we identified 33 rice-diverged GT genes with high expression in vegetative, above-ground tissues. We hypothesize that many of these GTs will have important roles in the biosynthesis of grass-specific cell wall components and thus are prime candidates for further functional analysis. We plan to update this

<table>
<thead>
<tr>
<th>Mutant library</th>
<th>No. of GTs with mutant lines</th>
<th>No. of mutant lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIAS Tos17a</td>
<td>75</td>
<td>276</td>
</tr>
<tr>
<td>OTL Tos17a</td>
<td>71</td>
<td>190</td>
</tr>
<tr>
<td>UCD Ds</td>
<td>67</td>
<td>124</td>
</tr>
<tr>
<td>RMD T-DNA*</td>
<td>157</td>
<td>196</td>
</tr>
<tr>
<td>TRIM T-DNA</td>
<td>108</td>
<td>118</td>
</tr>
<tr>
<td>OTL T-DNA*</td>
<td>141</td>
<td>122</td>
</tr>
<tr>
<td>Postech T-DNA</td>
<td>533</td>
<td>991</td>
</tr>
<tr>
<td>Postech AC</td>
<td>429 b</td>
<td>954 b</td>
</tr>
</tbody>
</table>

* a Phenotypic information for the mutant lines is available on the website. Hyperlinks are also provided in the Rice GT Database.
 b These numbers indicate putative activation tagging lines expected with their insertional positions.
database semi-annually and add additional features to the database, including: links to PubMed citations, protein–protein interaction data from experimental determination or computational prediction, new mutant lines and corresponding phenotype information for both GTs and their interacting proteins, MPSS data from new libraries, and new microarray expression data. We anticipate this database will provide a useful service to the plant researchers and accelerate biofuel research in particular.

METHODS

Identification of Rice GTs and Database Construction

We searched the CAZy database (www.cazy.org/) and downloaded all the rice GTs hosted in this database (Campbell et al., 1997; Coutinho et al., 2003). Because genes in CAZy are associated with different kinds of gene names, including RAP2 (Rice Annotation Project Version 2) IDs, NCBI IDs, common names and TIGR IDs, all identifiers were converted to TIGR Version 5 IDs using the RAP ID Converter (http://rapdb.dna.affrc.go.jp/tools/converter) and NCBI BLAST Version 2.2.17 searches (Altschul et al., 1990). Arabidopsis GTs from CAZy and identified by fold recognition (Egelund et al., 2004) were also used to scan all the annotated proteins in the rice (Oryza sativa ssp. japonica) genome at TIGR (Version 5) (Ouyang et al., 2007; Yuan et al., 2005), to find the corresponding rice homologs (i.e. homolog database). In addition, the GT-related domains from the Pfam database (http://pfam.sanger.ac.uk/) were used to scan the rice genome to identify putative GTs containing GT-related domains using HMMER 2.3.2 (i.e. domain search) (Finn et al., 2006). Finally, the GTs identified in the previous steps were used to search the corresponding paralogs using the TIGR Paralog Family Classification database (i.e. paralog search) (Lin et al., 2008). After assembling the initial putative rice GT list, the Pfam and Interpro databases (www.ebi.ac.uk/interpro) were used to check whether the candidates have GT-related domains (Finn et al., 2006; Mulder et al., 2007). Except as mentioned in the Results, genes lacking a GT-related domain and not annotated as GT-related genes in the TIGR annotation database were deleted from the current list. Additionally, five TE-related candidates were also discarded. A phylogenomic database was then constructed with ASP.NET and MSSQL, run on a Windows 2003 server. The http address is http://ricephylogenomics.ucdavis.edu/cellwalls/gt/

Phylogenetic Analysis

For each GT family with more than three members, the corresponding GT domain sequences were extracted according to the Pfam and Interpro domain assignments. We aligned GT domain sequences in these families using ClustalW Version 2.0 with default options (Larkin et al., 2007). The alignments were then corrected manually using the alignment editor software BioEdit Version 7.0.09 (www.mbio.ncsu.edu/BioEdit/bioedit.html). The sequences used for alignments are available in Supplemental Table 10. The unrooted phylogenetic tree was constructed with the neighbor-joining method executed in PHYLIP Version 3.67 (http://evolution.genetics.washington.edu/phylip.html) using only the domain sequences. Bootstrapping can provide an estimate of the confidence for each branch point, so 1000 bootstraps were adopted to infer the statistical support for the tree (Supplemental Figure 2).

Orthology Detection in Dicots

Inparanoid Version 2.0 was employed to evaluate the orthology relationships among rice and sequenced, annotated dicots on the whole-genome scale (Remm et al., 2001). The Arabidopsis genome sequences were downloaded from the Arabidopsis Information Resource (TAIR8, www.arabidopsis.org/), P. trichocarpa from the DoE Joint Genome Institute and Poplar Genome Consortium annotation v1.1 (http://genome.jgi-psf.org/Pptr1_1) (Tuskan et al., 2006), M. truncatula from the Medicago Genome Sequence Consortium (MGSC) Mt2.0 release (www.medicago.org/) (Young et al., 2005), and R. communis from the TIGR Castor bean Database (http://castorbear.tigr.org/).

Digital Expression Analysis (EST, MPSS, and Microarray)

EST Analysis

The Rice Gene Expression Anatomy Viewer (http://rice.plantbiology.msu.edu/dnav.shtml) provides the number of ESTs from several different rice tissues mapped onto TIGR gene models and was used for digital expression analysis of rice GTs (Jung et al., 2008c). Each of the TIGR locus IDs corresponding to all rice GT gene models was searched to find availability of corresponding EST evidence. The EST evidence was determined using the PASA program, which utilizes a number of alignment programs to maximally align transcripts to the genome (Haas et al., 2003). The minimal alignment allowed by the PASA program is 95% identity over 90% length of the transcript.

MPSS Analysis

Expression evidence from MPSS tags was determined from the rice MPSS project (http://mpss.udel.edu/rice/) mapped onto the TIGR rice gene models. We used only the sense strand signatures (Classes 1, 2, 5, and 7), which have only one hit on the rice pseudomolecules and show a perfect match (100% identity over 100% of the length of the tag) in the analysis. The normalized abundance (tags per million, TPM) of these signatures for a given gene in a given library represents a quantitative estimate of expression level of that gene. MPSS expression data for 17-bp signatures from 18 libraries representing 12 different tissues/organisms of rice were used. The description of these libraries is: NCA, 35 d callus; NCL, 14 d young leaves stressed in 4°C cold for 24 h; NCR, 14 d young roots stressed in 4°C cold for 24 h; NDL, 14 d young leaves stressed in drought for 5 d; NDR, 14 d young roots stressed in drought for 5 d; NGD, 10 d germinating seedlings grown in dark; NGS, 3 d germinating seed; NIP, 90 d immature panicle; NL4, 60 d mature leaves
(combination of replicates); NPO, 60 d crown vegetative meristematic tissue; NPO, mature pollen; NR2, 60 d mature roots (combination of replicates); NSL, 14 d young leaves stressed in 250 mM NaCl for 24 h; NSO, ovary and mature stigma; NSR, 14 d young roots stressed in 250 mM NaCl for 24 h; NST, 60 d stem; NYL, 14 d young leaves; NYR, 14 d young roots.

**Microarray Analysis**

The raw data for rice Affymetrix microarray experiment designed to profile the expression patterns of rice development were downloaded from the NCBI Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo) (Jain et al., 2007). The GEO accession number is GSE6893. Then MAS 5.0 method provided by the R package, affy, for the Affymetrix rice array was used to conduct background correction, normalization, probe-specific background correction, probe summarization and to convert probe level data to expression values (Affymetrix, 2001). The trimmed mean target intensity of each array was arbitrarily set to 500. These data were then log2 transformed. The rice Multi-platform Microarray Search tool (www.ricearray.org/matrix.search.shtml) was used to assign the corresponding Affymetrix probe sets for rice GTs (Jung et al., 2008c). We only included unique probe sets that match a single rice locus in the analysis. If several unique probe sets were available for a single rice GT gene model, we selected the probe set with the highest expression. The heatmap was generated by the TIGR MultiExperiment Viewer v4.1 (MeV, www.tm4.org/mev.html).

**Protein Topology and Sub-Cellular Localization Prediction**

We scanned rice GT proteins for the presence of a transmembrane domain(s) using TMHMM Version 2.0, which predicts the presence of transmembrane helices in amino acid sequences using HMM-based predictions (Krogh et al., 2001). SignalP Version 3.0, ChloroP Version 1.1, and TargetP Version 1.1 were used to predict the presence of signal peptide, chloroplast transit peptide, and the sub-cellular localization, respectively (Emanuelsson et al., 2007).

**Identification of Rice GT Gene-indexed Mutant Lines and Relating Rice Functional Genomic Databases**

Several rice mutant line libraries are available, including the National Institute of Agrobiological Sciences (NIAS) Tos17 Insertion Mutant Database (Miyao et al., 2003); the UCD Rice Transposon Flanking Sequence Tag Database with Ds Knockout (KO) lines (Kolesnik et al., 2004); the Oryza Tag Line (OTL) Database with Tos17 and T-DNA KO lines; the Rice Mutant Database (RMD) with T-DNA KO lines (Zhang et al., 2006); the Taiwan Rice Insertional Mutants Database (TRIM) with T-DNA KO lines; and the Postech Rice T-DNA Insertion Sequence Database with T-DNA KO and Activation (AC) lines (An et al., 2003; Jeong et al., 2006). The OryGenesDB database (http://orygenesdb.cirad.fr/index.html) was used to map flanking sequence tags (FSTs) from the different mutant libraries onto rice GTs (Droc et al., 2006). The flanking sequences have been placed in the TIGR Version 5 pseudomolecules by finding the highest hit based on an e^{-10} cut-off. The mapped insertions were then assigned to rice GT genes based on the insertion map locations relative to the TIGR genome annotations. In the OryGenesDB database, a gene was defined as beginning 800 bp 5' of the initiation codon and to the end of the 3'-UTR, where known. The Postech activation lines were obtained from the Postech Rice T-DNA Insertion Sequence Database (http://141.223.132.44/pfg/index.php) (Jeong et al., 2006).

**SUPPLEMENTARY DATA**

Supplementary Data are available at Molecular Plant Online.

**FUNDING**

This work was supported by a Department of Energy grant to the Joint BioEnergy Institute and a Plant Genomics Research Program National Science Foundation grant DBI-0313887 to P.C.R.

**ACKNOWLEDGMENTS**

We thank Drs Henrik Scheller and Wolf Frommer for their suggestions regarding this project and comments on the manuscript. We also thank Drs Blake C. Meyers and Kan Nobuta for sending MPSS data and Dr Gynheung An for providing the information on the knockout and activation tagging lines of rice GTs. L.E.B. was supported in part by a University of California Office of the President Postdoctoral Fellowship. K.H.J. was supported in part by a Korea Research Foundation Grant (KRF-2005-C00155). No conflicts of interest are declared.

**REFERENCES**


Egelund, J., et al. (2007). Molecular characterization of two Arabidopsis thaliana glycosyltransferase mutants, rra1 and rra2, which have a reduced residual arabino sugar content in a polymer tightly associated with the cellulose wall residue. Plant Mol. Biol. 64, 439–451.


