(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 November 2013 (14.11.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/170201 A2

(51) International Patent Classification:
A01H 5/00 (2006.01) *C12N 15/82* (2006.01)

(21) International Application Number:
PCT/US2013/040632

(22) International Filing Date:
10 May 2013 (10.05.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/645,537 10 May 2012 (10.05.2012) US

(71) Applicant: **THE REGENTS OF THE UNIVERSITY OF CALIFORNIA** [US/US]; 1111 Franklin Street, 12th Floor, Oakland, California 94607 (US).

(72) Inventors: **SCHELLER, Henrik Vibe**; 211 El Bonito Way, Millbrae, California 94030 (US). **HEAZLEWOOD, Joshua L.**; 2133 Sacramento Street, Berkeley, California 94702 (US).

(74) Agents: **LOCKYER, Jean M.** et al.; Kilpatrick Townsend & Stockton LLP, Two Embarcadero Center, 8th Floor, San Francisco, California 94111 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

[Continued on next page]

(54) Title: REGULATION OF GALACTAN SYNTHASE EXPRESSION TO MODIFY GALACTAN CONTENT IN PLANTS

(57) Abstract: The invention provides methods of engineering plants to modulate galactan content. The invention additionally provides compositions and methods comprising such plants.

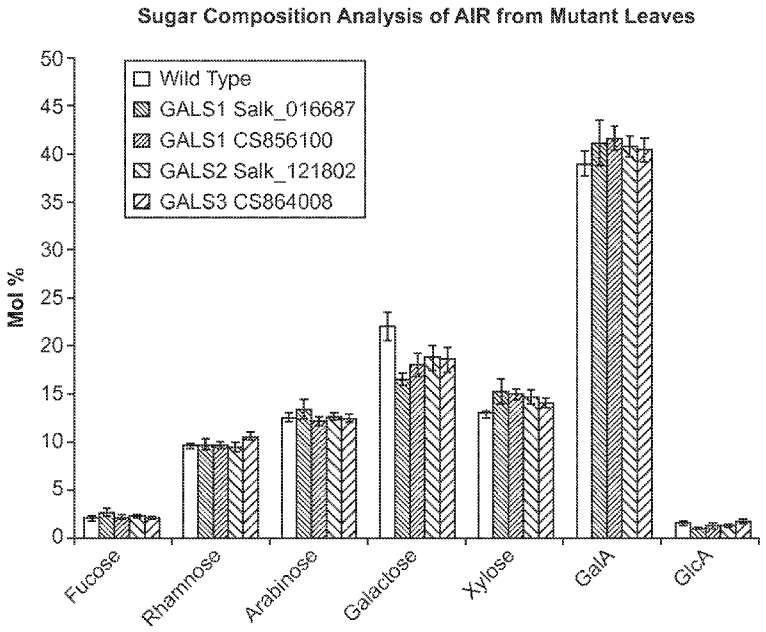

FIG. 3

WO 2013/170201 A2

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

Regulation of Galactan Synthase Expression to Modify Galactan Content in Plants

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims benefit of U.S. provisional application no. 61/645,537, filed May 10, 2012, which application is herein incorporated by reference.

STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH AND DEVELOPMENT

[0002] This invention was made with government support under Contract No. DE-AC02-05CH11231 awarded by the U.S. Department of Energy. The government has certain rights in this invention.

BACKGROUND OF THE INVENTION

[0003] Plant cell wall is the only source of cellulose for the paper industry and is a promising source of sugar for lignocellulosic biofuels. The utilization of plants to convert solar energy into transportable and storable energy will have a positive impact on the environment, since using plants can help to drastically reduce the utilization of fossil-derived fuels, can reduce carbon emission into the atmosphere, and even can contribute to carbon sequestration. However, even if lignocellulosic biofuels will be beneficial for the environment, the cost to produce them is still not cost-effective, mainly due to the expensive raw sugar derived from plant cell wall. The low density, recalcitrance to enzymatic hydrolysis, and low ratio of hexoses to pentoses in the biomass are the main contributors to the sugar cost because they impact transportation cost and require high amount of energy and chemicals. Therefore, improving the digestibility of the raw biomass and improving recovery of sugars that are more readily fermentable from biomass will have an important beneficial impact on the cost of lignocellulosic biofuels production.

[0004] Plant cell walls are predominantly composed of different polysaccharides, which can be grouped into cellulose, hemicelluloses and pectin. Pectin is a class of polysaccharide characterized by a high content of galacturonic acid residues and consists of two major types: homogalacturonan entirely composed of alpha-1,4-linked galacturonosyl residues, and rhamnogalacturonan I (RGI) composed of a backbone of alternating rhamnose and galacturonic acid residues with sidechains composed of arabinan and beta-1,4-galactan

(Mohnen, 2008; Harholt et al., 2010). Other domains of pectin include RGII, a complex structure with numerous different sugars, and xylogalacturonan, which is a type of HG with sidechains consisting of single xylosyl residues. It has been estimated that as many as 67 different transferases are required for bioynthesis of pectin (Mohnen, 2008; Harholt et al., 2010) but so far only one has been unambiguously indentified, namely the homogalacturonan galacturonosyltransferase GAUT1 (Sterling et al., 2006). A likely xylogalacturonan xylosyltransferase designated XGD1 has also been described but the final proof of activity of the isolated XGD1 protein has not been provided (Jensen et al., 2008). Biosynthesis of the arabinan sidechains on RGI involves the ARAD1 and ARAD2 proteins, but if they are arabinosyltransferases or work in a different way has not been determined (Harholt et al., 2006; Harholt et al., 2012). Beta-1,4-galactan constitutes a large part of pectin and of the total cell wall. However, little is known about the enzymes in plants responsible for its synthesis.

BRIEF SUMMARY OF THE INVENTION

[0005] This invention is based, in part, on the discovery that enzymes in the glycosyltransferases family 92 play a role in modulating galactan levels in plant cell wall. Plants harboring loss-of function mutations in genes encoding the GT92 family members of the invention, showed a decreased content of beta-1,4-galactan. The plants did not have an apparent growth phenotype, but pectin was more easily extracted from the cell walls of the mutants, and saccharification was improved. Overexpression of galactan synthase increased galactan content of cell walls. The invention thus provides methods employing plant GT92 family members for modulating galactan content in plants; and compositions and method of using such compositions.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] Figure 1 provides an alignment of various GT92 family members. The polypeptide sequence for each protein extends across the Figure 1 continuation pages. The SEQ ID NO for the protein is provided on the first page of Figure 1.

[0007] Figure 2 provides a schematic illustration of inserts in mutant lines.

[0008] Figure 3 provides illustrative data of the sugar composition in mutant and wildtype plants.

[0009] Figure 4 provides illustrative data using an antibody specific for beta-1,4-galactan (LM5) that showed less galactan in the mutants compared to wildtype plants and that petioles

had significantly less LM5 labeling in the mutants. Petioles were evaluated because the GT92 enzymes are relatively highly expressed in this tissue.

[0010] Figure 5 shows illustrative data demonstrating galactosyltransferase activity. Activity assays were performed with a beta-1,4-galacto-pentaose acceptor and microsome (left panel) or affinity purified GALS1 protein (right panel).

[0011] Figure 6 provides illustrative data showing monosaccharide composition of leaf cell wall of plants overexpressing GalS1 or expressing NST1-2A-AtUGE2 under the fiber specific pIRX5 promoter in a Col-0 background. '2A' is a sequence from foot-and-mouth-disease virus that can be used to express multiple proteins from the same ORF. Sugar levels are described as molar percentage, ± SE (n=6). From left to right for each sugar, the bars are: Col-0, GalS1 OE, pIRX5:UGE2#7; pIRX5:UGE2#21, and pIRX5:UGE2#25

[0012] Figure 7 provides illustrative data showing monosaccharide composition of leaf cell wall of plants coexpressing GalS1 and NST1-2A-AtUGE2 under the control of the fiber specific pIRX5 promoter or the constitutive 35S promoter. Asterisks mark the three lines that have significantly different galactose molar percentages (t-test, pBonf < 0.0083). Sugar levels are described as molar percentage, ± SE (n=6). From left to right for each sugar, the bars are: Col-0, GalS1 OE, 35S:F-UGE2 + GalS1 #1, 35S:F-UGE2 + GalS1 #2, 35S:F-UGE2 + GalS1 #5, pIRX5-UGE2 + GalS1 #4, pIRX5-UGE2 + GalS1 #3, pIRX5-UGE2 + GalS1 #14.

[0013] Figure 8 provides illustrative data showing monosaccharide composition of cell walls of stems of plants coexpressing GasIS1 and AtUGE2. As in Figures 7 and 8, expression of UGE2 under the pIRX5 promoter was done with a construct that simultaneously expresses the NST1 fiber-specific transcription factor, separated from UGE2 with a 2A domain. From left to right for each sugar, the bars are: Col-0, GalS1 OE, pIRX5:UGE2 + GalS1, 35S:F-UGE2 + GalS1, pIrX5:UGE2

[0014] Figure 9 proivdes data showing monosaccharide composition in cell walls of stems omitting xylose from the data. From left to right for each sugar, the bars are: Col-0, GalS1 OE, pIRX5:UGE2 + GalS1, 35S:F-UGE2 + GalS1, pIrX5:UGE2

DETAILED DESCRIPTION OF THE INVENTION

I. Definitions

[0015] As used herein, the term "galactan synthase" or "beta-1,4, galactan synthase" or "GALS" are used interchangeably to refer to an enzyme that is involved in the elongation of beta-1,4-galactan and has beta-1,4 galactosyltransferase activity. In the current invention, a galactan synthase is a glycosyltransferase in the family GT92. The term encompasses polymorphic variants, alleles, mutants, and interspecies homologs to the specific polypeptides described herein. A nucleic acid that encodes a galactan synthase refers to a gene, pre-mRNA, mRNA, and the like, including nucleic acids encoding polymorphic variants, alleles, mutants, and interspecies homologs of the particular amino acid sequences described herein. Thus, in some embodiments, a galactan synthase encodes a polypeptide having an amino acid sequence that has at least 50% amino acid sequence identity, or at least 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, preferably 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% or greater amino acid sequence identity, preferably over a region of at least about 25, 50, 100, 200 or more amino acids, or over the length of the entire polypeptide, to any one of the amino acid sequences shown in SEQ ID NOS:1, 2, or 3; or to any one of the plant amino acid sequences SEQ ID NOS:1-39 shown in Figure 1; or to any one of the amino acid sequence SEQ ID NO:1-27 shown in Figure 1. Examples of gene ids and accession numbers of galactan synthases are shown in the annotation for SEQ ID NOS:42-44. SEQ ID NOS:42-44 provide illustrations of GALS nucleic acids suitable for use in the invention.

[0016] As used herein, the term "transcription factor that regulates the production of components of a biosynthetic pathway" or "master transcription factor" refers to a transcription factor that regulates expression of one or of multiple genes in a biosynthetic pathway.

[0017] The term "downstream target," when used in the context of a downstream target of a transcription factor that regulates a component of a biosynthetic pathway of interest refers to a gene or protein whose expression is directly or indirectly regulated by the transcription factor. In some embodiments, the downstream target is a gene or protein that is directly or indirectly upregulated by the transcription factor. In some embodiments, the downstream target is a gene or protein that is directly or indirectly downregulated by the transcription factor.

[0018] The terms "increased level of activity," or "increased activity" refer interchangeably to an increase in the amount of activity of GALS protein in a plant engineered to increase

GALS compared to the amount of activity in a wild-type (*i.e.*, naturally occurring) plant. In some embodiments, increased activity results from increased expression levels. An increased level of activity or increased level of expression can be an increase in the amount of activity or expression of GALS in a plant genetically modified to overexpress GALS of at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% or greater compared to a wildtype plant. In some embodiments, the increased GALS activity or expression is localized to one or more tissues of the engineered plant, such as the xylem cells with secondary cell walls. Increased expression or activity of a GALS gene or protein can be assessed by any number of assays, including, but not limited to, measuring the level of RNA encoded by the GALS gene, the level of protein GALS protein, the levels of GALS enzymatic activity, or by measuring galactan content of a plant tissue.

[0019] The terms "reduced level of activity," "reduced activity" and "decreased activity" refer interchangeably to a reduction in the amount of activity of GALS protein in a plant engineered to decrease GALS compared to the amount of activity in a wild-type (*i.e.*, naturally occurring) plant. In some embodiments, reduced activity results from reduced expression levels. A reduced level of activity or a reduced level of expression can be a reduction in the amount of activity or expression of GALS of at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% or greater. In some embodiments, the reduced level of activity or reduced level of expression occurs, throughout all the tissues of the engineered plant. In some embodiments, the reduction in the amount of activity or expression is localized to one or more tissues of the engineered plant, such as the cell wall. In some embodiments, the GALS is not reduced in amount, but is modified in amino acid sequence so that the enzymatic activity is reduced directly or indirectly. Decreased expression or activity of a GALS gene or protein can be assessed by any number of assays, including, but not limited to, measuring the level of RNA encoded by the GALS gene, the level of protein GALS protein, the levels of GALS enzymatic activity, or by measuring galactan content of a plant tissue.

[0020] The terms "polynucleotide" and "nucleic acid" are used interchangeably and refer to a single or double-stranded polymer of deoxyribonucleotide or ribonucleotide bases read from the 5' to the 3' end. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases, nucleic acid analogs may be used that may have alternate backbones, comprising, *e.g.*, phosphoramidate, phosphorothioate, phosphorodithioate, or O-methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press); positive backbones; non-ionic backbones, and non-ribose backbones. Thus, nucleic acids or polynucleotides may also

include modified nucleotides that permit correct read-through by a polymerase. "Polynucleotide sequence" or "nucleic acid sequence" includes both the sense and antisense strands of a nucleic acid as either individual single strands or in a duplex. As will be appreciated by those in the art, the depiction of a single strand also defines the sequence of the complementary strand; thus the sequences described herein also provide the complement of the sequence. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses variants thereof (*e.g.*, degenerate codon substitutions) and complementary sequences, as well as the sequence explicitly indicated. The nucleic acid may be DNA, both genomic and cDNA, RNA or a hybrid, where the nucleic acid may contain combinations of deoxyribo- and ribo-nucleotides, and combinations of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthine hypoxanthine, isocytosine, isoguanine, etc.

[0021] The term "substantially identical," used in the context of two nucleic acids or polypeptides, refers to a sequence that has at least 50% sequence identity with a reference sequence. Percent identity can be any integer from 50% to 100%. Some embodiments include at least: 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99%, compared to a reference sequence using the programs described herein; preferably BLAST using standard parameters, as described below. For example, a polynucleotide encoding a GALS polypepitde may have a sequence that is at least 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identical to a sequence of SEQ ID NO:42, SEQ ID NO:43, or SEQ ID NO:44.

[0022] Two nucleic acid sequences or polypeptide sequences are said to be "identical" if the sequence of nucleotides or amino acid residues, respectively, in the two sequences is the same when aligned for maximum correspondence as described below. The terms "identical" or percent "identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same, when compared and aligned for maximum correspondence over a comparison window, as measured using one of the following sequence comparison algorithms or by manual alignment and visual inspection. When percentage of sequence identity is used in reference to proteins or peptides, it is recognized that residue positions that are not identical often differ by conservative amino acid substitutions, where amino acids residues are substituted for other amino acid residues with similar chemical properties (*e.g.*, charge or hydrophobicity) and therefore do not change the functional properties of the molecule. Where sequences differ in conservative substitutions,

the percent sequence identity may be adjusted upwards to correct for the conservative nature of the substitution. Means for making this adjustment are well known to those of skill in the art. Typically this involves scoring a conservative substitution as a partial rather than a full mismatch, thereby increasing the percentage sequence identity. Thus, for example, where an identical amino acid is given a score of 1 and a non-conservative substitution is given a score of zero, a conservative substitution is given a score between zero and 1. The scoring of conservative substitutions is calculated according to, *e.g.*, the algorithm of Meyers & Miller, *Computer Applic. Biol. Sci.* 4:11-17 (1988) *e.g.*, as implemented in the program PC/GENE (Intelligenetics, Mountain View, California, USA).

[0023] For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters.

[0024] A "comparison window," as used herein, includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 20 to 600, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequences for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, *Adv. Appl. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman & Wunsch, *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson & Lipman, *Proc. Nat'l. Acad. Sci. USA* 85:2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection.

[0025] Algorithms that are suitable for determining percent sequence identity and sequence similarity are the BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.* (1990) *J. Mol. Biol.* 215: 403-410 and Altschul *et al.* (1977) *Nucleic Acids Res.* 25: 3389-3402, respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (NCBI) web site. The algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in

the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al, supra*). These initial neighborhood word hits acts as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0) and N (penalty score for mismatching residues; always <0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a word size (W) of 28, an expectation (E) of 10, M=1, N=-2, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a word size (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)).

[0026] The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (*see, e.g.*, Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.01, more preferably less than about 10^{-5}, and most preferably less than about 10^{-20}.

[0027] Nucleic acid or protein sequences that are substantially identical to a reference sequence include "conservatively modified variants." With respect to particular nucleic acid sequences, conservatively modified variants refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given protein. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon, the codon can

be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation of the nucleic acid. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine) can be modified to yield a functionally identical molecule. Accordingly, each silent variation of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

[0028] As to amino acid sequences, one of skill will recognize that individual substitutions, in a nucleic acid, peptide, polypeptide, or protein sequence which alters a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration results in the substitution of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art.

[0029] The following six groups each contain amino acids that are illustsrative conservative substitutions for one another:

1) Alanine (A), Serine (S), Threonine (T);
2) Aspartic acid (D), Glutamic acid (E);
3) Asparagine (N), Glutamine (Q);
4) Arginine (R), Lysine (K);
5) Isoleucine (I), Leucine (L), Methionine (M), Valine (V); and
6) Phenylalanine (F), Tyrosine (Y), Tryptophan (W).

(*see, e.g.*, Creighton, *Proteins* (1984)).

[0030] Another indication that nucleotide sequences are substantially identical is if two molecules hybridize to each other, or a third nucleic acid, under stringent conditions. Stringent conditions are sequence dependent and will be different in different circumstances. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Typically, stringent conditions will be those in which the salt concentration is about 0.02 molar at pH 7 and the temperature is at least about 60°C. For example, stringent conditions for hybridization, such as RNA-DNA hybridizations in a blotting technique are those which include at least one wash in 0.2X SSC at 55°C for 20 minutes, or equivalent conditions.

[0031] The term "promoter," as used herein, refers to a polynucleotide sequence capable of driving transcription of a DNA sequence in a cell. Thus, promoters used in the polynucleotide constructs of the invention include cis- and trans- acting transcriptional control elements and regulatory sequences that are involved in regulating or modulating the timing and/or rate of transcription of a gene. For example, a promoter can be a cis- acting transcriptional control element, including an enhancer, a promoter, a transcription terminator, an origin of replication, a chromosomal integration sequence, 5' and 3' untranslated regions, or an intronic sequence, which are involved in transcriptional regulation. These cis-acting sequences typically interact with proteins or other biomolecules to carry out (turn on/off, regulate, modulate, *etc.*) gene transcription. Promoters are located 5' to the transcribed gene, and as used herein, include the sequence 5' from the translation start codon (i.e., including the 5' untranslated region of the mRNA, typically comprising 100-200 bp). Most often the core promoter sequences lie within 1-2 kb of the translation start site, more often within 1 kbp and often within 500 bp of the translation start site. By convention, the promoter sequence is usually provided as the sequence on the coding strand of the gene it controls. In the context of this application, a promoter is typically referred to by the name of the gene for which it naturally regulates expression. A promoter used in an expression construct of the invention is referred to by the name of the gene. Reference to a promoter by name includes a wildtype, native promoter as well as variants of the promoter that retain the ability to induce expression. Reference to a promoter by name is not restricted to a particular plants species, but also encompasses a promoter from a corresponding gene in other plant species.

[0032] A "constitutive promoter" in the context of this invention refers to a promoter that is capable of initiating transcription in nearly all cell types, whereas a "cell type-specific promoter" or "tissue-specific promoter" initiates transcription only in one or a few particular cell types or groups of cells forming a tissue. In some embodiments, a promoter is tissue - specific if the transcription levels initiated by the promoter in the cell wall are at least 2-fold, 3-fold, 4-fold, 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, 10-fold, 50-fold, 100-fold, 500-fold, 1000-fold higher or more as compared to the transcription levels initiated by the promoter in non-cell wall tissues

[0033] A polynucleotide is "heterologous" to an organism or a second polynucleotide sequence if it originates from a foreign species, or, if from the same species, is modified from its original form. For example, when a polynucleotide encoding a polypeptide sequence is said to be operably linked to a heterologous promoter, it means that the polynucleotide coding sequence encoding the polypeptide is derived from one species whereas the promoter

sequence is derived from another, different species; or, if both are derived from the same species, the coding sequence is not naturally associated with the promoter (*e.g.*, is a genetically engineered coding sequence, e.g., from a different gene in the same species, or an allele from a different ecotype or variety).

[0034] The term "operably linked" refers to a functional relationship between two or more polynucleotide (*e.g.*, DNA) segments. Typically, it refers to the functional relationship of a transcriptional regulatory sequence to a transcribed sequence. For example, a promoter or enhancer sequence is operably linked to a DNA or RNA sequence if it stimulates or modulates the transcription of the DNA or RNA sequence in an appropriate host cell or other expression system. Generally, promoter transcriptional regulatory sequences that are operably linked to a transcribed sequence are physically contiguous to the transcribed sequence, i.e., they are cis-acting. However, some transcriptional regulatory sequences, such as enhancers, need not be physically contiguous or located in close proximity to the coding sequences whose transcription they enhance.

[0035] The term "expression cassette" or "DNA construct" or "expression construct" refers to a nucleic acid construct that, when introduced into a host cell, results in transcription and/or translation of an RNA or polypeptide, respectively. Antisense or sense constructs that are not or cannot be translated are expressly included by this definition. In the case of both expression of transgenes and suppression of endogenous genes (*e.g.*, by antisense, RNAi, or sense suppression) one of skill will recognize that the inserted polynucleotide sequence need not be identical, but may be only substantially identical to a sequence of the gene from which it was derived. As explained herein, these substantially identical variants are specifically covered by reference to a specific nucleic acid sequence. One example of an expression cassette is a polynucleotide construct that comprises a polynucleotide sequence encoding a GALS protein operably linked to a heterologous promoter. In some embodiments, an expression cassette comprises a polynucleotide sequence encoding a GALS protein that is targeted to a position in a plant genome such that expression of the polynucleotide sequence is driven by a promoter that is present in the plant

[0036] The term "plant" as used herein can refer to a whole plant or part of a plant, *e.g.*, seeds, and includes plants of a variety of ploidy levels, including aneuploid, polyploid, diploid and haploid. The term "plant part," as used herein, refers to shoot vegetative organs and/or structures (*e.g.*, leaves, stems and tubers), branches, roots, flowers and floral organs (*e.g.*, bracts, sepals, petals, stamens, carpels, anthers), ovules (including egg and central cells), seed (including zygote, embryo, endosperm, and seed coat), fruit (*e.g.*, the mature

ovary), seedlings, and plant tissue (*e.g.*, vascular tissue, ground tissue, and the like), as well as individual plant cells, groups of plant cells (*e.g.*, cultured plant cells), protoplasts, plant extracts, and seeds. The class of plants that can be used in the methods of the invention is generally as broad as the class of higher and lower plants amenable to transformation techniques, including angiosperms (monocotyledonous and dicotyledonous plants), gymnosperms, ferns, bryophytes, and multicellular algae.

[0037] The term "biomass," as used herein, refers to plant material that is processed to provide a product, *e.g.*, a biofuel such as ethanol, or livestock feed, or a cellulose for paper and pulp industry products. Such plant material can include whole plants, or parts of plants, *e.g.*, stems, leaves, branches, shoots, roots, tubers, and the like.

[0038] The term "increased cell wall deposition" in the context of galactan deposition refers to an increased amount of galactan in a cell wall that is produced in an engineered plant of the present invention as compared to a wild-type (*i.e.*, naturally occurring) plant. In the current invention, galactan deposition is typically considered to be increased when the amount of galactan in the cell wall is increased by at least 10%, at least 20, 30%, 40%, 50%, 60%, 70%, 80%, 90% or more relative to the amount of galactan in the cell wall in a wild-type plant. The amount of galactan can be assessed using any method known in the art, including using an antibody that specifically binds galactan or enzymatic or chemical analyses.

[0039] The term "saccharification reaction" refers to a process of converting biomass, usually cellulosic or lignocellulosic biomass, into monomeric sugars, such as glucose and xylose.

[0040] The term "soluble sugar" refers to monomeric, dimeric, or trimeric sugar that is produced from the saccharification of biomass.

[0041] The term "increased amount," when referring to an amount of sugar or soluble sugar obtained from an engineered plant of the present invention, refers to an increase in the amount or yield of sugar that is obtained from saccharification of biomass per amount of starting material, in comparison to corresponding biomass from a wild-type (*i.e.*, naturally occurring) plant. In the context of the present invention, "corresponding biomass from a wild-type plant" refers to plant material that is from the same part of the plant as the biomass from a plant engineered to have modified galactan levels. As understood in the art, increased amount or increased yield is based upon comparisons of the same amount of corresponding plant material.

[0042] The term "conversion reaction," as used herein, refers to a reaction that converts biomass into a form of bioenergy. Examples of conversion reactions include, but are not limited to, combustion (burning), gasification, pyrolysis, and polysaccharide hydrolysis (enzymatic or chemical).

[0043] The term "increased production," when referring to an amount of bioenergy production obtained from an engineered plant of the present invention, refers to an increased amount of bioenergy that is produced from subjecting biomass from an engineered plant to a conversion reaction (*e.g.*, combustion, gasification, pyrolysis, or polysaccharide hydrolysis) as compared to the amount of bioenergy that is produced from corresponding biomass from a wild-type (*i.e.*, naturally occurring) plant.

II. Introduction

[0044] Galactan is one of the major cell wall polysaccharides. Galactans are composed of hexoses that are easily fermented, which is in contrast to the hemicellulose xylan, which is the naturally most abundant non-cellulosic component of biomass. In addition, beta-1,4-galactan is likely to be more easily degraded by enzymes than xylan, which has a more complex structure and is crosslinked with other cell wall components. Prior to this invention, no biosynthetic enzyme has been identified for galactan. In one aspect, the invention provides a method of engineering plants to increase galactan content, *e.g.*, to improve biofuel potential. Plants can be engineered to overexpress galactan by genetically modifying a plant to overexpress one or more GALS genes as described herein. Typically, overexpresson is targeted to cell wall using a tissue-specific promoter. An example of a method for fine-tuning GALS expression to increase expression in the cell wall is taught in PCT/US2012/023182, which is incorporated by reference.

[0045] A plant that is engineered to overexpress GALS may also be engineered to overexpress a UDP-galactose epimerase (more commonly referred to as a UDP-glucose epimerase). Such epimerases are well known in the art. Examples of epimerase genes are described by Barber *et al., J. Biol. Chem.* 281:17276–17285, 2006 and Kotake *et al., Biochem. J.* 424:169–177, 2009, each of which is incorporated by reference. An example of an epimerase polypeptide sequence (Kotake *et al.*,) is provided in SEQ ID NO:45.

[0046] In a further aspect, a plant may be further modified to alter the enzymes that synthesize galactan substrates. Such enzymes could include UDP-glucose pyrophosphorylase and other non-specific UDP-sugar pyrophosphorylases.

[0047] In a further aspect, the invention provides a method of decreasing beta-1,4-galactan in a plant, *e.g.*, to increase pectin yield in a plant usch as a tuber. Decreasing galactan content can be achieved by inhibiting expression of at least one GALS gene in the plant.

[0048] The invention additionally provides methods of using genetically modified plants that overexpress or have reduced levels of GALS activity and methods of using such plants.

GALS nucleic acid sequences

[0049] The invention employs various routine recombinant nucleic acid techniques. Generally, the nomenclature and the laboratory procedures in recombinant DNA technology described below are those well known and commonly employed in the art. Many manuals that provide direction for performing recombinant DNA manipulations are available, *e.g.*, Sambrook & Russell, Molecular Cloning, A Laboratory Manual (3rd Ed, 2001); and Current Protocols in Molecular Biology (Ausubel, *et al.*, John Wiley and Sons, New York, 2009).

[0050] GALS nucleic acid and polypeptide sequences suitable for use in the invention include GALS nucleic acid sequences that encode a plant GALS polypeptide as illustrated in any of SEQ NO:1-39, or a substantially identical variants. Such a variant typically has at least 60%, often at least 70%, or at least 75%, 80%, 85%, or 90% identity to any one of SEQ ID NOS:1-39. In some embodiments, the nucleic acid encodes a GALS polypeptide of one of SEQ ID NOS:1-28, or a substantially identical variant thereof. Such a variant typically has at least 60%, often at least 70%, or at leaset 75%, 80%, 85%, or 90% identity to any one of SEQ ID NOS:1-28. In some embodiments, the nucleic acid encodes a GALS polypeptide of one of SEQ ID NOS:1-3, or a substantially identical variant thereof. Such a variant typically has at least 60%, often at least 70%, or at least 75%, 80%, 85%, or 90% identity to SEQ ID NO:1, SEQ ID NO:2, or SEQ ID NO:3.

[0051] A comparison of GALS sequences is provided in Figure 1. As shown in Figure 1, there are highly conserved regions of the polypeptide sequences. For example, the sequence (F/Y/V)G(N/S/T)AAALFV(L/Q)MGAYRGGP (the corresponding sequence is shaded and underlined in SEQ ID NO:1 of Figure 1) is highly conserved throughout diverse plant GALS sequences. Additional highly conserved sequences include SKPIHVYGKPWYKCEWISN, KILPDWGYYGRVYTVVVVNCTF, GG(K/R)LI(L/V/I), Y(D/E)YLYCGSSL(Y/F)G, REWMAYHAWFFG, SHFVFHDAGG, QNIRDQ, GYYYNQFLIVNDCLHRYRYAANWTFFDVDEY, FTIEQNPMS, WGFEKLLFK, RRDRKYAIQ, RYYHYHNSI, and ELCRE all of which are indicated in with reference to SEQ ID NO:1 shown in Figure 1 by underlining and shading of SEQ ID NO:1. These

conserved sequences are not strictly conserved 100% across the various plant protein sequences. One of skill can obtain a GALS variant by using the sequence alignments to identify residues within the conserved sequences that would be expected to retain GALS function as well as residues outside of the conserved regions that would be tolerant to substitution.

[0052] GALS activity can be assessed using any number of assays, including assays that evaluate transfer of galactose onto an acceptor. A convenient assay incubates an enzyme preparation in the presence of UDP-Gal and beta-1,4-galacto-oligosaccharides, *e.g.*, beta-1,4-galactopentaose. The products consisting of the acceptor with addition of one or more galactose units can then be characterized by any number of methods. The simplest is to remove unincorporated UDP-Gal by ion exchange chromatography and analyze the produt by liquid scintillation counting (provided that radiolabelled UDP-Gal was used). An alternative method is to use mass spectrometry, paper chromatography or thin layer chromatography. I f a charged group is linked to the reducing end of the acceptor or product, it can also be analyzed by capillary electrophoresis or gel electrophoresis. If the charged group is also fluorescent it can be easily detected. A non-charged fluorescent labeled galacto-oligosaccharide has been used as acceptor in such assays where the products were separated by HPLC and detected by the fluorescent group (Ishii *et al., Planta* 219:310-318, 2004).

[0053] Genetic modification of a plant to overexpress GALS is often performed in conjunction with modifying the plant to overexpress UDP-galactose epimerase (UGE) (EC 5.1.3.2). UDP-galactose epimerase nucleic acid and polypeptide seuqences are well known in the art. Examples of UGE sequences that can be overexpressed are provided in U.S. Patent Application Publication Nos. 20030073828; 20070028332; and described by Barber *et al. J. Biol. Chem.* 281:17276-17285, 2006; and Kotake *et al., Biochem J.* 424:169-177, 2009; and Oomen *et al. Plant Science*166:1097-1104, 2004, each of which is incorporated by reference.

[0054] Isolation or generation GALS polynucleotide sequences (or UGE sequences) can be accomplished by a number of techniques. Cloning and expression of such technique will be addressed in the context of GALS genes. However, the same techniques can be used to isolate and express UGE family. In some embdoiments, oligonucleotide probes based on the sequences disclosed here can be used to identify the desired polynucleotide in a cDNA or genomic DNA library from a desired plant species. Probes may be used to hybridize with genomic DNA or cDNA sequences to isolate homologous genes in the same or different plant species.

[0055] Alternatively, the nucleic acids of interest can be amplified from nucleic acid samples using routine amplification techniques. For instance, PCR may be used to amplify the sequences of the genes directly from mRNA, from cDNA, from genomic libraries or cDNA libraries. PCR and other *in vitro* amplification methods may also be useful, for example, to clone nucleic acid sequences that code for proteins to be expressed, to make nucleic acids to use as probes for detecting the presence of the desired mRNA in samples, for nucleic acid sequencing, or for other purposes.

[0056] Appropriate primers and probes for identifying a GALS gene from plant cells such as moss or spikemoss, can be generated from comparisons of the sequences provided herein. For a general overview of PCR see PCR Protocols: A Guide to Methods and Applications. (Innis, M, Gelfand, D., Sninsky, J. and White, T., eds.), Academic Press, San Diego (1990).

[0057] GALS nucleic acid sequences for use in the invention includes genes and gene products identified and characterized by techniques such as hybridization and/or sequence analysis using exemplary nucleic acid sequences, *e.g.*, SEQ ID NO:42, SEQ ID NO:43, or SEQ ID NO:44.

Preparation of recombinant vectors

[0058] To use isolated sequences in the above techniques, recombinant DNA vectors suitable for transformation of plant cells such as crop plant cells are prepared. Techniques for transformation are well known and described in the technical and scientific literature. For example, a DNA sequence encoding a GALS polypepitde (described in further detail below), can be combined with transcriptional and other regulatory sequences which will direct the transcription of the sequence from the gene in the intended cells, *e.g.*, grass or other crop plant cells. In some embodiments, an expression vector that comprises an expression cassette that comprises the GALS gene further comprises a promoter operably linked to the GALS gene. In other embodiments, a promoter and/or other regulatory elements that direct transcription of the GALS gene are endogenous to the plant and an expression cassette comprising the GALS gene is introduced, *e.g.*, by homologous recombination, such that the heterologous GALS gene is operably linked to an endogenous promoter and is expression driven by the endogenous promoter.

[0059] Regulatory sequences include promoters, which may be either constitutive or inducible, or tissue-specific.

Tissue-Specific Promoters

[0060] In some embodiments, a plant promoter to direct expression of a GALS gene in a specific tissue is employed (tissue-specific promoters). Tissue specific promoters are transcriptional control elements that are only active in particular cells or tissues at specific times during plant development, such as in vegetative tissues or reproductive tissues.

[0061] Examples of tissue-specific promoters under developmental control include promoters that initiate transcription only (or primarily only) in certain tissues, such as vegetative tissues, cell walls, including e.g., roots or leaves. A variety of promoters specifically active in vegetative tissues, such as leaves, stems, roots and tubers are known. For example, promoters controlling patatin, the major storage protein of the potato tuber, can be used (see, *e.g.*, Kim, *Plant Mol. Biol.* 26:603-615, 1994; Martin, *Plant J.* 11:53-62, 1997). The ORF13 promoter from *Agrobacterium rhizogenes* that exhibits high activity in roots can also be used (Hansen, *Mol. Gen. Genet.* 254:337-343, 1997). Other useful vegetative tissue-specific promoters include: the tarin promoter of the gene encoding a globulin from a major taro (*Colocasia esculenta L. Schott*) corm protein family, tarin (Bezerra, *Plant Mol. Biol.* 28:137-144, 1995); the curculin promoter active during taro corm development (de Castro, *Plant Cell* 4:1549-1559, 1992) and the promoter for the tobacco root-specific gene TobRB7, whose expression is localized to root meristem and immature central cylinder regions (Yamamoto, Plant Cell 3:371-382, 1991).

[0062] Leaf-specific promoters, such as the ribulose biphosphate carboxylase (RBCS) promoters can be used. For example, the tomato RBCS1, RBCS2 and RBCS3A genes are expressed in leaves and light-grown seedlings, only RBCS1 and RBCS2 are expressed in developing tomato fruits (Meier, *FEBS Lett.* 415:91-95, 1997). A ribulose bisphosphate carboxylase promoters expressed almost exclusively in mesophyll cells in leaf blades and leaf sheaths at high levels (*e.g.*, Matsuoka, *Plant J.* 6:311-319, 1994), can be used. Another leaf-specific promoter is the light harvesting chlorophyll a/b binding protein gene promoter (see, *e.g.*, Shiina, *Plant Physiol.* 115:477-483, 1997; Casal, *Plant Physiol.* 116:1533-1538, 1998). The *Arabidopsis thaliana* myb-related gene promoter (Atmyb5) (Li, *et al., FEBS Lett.* 379:117-121 1996), is leaf-specific. The Atmyb5 promoter is expressed in developing leaf trichomes, stipules, and epidermal cells on the margins of young rosette and cauline leaves, and in immature seeds. Atmyb5 mRNA appears between fertilization and the 16 cell stage of embryo development and persists beyond the heart stage. A leaf promoter identified in maize (*e.g.*, Busk *et al.*, *Plant J.* 11:1285-1295, 1997) can also be used.

[0063] Another class of useful vegetative tissue-specific promoters are meristematic (root tip and shoot apex) promoters. For example, the "SHOOTMERISTEMLESS" and "SCARECROW" promoters, which are active in the developing shoot or root apical meristems, (*e.g.*, Di Laurenzio, *et al., Cell* 86:423-433, 1996; and, Long, et al., Nature 379:66-69, 1996); can be used. Another useful promoter is that which controls the expression of 3-hydroxy-3- methylglutaryl coenzyme A reductase HMG2 gene, whose expression is restricted to meristematic and floral (secretory zone of the stigma, mature pollen grains, gynoecium vascular tissue, and fertilized ovules) tissues (see, *e.g.*, Enjuto, Plant Cell. 7:517-527, 1995). Also useful are kn1-related genes from maize and other species which show meristem-specific expression, (see, *e.g.*, Granger, *Plant Mol. Biol.* 31:373-378, 1996; Kerstetter, *Plant Cell* 6:1877-1887, 1994; Hake, Philos. Trans. R. Soc. Lond. B. Biol. Sci. 350:45-51, 1995). For example, the *Arabidopsis thaliana KNAT1* promoter (see, *e.g.*, Lincoln, *Plant Cell* 6:1859-1876, 1994) can be used.

[0064] In some embodiments, the promoter is substantially identical to the native promoter of a promoter that drives expression of a gene involved in secondary wall deposition. Examples of such promoters are promoters from IRX1, IRX3, IRX5, IRX8, IRX9, IRX14, IRX7, IRX10, GAUT13, or GAUT14 genes. Specific expression in fiber cells can be accomplished by using a promoter such as the NST1 promoter and specific expression in vessels can be accomplished by using a promoter such as VND6 or VND7. (See, *e.g.*, PCT/US2012/023182 for illustrative promoter sequences).

[0065] One of skill will recognize that a tissue-specific promoter may drive expression of operably linked sequences in tissues other than the target tissue. Thus, as used herein a tissue-specific promoter is one that drives expression preferentially in the target tissue, but may also lead to some expression in other tissues as well.

Constitutive Promoters

[0066] A promoter, or an active fragment thereof, can be employed which will direct expression of a nucleic acid encoding a fusion protein of the invention, in all or most transformed cells or tissues, e.g. as those of a regenerated plant. Such promoters are referred to herein as "constitutive" promoters and are active under most environmental conditions and states of development or cell differentiation. Examples of constitutive promoters include those from viruses which infect plants, such as the cauliflower mosaic virus (CaMV) 35S transcription initiation region (see, *e.g.*, Dagless, *Arch. Virol.* 142:183-191, 1997); the 1'- or 2'- promoter derived from T-DNA of *Agrobacterium tumefaciens* (see, *e.g.*, Mengiste *supra*

(1997); O'Grady, Plant Mol. Biol. 29:99-108, 1995); the promoter of the tobacco mosaic virus; the promoter of Figwort mosaic virus (see, *e.g.*, Maiti, *Transgenic Res.* 6:143-156, 1997); actin promoters, such as the *Arabidopsis* actin gene promoter (see, *e.g.*, Huang, *Plant Mol. Biol.* 33:125-139, 1997); alcohol dehydrogenase (Adh) gene promoters (see, *e.g.*, Millar, *Plant Mol. Biol.* 31:897-904, 1996); ACT11 from *Arabidopsis* (Huang *et al., Plant Mol. Biol.* 33:125-139, 1996), Cat3 from *Arabidopsis* (GenBank No. U43147, Zhong *et al., Mol. Gen. Genet.* 251:196-203, 1996), the gene encoding stearoyl-acyl carrier protein desaturase from *Brassica napus* (Genbank No. X74782, Solocombe *et al., Plant Physiol.* 104:1167-1176, 1994), GPc1 from maize (GenBank No. X15596, Martinez *et al., J. Mol. Biol.* 208:551-565, 1989), Gpc2 from maize (GenBank No. U45855, Manjunath *et al., Plant Mol. Biol.* 33:97-112, 1997), other transcription initiation regions from various plant genes known to those of skill. See also Holtorf, "Comparison of different constitutive and inducible promoters for the overexpression of transgenes in *Arabidopsis thaliana*," *Plant Mol. Biol.* 29:637-646, 1995).

Inducible Promoters

[0067] In some embodiments, a plant promoter may direct expression of the nucleic acids under the influence of changing environmental conditions or developmental conditions. Examples of environmental conditions that may effect transcription by inducible promoters include anaerobic conditions, elevated temperature, drought or other environmental stress, or the presence of light. Examples of developmental conditions that may effect transcription by inducible promoters include senescence and embryogenesis. Such promoters are referred to herein as "inducible" promoters. For example, the invention can incorporate drought-specific promoter such as the drought-inducible promoter of maize (Busk *et al., Plant J,* 11: 1285-95, 1997); or alternatively the cold, drought, and high salt inducible promoter from potato (Kirch *Plant Mol. Biol.* 33:897-909, 1997).

[0068] Suitable promoters responding to biotic or abiotic stress conditions include the pathogen inducible PRP1-gene promoter (Ward *et al., Plant. Mol. Biol.* 22:361-366, 1993), the heat inducible hsp80-promoter from tomato (U.S. Pat. No. 5,187,267), cold inducible alpha-amylase promoter from potato (PCT Publication No. WO 96/12814) or the wound-inducible pinII-promoter (European Patent No. 375091). For other examples of drought, cold, and salt-inducible promoters, such as the RD29A promoter, see, *e.g.*, Yamaguchi-Shinozalei *et al., Mol. Gen. Genet.* 236:331-340, 1993 are also known.

[0069] Alternatively, plant promoters which are inducible upon exposure to plant hormones, such as auxins, may be used to express GALS genes. For example, the invention

can use the auxin-response elements E1 promoter fragment (AuxREs) in the soybean (*Glycine max L.*) (Liu, *Plant Physiol.* 115:397-407, 1997); the auxin-responsive *Arabidopsis* GST6 promoter (also responsive to salicylic acid and hydrogen peroxide) (Chen, Plant J. 10: 955-966, 1996); the auxin-inducible parC promoter from tobacco (Sakai, 37:906-913, 1996); a plant biotin response element (Streit, *Mol. Plant Microbe Interact.* 10:933-937, 1997); and, the promoter responsive to the stress hormone abscisic acid (Sheen, *Science* 274:1900-1902, 1996).

[0070] Plant promoters inducible upon exposure to chemicals reagents that may be applied to the plant, such as herbicides or antibiotics, are also useful for expressing a GALS gene in accordance with the ivvention. For example, the maize In2-2 promoter, activated by benzenesulfonamide herbicide safeners, can be used (De Veylder, *Plant Cell Physiol.* 38:568-577, 19997); application of different herbicide safeners induces distinct gene expression patterns, including expression in the root, hydathodes, and the shoot apical meristem. A GALS coding sequence can also be under the control of, *e.g.*, a tetracycline-inducible promoter, such as described with transgenic tobacco plants containing the *Avena sativa L.* (oat) arginine decarboxylase gene (Masgrau, *Plant J.* 11:465-473, 1997); or, a salicylic acid-responsive element (Stange, *Plant J.* 11:1315-1324, 1997; Uknes *et al., Plant Cel*l 5:159-169, 1993); Bi *et al., Plant J.* 8:235-245, 1995).

[0071] Examples of useful inducible regulatory elements include copper-inducible regulatory elements (Mett *et al., Proc. Natl. Acad. Sci. USA* 90:4567-4571, 1993); Furst *et al., Cell* 55:705-717, 1988); tetracycline and chlor-tetracycline-inducible regulatory elements (Gatz *et al., Plant J.* 2:397-404, 1992); Röder e*t al., Mol. Gen. Genet.* 243:32-38, 1994); Gatz, *Meth. Cell Biol.* 50:411-424, 1995); ecdysone inducible regulatory elements (Christopherson *et al., Proc. Natl. Acad. Sci. USA* 89:6314-6318, 1992; Kreutzweiser *et al., Ecotoxicol. Environ. Safety* 28:14-24, 1994); heat shock inducible regulatory elements (Takahashi *et al., Plant Physiol.* 99:383-390, 1992; Yabe *et al., Plant Cell Physiol.* 35:1207-1219, 1994; Ueda *et al., Mol. Gen. Genet.* 250:533-539, 1996); and lac operon elements, which are used in combination with a constitutively expressed lac repressor to confer, for example, IPTG-inducible expression (Wilde *et al., EMBO J.* 11:1251-1259, 1992). An inducible regulatory element useful in the transgenic plants of the invention also can be, for example, a nitrate-inducible promoter derived from the spinach nitrite reductase gene (Back et al., Plant Mol. Biol. 17:9 (1991)) or a light-inducible promoter, such as that associated with the small subunit of RuBP carboxylase or the LHCP gene families (Feinbaum et al., Mol. Gen. Genet. 226:449 (1991); Lam and Chua, Science 248:471 (1990)).

Expression using a positive feed back loop

[0072] In further embodiments, a plant can be engineered to overexpress GALS using a positive feedback loop to express GALS in a desired tissue. In such an embodiment, a promoter for use in a GALS expression construct is responsive to a transcription factor that mediates expression in the desired tissue. The GALS expression construct is used in a genentically modified plant comprising an expression construct encoding a transcription factor where expression is also driven by a promoter that is responsive to the transcription factor. Examples of such expression systems are provided in PCT/US2012/023182.

[0073] In some embodiments in which a positive feed back loop is employed, the plant is genetically modified to express a transcription factor that regulates the production of secondary cell wall. Examples of such transcription factors include NST1, NST2, NST3, SND2, SND3, MYB103, MBY85, MYB46, MYB83, MYB58, and MYB63 (*See, e.g.*, Mitsuda *et al.*, *Plant Cell* 17:2993-3006 (2005); Mitsuda *et al.*, *Plant Cell* 19:270-80 (2007); Ohashi-Ito *et al.*, *Plant Cell* 22:3461-73 (2010); Zhong *et al.*, *Plant Cell* 20:2763-82 (2008); Zhong *et al.*, *Plant Cell* 19:2776-92 (2007); Ko *et al.*, *Plant J.* 60:649-65 (2009); and McCarthy *et al.*, *Plant Cell Physiol.* 50:1950-64 (2009)).

[0074] Illustrative examples of gene and protein sequences and/or accession numbers for NST1, NST2, NST3, SND2, SND3, MYB103, MBY85, MYB46, MYB83, MYB58, and MYB63 are provided in PCT/US2012/023182.

[0075] In some embodiments, the polynucleotide encoding the transcription factor that regulates secondary cell wall production is operably linked to a promoter that is a downstream target of the transcription factor. Similarly, the GALS nucleic acid sequence is also linked to a promoter that is a downstream target of the transcription factor. The promoter may be the same promoter or different promoters. In such an embodiment, a promoter is suitable for use with the transcription factor that regulates secondary cell wall production if expression of the promoter is induced, directly or indirectly, by the transcription factor to be expressed, and if the promoter is expressed in the desired location, *e.g.*, the stem of the plant.

[0076] In some embodiments, a native IRX1, IRX3, IRX5, IRX8, IRX9, IRX14, IRX7, or IRX10, GAUT13, or GAUT14 promoter, or active variant thereof, is employed.

Additional embodimens for expressing GALS

[0077] In another embodiment, the GALS polynucleotide is expressed through a transposable element. This allows for constitutive, yet periodic and infrequent expression of the constitutively active polypeptide. The invention also provides for use of tissue-specific promoters derived from viruses including, e.g., the tobamovirus subgenomic promoter (Kumagai, *Proc. Natl. Acad. Sci. USA* 92:1679-1683, 1995); the rice tungro bacilliform virus (RTBV), which replicates only in phloem cells in infected rice plants, with its promoter which drives strong phloem-specific reporter gene expression; the cassava vein mosaic virus (CVMV) promoter, with highest activity in vascular elements, in leaf mesophyll cells, and in root tips (Verdaguer, *Plant Mol. Biol.* 31:1129-1139, 1996).

[0078] A vector comprising GALS nucleic acid sequences will typically comprise a marker gene that confers a selectable phenotype on the cell to which it is introduced. Such markers are known. For example, the marker may encode antibiotic resistance, such as resistance to kanamycin, G418, bleomycin, hygromycin, and the like.

[0079] GALS nucleic acid sequences of the invention are expressed recombinantly in plant cells as described. As appreciated by one of skill in the art, expression constructs can be designed taking into account such properties as codon usage frequencies of the plant in which the GALS nucleic acid is to be expressed. Codon usage frequencies can be tabulated using known methods (see, e.g., Nakamura *et al. Nucl. Acids Res.* 28:292, 2000). Codon usage frequency tables are available in the art (*e.g.*, from the Codon Usage Database at the internet site www.kazusa.or.jp/codon/.)

[0080] When two or more of GALS, UGE or transcription factors are expressed in combination, they can be expressed from individual promoters. In some embodiments, two or more proteins are expressed from a single promoter, *e.g.*, by incorporating a 2A domain between the two coding sequences.

[0081] Additional sequence modifications may be made that are also known to enhance gene expression in a plant. These include elimination of sequences encoding spurious polyadenylation signals, exon-intron splice site signals, transposon-like repeats, and other such well-characterized sequences that may be deleterious to gene expression. The G-C content of the sequence may be adjusted to levels average for a given cellular host, as calculated by reference to known genes expressed in the host cell. When possible, the sequence may also be modified to avoid predicted hairpin secondary mRNA structures.

Production of Transgenic Plants

[0082] As detailed herein, the present invention provides for transgenic plants comprising recombinant expression cassettes either for expressing heterologous GALS. It should be recognized that the term "transgenic plants" as used here encompasses the plant or plant cell in which the expression cassette is introduced as well as progeny of such plants or plant cells that contain the expression cassette, including the progeny that have the expression cassette stably integrated in a chromosome.

[0083] Once an expression cassette comprising a polynucleotide encoding a GALS (or a polynucleotide sequence designed to suppress or inhibit GALS expression as described below) has been constructed, standard techniques may be used to introduce the polynucleotide into a plant in order to modify gene expression. *See, e.g.*, protocols described in Ammirato et al. (1984) Handbook of Plant Cell Culture--Crop Species. Macmillan Publ. Co. Shimamoto et al. (1989) Nature 338:274-276; Fromm et al. (1990) Bio/Technology 8:833-839; and Vasil et al. (1990) Bio/Technology 8:429-434.

[0084] Transformation and regeneration of plants is known in the art, and the selection of the most appropriate transformation technique will be determined by the practitioner. Suitable methods may include, but are not limited to: electroporation of plant protoplasts; liposome-mediated transformation; polyethylene glycol (PEG) mediated transformation; transformation using viruses; micro-injection of plant cells; micro-projectile bombardment of plant cells; vacuum infiltration; and *Agrobacterium tumeficien*s mediated transformation. Transformation means introducing a nucleotide sequence in a plant in a manner to cause stable or transient expression of the sequence. Examples of these methods in various plants include: U.S. Pat. Nos. 5,571,706; 5,677,175; 5,510,471; 5,750,386; 5,597,945; 5,589,615; 5,750,871; 5,268,526; 5,780,708; 5,538,880; 5,773,269; 5,736,369 and 5,610,042.

[0085] Transformed plant cells derived by any of the above transformation techniques can be cultured to regenerate a whole plant that possesses the transformed genotype and thus the desired phenotype such as enhanced drought-resistance. Such regeneration techniques rely on manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker which has been introduced together with the desired nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., Protoplasts Isolation and Culture, Handbook of Plant Cell Culture, pp. 124-176, MacMillilan Publishing Company, New York, 1983; and Binding, Regeneration of Plants, Plant Protoplasts, pp. 21-73, CRC Press, Boca Raton, 1985. Regeneration can also be

obtained from plant callus, explants, organs, or parts thereof. Such regeneration techniques are described generally, *e.g.*, in Klee *et al. Ann. Rev. of Plant Phys.* 38:467-486, 1987.

[0086] One of skill will recognize that after the expression cassette is stably incorporated in transgenic plants and confirmed to be operable, it can be introduced into other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

[0087] The techniques described herein for obtaining and expressing GALS nucleic acid sequences in plant cells can also be employed to express nucleic acid sequences that encode UGE family members in order to modify plants to overexpress UGE proteins.

[0088] The expression constructs of the invention can be used to increase the galactan content of cell walls of essentially any plant. The plant may be a monocotyledonous plant or a dicotyledonous plant. In some embodiments of the invention, the plant is a green field plant. In some embodiments, the plant is a gymnosperm or conifer. Thus, the invention has use over a broad range of plants, including species from the genera Asparagus, Atropa, Avena, Brassica, Cannabis, Citrus, Citrullus, Camelina, Capsicum, Cucumis, Cucurbita, Daucus, Fragaria, Glycine, Gossypium, Helianthus, Heterocallis, Hordeum, Hyoscyamus, Lactuca, Linum, Lolium, Lycopersicon, Malus, Manihot, Majorana, Medicago, Nicotiana, Oryza, Panicum, Pannesetum, Persea, Pisum, Pyrus, Prunus, Raphanus, Secale, Senecio, Sinapis, Solanum, Sorghum, Trigonella, Triticum, Vitis, Vigna, and, Zea. In some embodiments, the plant is corn, switchgrass, sorghum, miscanthus, sugarcane, poplar, pine, wheat, rice, soy, cotton, barley, turf grass, tobacco, potato, bamboo, rape, sugar beet, sunflower, willow, and eucalyptus. In further embodiments, the plant is reed canarygrass (*Phalaris arundinacea*), *Miscanthus x giganteus*, *Miscanthus* sp., sericea lespedeza (*Lespedeza cuneata*), millet, ryegrass (*Lolium multiflorum*, *Lolium* sp.), timothy, Kochia (*Kochia scoparia*), forage soybeans, alfalfa, clover, sunn hemp, kenaf, bahiagrass, bermudagrass, dallisgrass, pangolagrass, big bluestem, indiangrass, fescue (*Festuca* sp.), *Dactylis* sp., *Brachypodium distachyon*, smooth bromegrass, orchardgrass, or Kentucky bluegrass among others. In some embodiments, the plant is an ornamental plant. In some embodiment, the plant is a vegetable- or fruit-producing plant. In some embodiments, the plant is a plant that is suitable for generating biomass, including plants as noted above, *e.g.*, Arabidopsis, poplar, eucalyptus, rice, corn, switchgrass, sorghum, millet, miscanthus, sugarcane, pine, alfalfa, wheat, soy, barley, turfgrass, tobacco, hemp, bamboo, rape, sunflower, willow, Jatropha, and Brachypodium.

[0089] In some embodiments, the plant into which the expression construct comprising a nucleic acid sequence that encodes GALS (or that is designed to inhibit expresson of GALS) is introduced is the same species of plant from which the GALS sequence, and/ or the promoter driving expression of the GALS sequence, is obtained. In some embodiments, the plant into which the expression construct is introduced is a different species of plant compared to the species from which the GALS and/or promoter sequence was obtained.

[0090] Plants that overexpress GALS can be identified using any known assay, including analysis of RNA, protein, or galactan composition With respect to this aspect of the invention, the plants have enhanced galactan levels. Galactan levels can be determined directly or indirectly. For example, in some embodiments, galactan is assessed using an immunoassay employing an antibody that specifically binds beta-1,4, galactan to determine galactan levels. In some embodiments, GALS enzymatic activity can be directly measured in a plant by determining the activity of the enzyme to transfer galactose to an acceptor.

Modification of plants to decrease galactan production

[0091] In one aspect, the invention also provides a plant in which expression of GALS is inhibited, thereby resulting in reduced elvels of galactan in the plant. In some embodiments, the plant is modified to have a level of GALS activity that is reduced throughout the entire plant. In some embodiments, the plant is modified to reduce GALS activity in a subset of cells or tissues of the plant. The genetic background of the plant can be modified according to any method known in the art, such as antisense, siRNA, microRNA, dsRNA, sense suppression, mutagenesis, or use of a dominant negative inhibition strategy. In some embodiments, the level of expression of the protein is reduced.

Gene silencing techniques

[0092] In some embodiments, expression of a GALS is inhibited by an antisense oligonucleotide. In antisense technology, a nucleic acid segment from the desired gene is cloned and operably linked to a promoter such that the antisense strand of RNA will be transcribed. The expression cassette is then transformed into plants and the antisense strand of RNA is produced. In plant cells, it has been suggested that antisense RNA inhibits gene expression by preventing the accumulation of mRNA which encodes the enzyme of interest, see, e.g., Sheehy *et al.*, *Proc. Nat. Acad. Sci. USA*, 85:8805-8809 (1988); Pnueli *et al.*, *The Plant Cell* 6:175-186 (1994); and Hiatt *et al.*, U.S. Patent No. 4,801,340.

[0093] The antisense nucleic acid sequence transformed into plants will be substantially identical to at least a portion of the endogenous gene or genes to be repressed. The sequence, however, does not have to be perfectly identical to inhibit expression. Thus, an antisense or sense nucleic acid molecule encoding only a portion of a GALS-encoding sequence can be useful for producing a plant in which expression of GALS is inhibited. For antisense suppression, the introduced sequence also need not be full length relative to either the primary transcription product or fully processed mRNA. Generally, higher homology can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective. In some embodiments, a sequence of at least, *e.g.*, 20, 25, 30, 50, 100, 200, or more continuous nucleotides (up to mRNA full length) substantially identical to a GALS mRNA, or a complement thereof, can be used.

[0094] Catalytic RNA molecules or ribozymes can also be used to inhibit expression of a gene encoding a GALS polypeptide. It is possible to design ribozymes that specifically pair with virtually any target RNA and cleave the phosphodiester backbone at a specific location, thereby functionally inactivating the target RNA. In carrying out this cleavage, the ribozyme is not itself altered, and is thus capable of recycling and cleaving other molecules, making it a true enzyme. The inclusion of ribozyme sequences within antisense RNAs confers RNA-cleaving activity upon them, thereby increasing the activity of the constructs.

[0095] A number of classes of ribozymes have been identified. One class of ribozymes is derived from a number of small circular RNAs that are capable of self-cleavage and replication in plants. The RNAs replicate either alone (viroid RNAs) or with a helper virus (satellite RNAs). Examples include RNAs from avocado sunblotch viroid and the satellite RNAs from tobacco ringspot virus, lucerne transient streak virus, velvet tobacco mottle virus, solanum nodiflorum mottle virus and subterranean clover mottle virus. The design and use of target RNA-specific ribozymes is described in Haseloff *et al. Nature*, 334:585-591 (1988).

[0096] Another method by which expression of a gene encoding a GALS polypeptide can be inhibited is by sense suppression (also known as co-suppression). Introduction of expression cassettes in which a nucleic acid is configured in the sense orientation with respect to the promoter has been shown to be an effective means by which to block the transcription of target genes. For an example of the use of this method to modulate expression of endogenous genes, see Napoli *et al., The Plant Cell* 2:279-289 (1990); Flavell, *Proc. Natl. Acad. Sci., USA* 91:3490-3496 (1994); Kooter and Mol, *Current Opin. Biol.* 4:166-171 (1993); and U.S. Patents Nos. 5,034,323, 5,231,020, and 5,283,184.

[0097] Generally, where inhibition of expression is desired, some transcription of the introduced sequence occurs. The effect may occur where the introduced sequence contains no coding sequence *per se*, but only intron or untranslated sequences homologous to sequences present in the primary transcript of the endogenous sequence. The introduced sequence generally will be substantially identical to the endogenous GALS sequence intended to be repressed. This minimal identity will typically be greater than about 65%, but a higher identity can exert a more effective repression of expression of the endogenous sequences. In some embodiments, sequences with substantially greater identity are used, *e.g.*, at least about 80%, at least about 95%, or 100% identity are used. As with antisense regulation, further discussed below, the effect can be designed and tested to apply to any other proteins within a similar family of genes exhibiting homology or substantial homology.

[0098] For sense suppression, the introduced sequence in the expression cassette, needing less than absolute identity, also need not be full length, relative to either the primary transcription product or fully processed mRNA. Furthermore, the introduced sequence need not have the same intron or exon pattern, and identity of non-coding segments will be equally effective. In some embodiments, a sequence of the size ranges noted above for antisense regulation is used, i.e., 30-40, or at least about 20, 50, 100, 200, 500 or more nucleotides.

[0099] Endogenous gene expression may also be suppressed by means of RNA interference (RNAi) (and indeed co-suppression can be considered a type of RNAi), which uses a double-stranded RNA having a sequence identical or similar to the sequence of the target gene. RNAi is the phenomenon in which when a double-stranded RNA having a sequence identical or similar to that of the target gene is introduced into a cell, the expressions of both the inserted exogenous gene and target endogenous gene are suppressed. The double-stranded RNA may be formed from two separate complementary RNAs or may be a single RNA with internally complementary sequences that form a double-stranded RNA. Although complete details of the mechanism of RNAi are still unknown, it is considered that the introduced double-stranded RNA is initially cleaved into small fragments, which then serve as indexes of the target gene in some manner, thereby degrading the target gene. RNAi is known to be also effective in plants (see, e.g., Chuang, C. F. & Meyerowitz, E. M., *Proc. Natl. Acad. Sci. USA* 97: 4985 (2000); Waterhouse *et al., Proc. Natl. Acad. Sci. USA* 95:13959-13964 (1998); Tabara *et al. Science* 282:430-431 (1998); Matthew, *Comp Funct Genom* 5: 240–244 (2004); Lu, *et al., Nucleic Acids Res.* 32(21):e171 (2004)).

[0100] Thus, in some embodiments, inhibition of a gene encoding a GALS polypeptide is accomplished using RNAi techniques. For example, to achieve suppression of the expression

of a DNA encoding a protein using RNAi, a double-stranded RNA having the sequence of a DNA encoding the protein, or a substantially similar sequence thereof (including those engineered not to translate the protein) or fragment thereof, is introduced into a plant of interest. As used herein, RNAi and dsRNA both refer to gene-specific silencing that is induced by the introduction of a double-stranded RNA molecule, see e.g., U.S. Pat. Nos. 6,506,559 and 6,573,099, and includes reference to a molecule that has a region that is double-stranded, *e.g.*, a short hairpin RNA molecule. The resulting plants may then be screened for a phenotype associated with the target protein, for example, screening for an increase in the extractability of sugar from the plants as compared to wild-type plants, and/or by monitoring steady-state RNA levels for transcripts encoding the protein. Although the genes used for RNAi need not be completely identical to the target gene, they may be at least 70%, 80%, 90%, 95% or more identical to the target gene sequence. *See, e.g.*, U.S,. Patent Publication No. 2004/0029283. The constructs encoding an RNA molecule with a stem-loop structure that is unrelated to the target gene and that is positioned distally to a sequence specific for the gene of interest may also be used to inhibit target gene expression. *See, e.g.*, U.S. Patent Publication No. 2003/0221211.

[0101] The RNAi polynucleotides may encompass the full-length target RNA or may correspond to a fragment of the target RNA. In some cases, the fragment will have fewer than 100, 200, 300, 400, 500 600, 700, 800, 900 or 1,000 nucleotides corresponding to the target sequence. In addition, in some embodiments, these fragments are at least, *e.g.*, 50, 100, 150, 200, or more nucleotides in length. In some cases, fragments for use in RNAi will be at least substantially similar to regions of a target protein that do not occur in other proteins in the organism or may be selected to have as little similarity to other organism transcripts as possible, *e.g.*, selected by comparison to sequences in analyzing publicly-available sequence databases.

[0102] Expression vectors that continually express siRNA in transiently- and stably-transfected have been engineered to express small hairpin RNAs, which get processed *in vivo* into siRNAs molecules capable of carrying out gene-specific silencing (Brummelkamp *et al.*, *Science* 296:550-553 (2002), and Paddison, *et al.*, *Genes & Dev.* 16:948-958 (2002)). Post-transcriptional gene silencing by double-stranded RNA is discussed in further detail by Hammond et al. *Nature Rev Gen* 2: 110-119 (2001), Fire *et al. Nature* 391: 806-811 (1998) and Timmons and Fire *Nature* 395: 854 (1998).

[0103] Yet another way to suppress expression of an endogenous GALS gene is by recombinant expression of a microRNA that suppresses a target GALS. Artificial

microRNAs are single-stranded RNAs (e.g., between 18-25-mers, generally 21-mers), that are not normally found in plants and that are processed from endogenous miRNA precursors. Their sequences are designed according to the determinants of plant miRNA target selection, such that the artificial microRNA specifically silences its intended target gene(s) and are generally described in Schwab *et al*, *The Plant Cell* 18:1121-1133 (2006) as well as the internet-based methods of designing such microRNAs as described therein. *See also*, US Patent Publication No. 2008/0313773.

[0104] Another example of a method to reduce levels of GALS employs riboswitch techniques (see, e.g., U.S. Patent Application Publication Nos. US20100286082, and US20110245326).

Plants having mutant backgrounds

[0105] In some embodiments, the level of expression of GALS is reduced by generating a plant that has a mutation in a gene encoding the GALS enzyme. One method for abolishing or decreasing the expression of a gene encoding GALS is by insertion mutagenesis using the T-DNA of *Agrobacterium tumefaciens*. After generating the insertion mutants, the mutants can be screened to identify those containing the insertion in the gene of interest. Mutants containing a single mutation event at the desired gene may be crossed to generate homozygous plants for the mutation (Koncz et al. (1992) Methods in Arabidopsis Research. World Scientific).

[0106] Alternatively, random mutagenesis approaches may be used to generate new alleles that will generate truncated or defective (non-functional or poorly active) enzymes or unstable RNA, or to disrupt or "knock-out" the expression of a gene encoding a GALS enzyme using either chemical or insertional mutagenesis or irradiation. For example, a procedure known as TILLING (see, *e.g.* Colbert *et al., Plant Physiol* 126:480-484, 2001; McCallum *et al., Nature Biotechnology* 18:455-457, 2000). may be used. In this method, mutations are induced in the seed of a plant of interest. The resulting plants are grown and self-fertilized, and the progeny are assessed, *e.g.,* by PCR, to identify whether a mutated plant has a mutation in the gene of interest, or by evaluating whether the plant has reduced galactan content in a part of the plant that expressed the gene of interest.

[0107] An expression cassette comprising a polynucleotide encoding the GALS, or transcription factor regulating the production of secondary cell wall and operably linked to a promoter, as described herein, can be expressed in various kinds of plants. The plant may be

a monocotyledonous plant or a dicotyledonous plant. In some embodiments of the invention, the plant is a green field plant. In some embodiments, the plant is a gymnosperm or conifer.

[0108] In some embodiments, the plant is a plant that is suitable for generating biomass. Examples of suitable plants include, but are not limited to, Arabidopsis, poplar, eucalyptus, rice, corn, switchgrass, sorghum, millet, miscanthus, sugarcane, pine, alfalfa, wheat, soy, barley, turfgrass, tobacco, hemp, bamboo, rape, sunflower, willow, Jatropha, and Brachypodium.

[0109] In some embodiments, the plant into which the expression cassette is introduced is the same species of plant as the promoter and/or as the polynucleotide encoding GALS or transcription factor (*e.g.*, a vessel-specific promoter, GALS enzyme, and/or transcription factor from Arabidopsis is expressed in an Arabidopsis plant). In some embodiments, the plant into which the expression cassette is introduced is a different species of plant than the promoter and/or than the polynucleotide encoding GALS or transcription factor (*e.g.*, a vessel-specific promoter, GALS enzyme, and/or transcription factor from Arabidopsis is expressed in a poplar plant). *See, e.g.,* McCarthy *et al.*, *Plant Cell Physiol.* 51:1084-90 (2010); and Zhong *et al.*, *Plant Physiol.* 152:1044-55 (2010).

Methods of Using Plants Having Modified GALS expression

[0110] Plants, parts of plants, or plant biomass material from plants having modified GALS expression can be used for a variety of purposes. In embodiments, in which GALS is overexpressed, the plants, parts of plants, or plant biomass material may be used in a conversion reaction to generate an increased amount of bioenergy as compared to wild-type plants. For example, the plants, parts of plants, or plant biomass material can be used in a saccharification reaction to generate an increased amount of soluble and fermentable sugar compared to wild-type plants. In some embodiments, the plants, parts of plants, or plant biomass material are used to increase biomass yield or simplify downstream processing for wood industries (such as paper, pulping, and construction) as compared to wild-type plants. In some embodiments, the plants, parts of plants, or plant biomass material are used to increase the quality of wood for construction purposes. In some embodiments the plants, or parts of plants are used to improve the quality of textile fiber or simplify the downstream processing for textile industry. In some embodiments the plants, or parts of plants, are used as a raw material for pectin production.

[0111] Methods of conversion, for example biomass gasification, are known in the art. Briefly, in gasification plants or plant biomass material (*e.g.*, leaves and stems) are ground

into small particles and enter the gasifier along with a controlled amount of air or oxygen and steam. The heat and pressure of the reaction break apart the chemical bonds of the biomass, forming syngas, which is subsequently cleaned to remove impurities such as sulfur, mercury, particulates, and trace materials. Syngas can then be converted to products such as ethanol or other biofuels.

[0112] Methods of enzymatic saccharification are also known in the art. Briefly, plants or plant biomass material (*e.g.*, leaves and stems) are optionally pre-treated with hot water, dilute acid, alkali, or ionic liquid followed by enzymatic saccharification using a mixture of cellulases and hemicellulases and pectinases in buffer and incubation of the plants or plant biomass material with the enzymatic mixture. Following incubation, the yield of the saccharification reaction can be readily determined by measuring the amount of reducing sugar released, using a standard method for sugar detection, *e.g.* the dinitrosalicylic acid method well known to those skilled in the art. Plants engineered in accordance with the invention provide a higher sugar yield as compared to wild-type plants.

[0113] In some embodiments, plants in which GALS expression is inhibited that have decreased levels of galactan relative to wild type plants can be used in applications in which it may be desirable to increase the pectin yield and quality. For example in some embodiments, tubers may be genetically modified to inhibit expression of one or more GALS genes, thereby decreasing levels of galactan.

EXAMPLES

[0114] The following examples are provided to illustrate, but not limit the claimed invention.

1. Identification and expression of GALS

[0115] Plant cell walls are predominantly composed of different polysaccharides, which can be grouped into cellulose, hemicelluloses and pectin. Pectin is a class of polysaccharide characterized by a high content of galacturonic acid residues and consists of two major types: homogalacturonan entirely composed of alpha-1,4-linked galacturonosyl residues, and rhamnogalacturonan I (RGI) composed of a backbone of alternating rhamnose and galacturonic acid residues with sidechains composed of arabinans and galactans. In example 1, we identified galactan synthase enzymes in palnts. In this example, we investigated enzymes in glycosyltransferases family 92 (GT92), which has three members in *Arabidopsis*. Loss-of-function mutants in the corresponding genes showed a decreased content of beta-1,4-

galactan. The plants did not have an obvious growth phenotype but pectin was more easily extracted from the cell walls of the mutants, and saccharification was improved. The GT92 enzymes were shown to be ubiquitously expressed and located in the Golgi apparatus. Heterologous expression of one of the proteins showed a high activity in transferring galactose residues from UDP-Gal onto beta-1,4-galactopentaose, confirming the identity of the GT92 enzyme as beta-1,4-galactan synthase.

Results

Glycosyltransferase family GT92 contains beta-1,4-galactosyltransferases

[0116] To identify candidate enzymes for beta-1,4-galactan synthase, we explored the CAZy database of carbohydrate active enzymes (www.cazy.org) (Cantarel et al., 2009). Glycosyltransferases in CAZy are divided into 91 families, 42 of which are represented in angiosperms. Many glycosyltransfases in a range of families have already been investigated in loss-of-function mutants, in many cases without clear indication of the role of the specific glycosyltransferases under study, and previous studies have not led to identification of good candidates for beta-1,4-galactan synthase. GT family 92 was added to CAZy recently with the identification of beta-1,4-galactosyltransferases in animals. The GT92 proteins are reported from pigeon, but not from chicken and not from mammals, where they catalyze the transfer of beta-1,4-linked galactose unto beta-1,4-linked galactose in N-glycan structures. In C. elegans, a member of GT92 has been shown to be a beta-1,4-galactosyltransferase that adds galactose onto core fucose in N-linked glycans. All plants that have had their genomes sequenced have members of GT92, but beta-1,4-galactose is not known from plant glycoproteins. Furthermore, GT92 genes have been identified in transcriptomic studies of tension wood, which is known to be rich in galactan. We investigated the role of GT92 proteins in Arabidopsis. In this species there are three members of GT92.

Arabidopsis loss-of-function mutants in GT92 genes are deficient in beta-1,4-galactan

[0117] Mutants with T-DNA insertions in exons were identified for all three genes and required from the Arabidopsis Biological Resource Center (abrc.osu.edu) and the European Arabidopsis Stock Centre (arabidopsis.info). Homozygous mutants were identified by PCR. Based on the results shown below, the enzymes were designated Galactan Synthase 1 (GALS1, At2g33570), GALS2 (At5g44670) and GALS3 (At4g20170).

[0118] None of the mutants showed any obvious growth or developmental alteration. Cell walls were prepared from leaves and stems and analyzed for sugar composition. Mutants in

any of the three genes resulted in a decrease in total cell wall galactose content whereas no other sugar was changed (Fig. 2). Galactose is a component of several different cell wall components besides galactan. However, dot blots using the LM5 antibody, which is specific for beta-1,4-galactan, showed an effect in the mutants (Fig. 4), whereas an antibody recognizing arabinogalactan proteins did not show any difference in binding. The specific effect on beta-1,4-galactan was further shown by immunofluorescence microscopy of petioles, which had significantly less LM5 labeling in the mutants (Fig. 4). These data indicated that the GT92 enzymes were specifically involved in biosynthesis of beta-1,4-galactan.

GALS1, GALS2 and GALS3 are Golgi localized enzymes with overlapping expression pattern

[0119] The GALS proteins are predicted to be Type II membrane proteins targeted to the secretory pathway. The subcellular localization was investigated by heterologous expression of YFP fusion proteins in *Nicotiana benthamiana* and confocal laser scanning microscopy. This analysis showed that GALS proteins were targeted to Golgi vesicles, consistent with a role in pectin biosynthesis.

[0120] To investigate the expression of GALS in more detail we expressed the beta-glucuronidase gene under control of the native promoters of the three genes. The data confirmed that all three genes are expressed in several different tissues, but also showed some important differences in expression pattern.

GALS1 is a beta-1,4-galactosyltransferase

[0121] Because the animal GT92 enzymes include beta-1,4-glactosyltransferases and the Arabidopsis mutants had specific deficiency in beta-1,4-galactan, the Arabidopsis GT92 enzymes were strong candidates for beta-1,4-galactan synthase. To further investigate this, the GALS1 protein was heterologously expressed in *N. benthamiana* as fusion protein with N-terminal FLAG or YFP tags. Galactan synthase assays using microsomal proteins and endogenous acceptors showed a very high activity. The identity of the product as beta-1,4-galactan was confirmed by digestion with beta-1,4-galactanase. To further characterize the galactosyltransferase activity we chemically synthesized beta-1,4-galactopentaose. The purity and identity of the beta-1,4-galactopentaose were confirmed by NMR (data not shown). After solubilization of microsomal membranes with Triton X-100 and incubation in the presence of the acceptor, the unincorporated UDP-Galactose was removed by anion exchange chromatography. The results showed a highly significant activity of transfer of galactose onto the acceptor with microsomes from plants expressing GALS1 (Fig. 5). The

galactan synthase activity was high, but not very stable as seen from the comparison of 20 min and 1 hr incubation times.

[0122] While the data using microsomal protein strongly indicated that GALS1 was indeed beta-1,4-galactan synthase, it must always be considered that microsomes contain many proteins, including endogenous galactan synthase, and that the expression of a heterologous protein could alter the background of endogenous enzyme activities. This is not normally a problem for this type of assay, but must nevertheless be considered. We therefore affinity purified the FLAG-GALS1 fusion protein. The purifed protein retained acceptor dependent activity, whereas the control reactions (mock purified protein from p19 expressing microsomes) had no detectable activity (Fig. 5).

[0123] Evlauation of GALS1 overexpressors demonstrated that the plants have 250% galactose in the cell wall compared to the control (data not shown).

Discussion

[0124] Pectin synthesis requires many different proteins for it synthesis and identifying these proteins has been very challenging. Only one enzyme – GAUT1 – has had its activity unambiguously shown previously (Sterling et al., 2006). The XGD1 protein appears to add xylose to the backbone of homogalaturonan, but the activity was only shown with crude membrane preparations and not with isolated protein (Jensen et al., 2008). Here we provide clear biochemical evidence that glycosyltransferases of family GT92 in plants are beta-1,4-galactan synthases. The GALS1 enzyme showed a high activity with galactooligosaccharide acceptors and is hence capable of elongating beta-1,4-galactans. It is not clear if the same enzymes will add the first galactose residue onto the RGI backbone, but we find that unlikely due to the very different properties of the acceptor polysaccharide. Most likely one or two as yet unidentified galactosyltransferases are required to initiate the beta-1,4-galactan sidechains.

[0125] GT92 proteins are encoded in the genomes of all plants that have been investigated and are found in a limited range of animal species. They are not found in fungi, nor in green algae such as *Chlamydomonas reinhardtii*, but they are reported in apicomplexan of the *Cryptosporidium* genus. In animals the proteins are known to add beta-1,4-galactose to various N-glycans, but this particular decoration is not well conserved and the phylogeny indicates that GT92 members have been lost in many taxonomic groups. In plants, the proteins are also beta-1,4-galactosyltransferases but have evolved specifically to synthesize pectic beta-1,4-galactans. The GT92 proteins contain Domain of Unknown Function 23

(DUF23), which is also found in a few bacteria. However, these bacterial proteins are quite divergent form the eukaryotic proteins and none of them have been characterized.

[0126] The three *Arabidopsis* GALS show overlapping but not identical expression. This explains why biochemical phenotypes could be observed for mutants in all three genes while all the mutants retained significant amounts of residual galactan. Mutants with combinations of two or three gene mutations can be generated to observe larger reductions in galactan. T he mutants did not show any obvious phenotype on growth.

2. Overexpression of galactan synthase and epimerase increases galactan accumulation in plants

[0127] Overexpression of galactan synthase (GALS1 and homologs of this) led to increased galactan accumulation in plants. This example provides illustrative data showing that even higher accumulation can be achieved by coexpressing galactan synthase and an enzyme that produces the substrate of galactan synthase, *i.e.* UDP-glucose 4-epimerase (UGE).

[0128] Plants have different isoforms of UGE, which may differ in substrate specificity (some also convert UDP-xylose) and in specificity for certain polysaccharide products. For example, the *Arabidopsis thaliana* genome encodes five isoforms of UGE. These belong to a larger family of epimerases that also include UDP-glucuronic acid epimerases and UDP-xylose epimerases. The sequence of *Arabidopsis* UGE2 (see, e.g., Kotake *et al., Biochem. J.* 2009, *supra*) is provided as SEQ ID NO:45. Similar, although not necessarily identical, results with respect to further increasing galactan accumulation can be expected with different isoforms of UGE, e.g. UGE1, UGE2, UGE3, UGE4 and UGE5 from *Arabidopsis thaliana*. UGE2 was chosen for this example because it has low activity with UDP-xylose and because it has previously been implicated in pectin biosynthesis (whereas, *e.g.* UGE4 which has been implicated in arabinogalactan and xyloglucan biosynthesis).

[0129] Since an optimal biofuel feedstock would contain less xylose (pentose) and less lignin (recalcitrant), it would be desirable to generate a plant where hexoses, such as galactan, replace the xylans. Plants with a decrease in xylan have been engineered by expressing xylan GTs under control of vessel-specific promoters in xylan deficient mutants (*see, e.g.*, Petersen *et al.*, 2012). Similary, plants with reduced lignin levels have been engineered by replacing the promoter of a key lignin gene (C4H) with a vessel-specific promoter (Yang *et al.*, 2012). These low-lignin plants additionally contain an artificial positive feedback loop, with the NST1 transcription factor under control of the fiber-specific pIRX8 promoter, which it induces. This positive feedback loop results in plants with enhanced polysaccharide

deposition in stems. An approach to add galactan to fibers cells of plants was therefore employed in this example where AtUGE2 was expressed in combination with the mentioned positive feedback loop, under control of the fiber-specific pIRX5 promoter.

[0130] Thus, two different strategies were employed to increase galactan (and thereby hexose) levels beyond what could be achieved by overexpression of GALS1 alone: Simultaneous overexpression of a UGE with GalS1, and tissue-specific expression of a UGE combined with a positive feedback loop for increased polysaccharide deposition.

Cloning strategy – construction of transformation vectors

[0131] *Arabidopsis thaliana* UGE2 was overexpressed in *A. thaliana*. In order to track the expression of the transgenic protein, constructs were designed to contain an N-terminal FLAG-tag or an N-terminal GFP-tag as well as the N-terminal FLAG-tag. The FLAG-tag was introduced by a PCR reaction using overlapping primers, while the GFP tag is part of the vector pMDC43. Both vectors pMDC32 and pMDC43 contain two consecutive copies of the Cauliflower Mosaic Virus 35S promoter. An empty vector control was generated by inserting a non-coding 15bp sequence in the pMDC32 vector. Finally, a construct for fiber-specific expression of AtUGE2 was created with the pIRX5 promoter. IRX5 is a component of the cellulose synthase complex, and thus expression under its promoter leads to secondary cell wall expression (Taylor, 2003). The construct was designed so that after the promoter is the coding sequence of the transcription factor NAC Secondary Wall Thickening Promoting Factor 1 (NST1), followed by the autoproteolytic peptide 2A (Halpin et al., 1999), and then AtUGE2. NST1 is an important factor for activation of secondary wall biosynthesis including the deposition of cellulose in the interfascicular fibers, and affects several cellulose synthesis related genes, including inducing pIRX5. Therefore, by expressing this construct, a positive feedback loop is created, where expression of NST1 leads to an increased expression of the secondary cell wall synthesizing genes including pIR5X, which in turn leads to higher NST1 expression and as a result increased cell wall deposition.

[0132] All of the 35S constructs were transformed into Col-0 plants as controls. The 35S promoter FLAG-tagged AtUGE2 was also transformed in GalS1-overexpressing transformants in a Col-0 background. Likewise, the IRX5 promoter construct was transformed into GalS1 overexpressors. The GalS1 overexpressing lines containing BASTA resistance and YFP tagged GalS1 are described in Liwanag et al. (2012).

Monosaccharide composition

[0133] The monosaccharide composition of the cell wall was determined by high-performance anion exchange chromatography (HPAEC) as described (Harholt *et al.*, 2006). Destarched, TFA-hydrolyzed, Alcohol Insoluble (AIR) samples were run on an Ion Chromatography System 3000 (Dionex) using a CarboPac PA20 anion exchange column (Dionex) and gold electrode to determine the monosaccharide composition. Sugar standard solutions of 5, 10, 25, 50, 100, 150 and 200 µM of glucose, fucose rhamnose, arabinose, galactose, xylose, galacturonic acid and glucuronic acid (Sigma-Aldrich) were run as references. A NaOH gradient program from 10 mM to 45mM NaOH was used to elute the sugars.

Results

[0134] Transformants were selected on plates with the appropriate antibiotic and transferred to soil. Among several recovered transformants, plants with high expression of the transgene were identified by RT-PCR. The selected plants were grown to maturity and the seeds harvested for analysis of cell wall composition in the subsequent plant generation.

[0135] The cell walls of control plants expressing only UGE2 under the 35S promoter or the pIRX:NST1-2A-UGE2 construct were analyzed and compared to wild-type plants and GALS1 overexpressor plants. Only GALS1 overexpressing plants showed an increase in cell wall galactose (Fig. 6). Thus, under these conditions, expression of UGE2 or NST1-2A-UGE2 alone did not result in increased galactan deposition in leaves.

[0136] When leaves were analyzed from plants that coexpressed GALS1 and UGE2, no increase was observed above that achieved with GALS1 alone (Fig. 7). However, leaves from plants coexpressing GALS1 and the NST1-2A-UGE2 construct had significantly higher galactose content in the cell walls over what was observed with GALS1 alone.

[0137] The pIRX5 promoter is not very active in leaves, and we therefore investigated cell walls of stems, which is also more relevant for biofuel and biorefinery applications. The results (Fig. 8) showed that coexpression of both UGE2 and GALS1 and of NST1-2A-UGE2 and GALS1 lead to increased cell galactose comaprred to expression of GALS1 alone. The plants responded by incorporating relatively less xylose in the cell walls. To better see the change in cell wall monosaccharides besides xylose, the data was also shown without xylose (Fig. 9). Clearly, the coexpression increased the galactose content substantially over that achieved with GALS1 overexpression alone. GALS1 overexpression alone with the 35S

promoter does not lead to significant increase in galactose in stem cell walls. The figures also show that overexpression of the pIRX5:NST1-2A-UGE2 construct alone did not lead to increased galactose in stems.

[0138] The GALS1 gene in these experiments was expressed with the 35S promoter. For a higher expression in fiber cells and improved galactan accumulation in this cell type, the GALS1 gene or a homologous GALS gene is expressed under a promoter with high activity in fiber cells, e.g. pIRX8, pIRX5, or other as described in Loque D, Scheller HV (2012) Spatially modified gene expression in plants. PCT/US2012/023182. The GALS gene is expressed under a separate promoter distinct from the NST1-2A-UGE2 sequence or in one construct such as pIRX5:GALS1-2A-NST1-2A-UGE2 where all three open reading frames are transcribed from the same promoter.

Citations for References Cited by Author, Date

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res 37: D233-238

Harholt J, Jensen JK, Sorensen SO, Orfila C, Pauly M, Scheller HV (2006) ARABINAN DEFICIENT 1 is a putative arabinosyltransferase involved in biosynthesis of pectic arabinan in Arabidopsis. Plant Physiol 140: 49-58

Harholt J, Jensen JK, Verhertbruggen Y, Sogaard C, Bernard S, Nafisi M, Poulsen CP, Geshi N, Sakuragi Y, Driouich A, Knox JP, Scheller HV (2012) ARAD proteins associated with pectic Arabinan biosynthesis form complexes when transiently overexpressed in planta. Planta

Harholt J, Suttangkakul A, Vibe Scheller H (2010) Biosynthesis of pectin. Plant Physiol 153: 384-395

Jensen JK, Sorensen SO, Harholt J, Geshi N, Sakuragi Y, Moller I, Zandleven J, Bernal AJ, Jensen NB, Sorensen C, Pauly M, Beldman G, Willats WG, Scheller HV (2008) Identification of a xylogalacturonan xylosyltransferase involved in pectin biosynthesis in Arabidopsis. Plant Cell 20: 1289-1302

Mohnen D (2008) Pectin structure and biosynthesis. Curr Opin Plant Biol 11: 266-277

Sterling JD, Atmodjo MA, Inwood SE, Kumar Kolli VS, Quigley HF, Hahn MG, Mohnen D (2006) Functional identification of an Arabidopsis pectin biosynthetic homogalacturonan galacturonosyltransferase. Proc Natl Acad Sci USA 103: 5236-5241

Kotake, T., Takata, R., Verma, R., Takaba, M., Yamaguchi, D., Orita, T., Kaneko, S., Matsuoka, K., Koyama, T., Reiter, W.D., Tsumuraya, Y. (2009). Bifunctional cytosolic UDP-glucose 4-epimerases catalyse the interconversion between UDP-D-xylose and UDP-L-arabinose in plants. Biochemical Journal, 424, 169-177.

Petersen, P. D., Lau, J., Ebert, B., Yang, F., Verhertbruggen, Y., Kim, J. S., Varanasi, P., Suttangkakul, A., Auer, M., Loque, D., Scheller, H. V. (2012). Engineering of plants with improved properties as biofuels feedstocks by vessel-specific complementation of xylan biosynthesis mutants. Biotechnology for Biofuels, 5(84).

Taylor, N. G. (2008). Cellulose biosynthesis and deposition in higher plants. New Phytologist, 178(2), 239-252.

Yang, F., Mitra, P., Zhang, L., Prak, L., Verhertbruggen, Y., Kim, J., Sun, L., Zheng, K., Tang, K., Auer, M, Scheller, H.V., Loqué, D. (2012). Engineering secondary cell wall deposition in plants. Plant Biotechnology Journal.

Halpin, C. et al. (1999) Self-processing 2A-polyproteins – a system for co-ordinate expression of multiple proteins in transgenic plants. The Plant Journal 17, 453- 459

[0139] It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All publications, patents, accession numbers, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

ILLUSTRATIVE SEQUENCES

SEQ ID NO:1 GALS1 (ARABIDOPSIS THALIANA_ATG2G33570)

ACCESSION NUMBER: NP_565768.1 GENE ID: 817922

MRKEVLPPVLSTTTVCFEKKPIIATLLALSLVMIVWNLPPYYHNLISTARPCSAVTTTTTTTTLLSSSNFTS
AENFTTSLSTTTAAASQKYDSTPSDPNKRVFQPFGNAAALFVLGLYCEAYRGGPTTFSVIGLASKPIH
VYGKPWYKCEWISNNGTSIRAKAQKILPDWGYGRVYTVVVVNCTFNSNPNSDNTGGKLILNAYYNES
PKLFERFTTLEESAGIYDESKYSPPYQYDYLYCGSSLYGNVSASRMREWMAYHAWFFGDKSHFVFH
DAGGVSPEVRKVLEPWIRAGRVTVQNIRDQSQYDGYYYNQFLIVNDCLHRYRYAANWTFFFDVDEYI
YLPHGNTLESVLDEFSVNTQFTIEQNPMSSVLCINDSSQDYPRQWGFEKLLFKDSRTKIRRDRKYAIQ
AKNAFATGVHMSENIVGKTLHKTETKIRYYHYHNTITVHEELCREMLPNSAKKKVTLYNKLPYVYDDN
MKKLVKTIKEFEQKKLGTDVKNFS*

SEQ ID NO:2 GALS2 (ARABIDOPSIS THALIANA_AT5G44670)

ACCESSION NO: NP_199280.1 GENE ID: 834496

MAKERDQNTKDKNLLICFLWNFSAELKLALMALLVLCTLATLLPFLPSSFSISASELRFCISRIAVNSTSV
NFTTVVEKPVLDNAVKLTEKPVLDNGVTKQPLTEEKVLNNGVIKRTFTGYGWAAYNFVLMNAYRGGV
NTFAVIGLSSKPLHVYSHPTYRCEWIPLNQSDNRILTDGTKILTDWGYGRVYTTVVVNCTFPSNTVINP
KNTGGTLLLHATTGDTDRNITDSIPVLTETPNTVDFALYESNLRRREKYDYLYCGSSLYGNLSPQRIRE
WIAYHVRFFGERSHFVLHDAGGITEEVFEVLKPWIELGRVTVHDIREQERFDGYYHNQFMVVNDCLH
RYRFMAKWMFFFDVDEFIYVPAKSSISSVMVSLEEYSQFTIEQMPMSSQLCYDGDGPARTYRKWGF
EKLAYRDVKKVPRRDRKYAVQPRNVFATGVHMSQHLQGKTYHRAEGKIRYFHYHGSISQRREPCRH
LYNGTRIVHENNPYV LDTTMRDIGLAVKTFEIRTIGDRLLRTRQ*

SEQ ID NO:3 GALS3 (ARABIDOPSIS THALIANA_AT4G20170)

PROTEIN ACCESSION NO: NP_193750.1 GENE ID: 827763

MAMVKEKEQNTKDKKLLVGVIWNFSAELKLTFMALLVLCTLATLLPFIPSSFSLSTSDFRFCISRFSSAV
PLNTTTTVEESSSSPSPEKNLDRVLDNGVIKRTFTGYGSAAYNFVSMSAYRGGVNSFAVIGLSSKPLH
VYGHPSYRCEWVSLDPTQDPISTTGFKILTDWGYGRIYTTVVVNCTFSSISAVNPQNSGGTLILHATT
GDPTLNLTDSISVLTEPPKSVDFDLYNSTKKTKKYDYLYCGSSLYGNLSPQRVREWIAYHVRFFGERS
HFVLHDAGGIHEEVFEVLKPWIELGRVTLHDIRDQERFDGYYHNQFMIVNDCLHRYRFMTKWMFFFD
VDEFLHVPVKETISSVMESLEEYSQFTIEQMPMSSRICYSGDGPARTYRKWGIEKLAYRDVKKVPRR
DRKYAVQPENVFATGVHMSQNLQGKTYHKAESKIRYFHYHGSISQRREPCRQLFNDSRVVFENTPY
VLDTTICDVGLAVRTF ELRTIGDRLLRTRQ*

SEQ ID NO: 42 GALS1 (ARABIDOPSIS THALIANA AT2G33570)

ACCESSION NUMBER: NM_128917' 1838 BP MRNA

1 CTAATTTCTA CACGCCGTGT CGGCAAAGCC TCTCGTCACT TCTCTCTGAC GCTTGTCGTC
61 ACTTTTGAAT TTTTTTAATT TTTAAATAAT TGATAACCGA AACGGTGCGT TTTACTCACC
121 GTCGTCGGGA AAAAAAACAT GAGGAAGGAA GTTTTGCCGC CGGTGTTATC AACCACCACA

```
181  GTATGTTTCG AGAAGAAACC AATAATTGCT ACATTACTAG CTCTCTCTCT CGTCATGATT
241  GTCTGGAACC TTCCTCCTTA CTACCACAAC CTCATCTCCA CCGCTCGTCC CTGCTCCGCC
301  GTCACCACCA CCACCACCAC CACCTTACTC TCCTCATCGA ACTTCACTTC GGCGGAGAAT
361  TTCACCACCT CTCTCTCAAC GACAACTGCA GCAGCTTCTC AGAAGTACGA TTCAACTCCC
421  TCAGATCCGA ACAAACGCGT TTTCCAACCG TTCGGAAACG CGGCGGCGTT ATTCGTACTA
481  ATGGGAGCTT ACCGCGGCGG TCCAACGACG TTTTCCGTTA TCGGACTCGC GTCGAAACCG
541  ATCCACGTTT ACGGAAAACC ATGGTACAAG TGTGAGTGGA TATCTAACAA TGGAACTTCG
601  ATTCGAGCTA AAGCACAGAA GATTCTACCA GATTGGGGAT ACGGACGAGT CTACACCGTC
661  GTCGTCGTCA ATTGCACTTT CAATTCAAAC CCTAACTCCG ATAACACCGG AGGTAAACTC
721  ATTCTCAACG CTTACTACAA CGAATCTCCC AAACTCTTTG AACGATTCAC TACGTTAGAA
781  GAATCAGCTG GAATCTACGA CGAATCGAAA TACTCGCCGC CGTATCAGTA CGATTACCTC
841  TATTGTGGCT CGTCACTGTA CGGTAACGTG AGCGCGTCGC GTATGAGAGA GTGGATGGCT
901  TACCACGCTT GGTTCTTTGG TGACAAATCG CATTTTGTTT TCCACGATGC TGGTGGTGTG
961  TCGCCGGAAG TTAGGAAGGT TCTTGAGCCG TGGATTCGAG CTGGGAGGGT CACGGTTCAG
1021 AATATTCGGG ATCAGTCGCA GTATGATGGT TACTACTATA ATCAGTTTCT TATTGTTAAT
1081 GATTGCTTGC ATCGGTATCG ATACGCTGCG AATTGGACCT TCTTCTTCGA TGTCGATGAG
1141 TATATCTATT TGCCGCATGG TAATACACTC GAATCCGTGC TCGATGAGTT CTCGGTTAAC
1201 ACGCAGTTTA CGATTGAGCA GAATCCAATG TCTAGTGTTC TTTGCATAAA CGACTCTTCT
1261 CAAGATTATC CAAGGCAATG GGGATTTGAG AAATTGTTAT TTAAGGATTC AAGGACGAAG
1321 ATACGACGTG ATAGAAAATA TGCAATCCAA GCGAAGAACG CATTTGCTAC AGGAGTTCAT
1381 ATGTCTGAAA ACATTGTAGG CAAAACACTA CACAAGACAG AGACAAAGAT TCGTTATTAC
1441 CATTACCACA ACACCATAAC TGTGCATGAG GAGCTTTGTA GAGAGATGTT ACCAAATTCA
1501 GCCAAGAAGA AGGTGACATT GTACAATAAG CTTCCGTATG TGTATGATGA CAACATGAAG
1561 AAGCTAGTGA AGACGATTAA AGAGTTTGAG CAGAAAAAAC TTGGGACGGA TGTGAAGAAT
1621 TTCTCATGAC CATAATATAG CTGTAATCTC TCTGATAAGC ATTTTGTCTA TAAAGGTATA
1681 GTTGTTTCTA CATTACATGT ATCATTTTTC ATTCTGTTTT GTCCTCTTTT ACTATTTCAT
1741 TAATGACTTT GATCAATATT TTTGAAAATT ACTTGTGTTT TCTTTTGTTA TGTATTGAAC
1801 TTAATAGAAA TTAGAGTTAC TCAAGACCTT GGACATAC //
```

SEQ ID NO:43 GALS2 (ARABIDOPSIS THALIANA AT5G44670)

ACCESSION NUMBER: NM_123834; 1983 BP MRNA

```
1    AAATTCTCCA AATTTTTTCT TCTCTCTCTC TTCCCTCTGT CGCTTCACTT TCCCTCTGCT
61   TCATAGTTCA TACGATTCTT CGATTCGTCT TCTTCAATCA GTGAAGAAGA ACTCAAAAGA
121  TGGCTAAAGA GAGAGACCAA AACACTAAAG ACAAAAACCT CCTCATCTGT TTCTTATGGA
181  ACTTCTCCGC CGAGCTTAAG CTAGCTCTAA TGGCGTTACT CGTTCTCTGC ACTTTAGCTA
241  CTCTCCTCCC TTTTCTACCT TCTTCTTTCT CCATCTCCGC TTCCGAACTC CGTTTCTGCA
301  TCTCACGCAT CGCCGTAAAC TCCACCTCCG TCAACTTCAC CACCGTCGTC GAAAAGCCAG
361  TCTTAGATAA CGCTGTCAAG TTAACTGAGA AGCCGGTGTT GGATAACGGC GTTACGAAAC
421  AGCCGTTAAC TGAAGAGAAG GTGTTAAATA ACGGCGTTAT TAAACGGACG TTTACTGGTT
481  ACGGCTGGGC AGCTTATAAC TTCGTGTTAA TGAACGCTTA CAGAGGCGGC GTTAACACAT
541  TCGCCGTTAT CGGTTTATCA TCTAAACCAC TTCACGTTTA CTCTCATCCC ACTTACCGTT
```

601 GCGAATGGAT TCCACTAAAC CAATCCGATA ACCGGATTTT AACCGACGGT ACCAAAATCT
661 TAACCGATTG GGGTTACGGT AGAGTTTACA CAACCGTCGT CGTAAACTGT ACTTTTCCGT
721 CAAACACCGT GATAAACCCT AAAAACACCG GAGGTACTCT TCTCCTCCAC GCAACCACCG
781 GAGATACAGA CCGGAACATC ACCGATTCAA TTCCGGTACT CACCGAAACT CCAAACACCG
841 TCGATTTTGC TCTCTACGAA TCCAATCTCC GCCGGCGAGA GAAGTACGAT TATCTCTATT
901 GTGGATCTTC TCTCTACGGC AACTTATCAC CACAGAGAAT CAGAGAATGG ATCGCTTACC
961 ATGTAAGGTT CTTCGGTGAA AGATCTCATT TTGTTCTACA TGACGCCGGA GGGATTACAG
1021 AGGAAGTGTT TGAGGTTTTA AAGCCATGGA TAGAGCTTGG GAGAGTTACT GTTCATGATA
1081 TTAGAGAACA AGAGAGATTT GATGGTTATT ATCATAATCA ATTCATGGTG GTGAATGATT
1141 GTTTGCATAG GTATAGATTC ATGGCGAAGT GGATGTTTTT CTTCGATGTT GATGAGTTTA
1201 TTTATGTTCC GGCGAAGAGT TCGATTTCGT CGGTGATGGT ATCTTTGGAG GAATATTCTC
1261 AGTTTACTAT TGAACAGATG CCTATGAGTA GTCAGCTTTG TTACGACGGT GATGGTCCGG
1321 CGAGGACTTA CAGGAAATGG GGATTTGAGA AATTGGCGTA TAGAGATGTG AAGAAAGTAC
1381 CACGACGGGA TAGGAAGTAT GCGGTTCAAC CGCGGAACGT ATTTGCGACA GGGGTTCACA
1441 TGTCTCAGCA TCTACAAGGG AAGACGTATC ACAGAGCGGA AGGGAAAATA CGCTATTTTC
1501 ACTACCATGG TTCAATCTCG CAGCGTCGTG AGCCTTGTCG TCATCTTTAT AACGGTACCC
1561 GTATCGTTCA TGAGAACAAT CCTTACGTGC TTGATACCAC AATGCGTGAT ATTGGTCTCG
1621 CGGTGAAGAC GTTTGAGATT AGGACGATTG GAGATCGCTT GCTTAGGACG AGACAATGAA
1681 GGCAGGAGAA GAATGGTTAA AGACATGTTA TCATCATTAT GCGTTGTAAC GTAAATCTTT
1741 TAGAGTATTA TTTAGGCGAA TGTAACAATT TTCATGGTTT TTTGTTTAGT ATATTCTTTT
1801 ATTGTATTAT AAAATGGGTT CGTACATAGA GATCATCATA CAGCTCAGAT TCTTGGTATA
1861 TAAGCATCTT TTTTATGGGC TTTATAATTT TTTCCGTTAT TTATGGAAAA GTGCTTTATA
1921 TAAATTAGTG AAAGTTGTTG TGGTCTTCCA TGGATCTTTG TCGTGTTAAT TAAAAGTTTC
1981 CAC //

SEQ ID NO:44 GALS3 (ARABIDOPSIS THALIANA AT4G20170)

ACCESSION NUMBER: NM_118136 1993 BP MRNA

1 AAAAGTGAGA GACACACAAC TTCGGAGCGA ATCTATTCTT CTTCTTCTTC TTCTTCTTCT
61 TCTTCTTCTT CCTCCGTTTT TTTCATCTTC TTCTCTGTTT CGAGAGATCC ACTAGTGAAA
121 GAGTCAGCAC CATGGCCATG GTCAAAGAGA AAGAACAAAA CACTAAAGAC AAAAAACTCC
181 TCGTCGGCGT CATTTGGAAC TTCTCCGCCG AGCTCAAGCT CACTTTCATG GCGTTACTTG
241 TTCTCTGCAC TTTAGCTACT CTCTTACCTT TCATACCTTC TTCATTCTCT CTCTCCACTT
301 CCGATTTCCG CTTCTGCATC TCACGCTTCT CCTCCGCCGT CCCTCTCAAC ACCACCACCA
361 CCGTAGAAGA ATCATCATCC TCACCGTCAC CGGAGAAGAA CCTAGATCGA GTTTTGGATA
421 ACGGAGTTAT TAAACGGACG TTTACTGGCT ACGGCTCAGC AGCTTATAAC TTCGTCTCAA
481 TGAGTGCTTA CAGAGGCGGC GTTAACTCAT TCGCCGTTAT CGGATTATCA TCAAAACCAT
541 TACACGTGTA CGGTCATCCT TCGTATAGAT GCGAATGGGT CTCATTAGAC CCGACTCAAG
601 ATCCGATTTC AACAACCGGG TTTAAAATCT TAACCGATTG GGGTTACGGA CGGATCTACA
661 CAACAGTCGT CGTTAACTGT ACTTTCTCAT CAATCTCCGC CGTGAATCCA CAAAACTCCG
721 GTGGAACTCT CATCCTCCAC GCCACCACCG GAGATCCAAC TCTCAATCTC ACCGATTCAA
781 TCTCAGTCCT AACCGAACCT CCCAAATCCG TCGATTTCGA TCTCTATAAC TCCACGAAGA

```
841  AGACGAAGAA GTACGATTAT CTCTATTGCG GATCGTCCTT ATACGGTAAC CTAAGTCCGC
901  AACGAGTTAG AGAATGGATC GCTTACCACG TTAGATTCTT CGGTGAACGG TCACATTTCG
961  TGCTACACGA CGCCGGAGGG ATTCATGAGG AAGTGTTCGA GGTTTTAAAG CCATGGATTG
1021 AGCTAGGGAG AGTGACGTTA CATGATATTA GAGATCAAGA ACGATTCGAT GGATATTATC
1081 ATAATCAGTT CATGATAGTG AATGATTGTT TGCATAGGTA TAGATTCATG ACGAAGTGGA
1141 TGTTCTTCTT TGATGTTGAT GAGTTTTTAC ATGTTCCAGT GAAAGAGACG ATTTCGTCTG
1201 TGATGGAATC TTTGGAGGAA TATTCTCAGT TTACTATTGA ACAGATGCCT ATGAGTAGTC
1261 GGATTTGTTA TTCCGGTGAT GGTCCGGCGA GAACTTACAG GAAATGGGGA ATTGAGAAAC
1321 TGGCATATAG AGACGTCAAG AAGGTTCCAA GACGGGATCG AAAATACGCT GTCCAGCCGG
1381 AGAATGTATT CGCGACAGGC GTACACATGT CTCAGAATCT ACAAGGGAAA ACATACCACA
1441 AGGCTGAAAG CAAAATCCGT TACTTCCACT ACCATGGTTC GATCTCTCAG CGCCGCGAGC
1501 CTTGTCGTCA ACTTTTTAAC GATTCTCGAG TCGTGTTCGA GAACACTCCT TATGTGCTAG
1561 ACACTACAAT ATGTGATGTT GGCCTTGCTG TGAGAACGTT CGAGTTGAGA ACGATCGGTG
1621 ATCGGCTGCT ACGGACAAGA CAATGAAGAG ATGGCAAAAA TGAATAGTGA ATGTAATCAA
1681 TCTTTAGAAA GAAGAATTAG AAGGTGTTAA GATGAGTTAC TTTGTATTAT TTTCTTTTGG
1741 GGGTATATTC TTTTATTGTA TCATATAATT TGGGTAATGG GTTCATTAAT ACAGCTTGAA
1801 AATACTCTTT GGTATATATA TTCTGTATGA TGTATGATTT AGAAAAAAGG TCTCTGAGTA
1861 TATAATCTAG TGATGATAAT TGTGGAGATC AAGTAATATC ACTGTTTGTA TTTGATTACT
1921 GTACTCTTAG TTGACAAAAA GAAAATGTCA ATATCCATTG GTGTTACTCC AGTAATCCAT
1981 ATGGAACGTT GAT //
```

SEQ ID NO:45 UDP-D-GLUCOSE/UDP-D-GALACTOSE 4-EPIMERASE 2 (ARABIDOPSIS THALIANA)

GI|332659427|GB|AEE84827.1

```
MAKSVLVTGGAGYIGSHTVLQLLEGGYSAVVVDNYDNSSAASLQRVKKLAGENGNRLSFHQVDLRDRPAL
EKIFSETKFDAVIHFAGLKAVGESVEKPLLYYNNNIVGTVTLLEVMAQYGCKNLVFSSSATVYGWPKEVP
CTEESPISATNPYGRTKLFIEEICRDVHRSDSEWKIILLRYFNPVGAHPSGYIGEDPLGVPNNLMPYVQQ
VAVGRRPHLTVFGTDYKTKDGTGVRDYIHVMDLADGHIAALRKLDDLKISCEVYNLGTGNGTSVLEMVAA
FEKASGKKIPLVMAGRRPGDAEVVYASTEKAERELNWKAKNGIEEMCRDLWNWASNNPYGYNSSSNGSSS
```

WHAT IS CLAIMED IS:

1. A method of engineering a plant to increase the galactan content in a desired tissue, the method comprising:

introducing an expression cassette into the plant, wherein the expression cassette comprises a polynucleotide encoding a β-1,4-galactan synthase (GALS) operably linked to a heterologous promoter, wherein the GALS has at least 70% identity to a sequence selected from SEQ ID NOS:1-27, and

culturing the plant under conditions in which the transcription factor is expressed.

2. The method of claim 1, wherein the plant is genetically modified to overexpress a UDP-galactose epimerase in the same plant tissue in which GALS is overexpressed.

3. The method of claim 1, wherein the plant is genetically modified to overexpress a transcription factor that induces expression through the heterologous promoter.

4. The method of claim 3, wherein expression of the transcription factor is driven by a promoter that is activated by the transcription factor.

5. The method of claim 1, wherein the plant is genetically modified to overexpress a UDP-galactose epimerase in the same plant tissue in which GALS is overexpressed and is genetically modified to overexpress a transcription factor that induces expression through the heterologous promoter.

6. The method of claim 5, wherein expression of the GALS, UDP-galactose epimerase and the transcription factor is driven by a promoter that is activated by the transcription factor.

7. The method of claim 1, wherein promoter is an IRX1, IRX3, IRX5, IRX8, IRX9, IRX14, IRX7, IRX10, GAUT13, GAUT14, or CESA4 promoter.

8. The method of claim 7, wherein the promoter is an IRX5 or IRX8 promoter.

9. The method of any one of claims 1 to 8, wherein the transcription factor is NST1.

10. The method of 1, wherein wherein the GALS has at least 70% identity to SEQ ID NO:1, SEQ ID NO:2, or SEQ ID NO:3.

11. The method of claim 10, wherein the GALS has at least 70% identity to SEQ ID NO:1.

12. The method of any one of claims 1 to 11, wherein the plant is Arabidopsis, poplar, eucalyptus, rice, corn, cotton, switchgrass, sorghum, millet, miscanthus, sugarcane, pine, alfalfa, wheat, soy, barley, turfgrass, tobacco, hemp, bamboo, rape, sunflower, willow, or Brachypodium.

13. A plant engineered by the method of any of claims 1 to 12, or a progeny of the plant.

14. A plant cell from the plant of claim 13.

15. Seed from the plant of claim 13.

16. A plant cell comprising a polynucleotide encoding a GALS operably linked to a heterologous promoter.

17. Biomass comprising plant tissue from the plant of claim 13.

18. A method of obtaining an increased amount of soluble sugars from a plant in a saccharification reaction, the method comprising:

subjecting the plant of claim 13 to a saccharification reaction, thereby increasing the amount of soluble sugars that can be obtained from the plant as compared to a wild-type plant.

19. A method of engineering a plant having decreased galactan content, the method comprising:

introducing a polynucleotide that inhibits expression of a gene encoding GALS into the plant, wherein the polynucleotide is operably linked to a heterologous promoter; and

culturing the plant under conditions in which the polynucleotide is expressed.

20. The method of claim 12, wherein the plant is a tuber.

21. A plant comprising a polynucleotide that inhibits expression of a gene encoding GALS into the plant, wherein the polynucleotide is operably linked to a heterologous promoter.

22. A plant comprising a heterologous polynucleotide encoding a GALS protein.

```
SEQ ID NO:
Sbi_Sb04g030080.1|Sb04g030080.                  ----------MKYNTR----------KDA-GGGGGAPFAVP-  4
Zma_GRMZM2G164912_T01|GRMZM2G1                  ----------MKYNTR----------KDASGGGGGGAPFAIP-  5
Sit_Si017001m|Si017001m|Si0170                  ----------MKYNAR----------KDAGGGGGGGAPFAIP-  6
Osa_LOC_Os02g48190.1|LOC_Os02g                  ----------MKNGAR----------KDAGGGGGGG--VFVP-  7
Bdi_Bradi3g53170.1|Bradi3g5317                  ----------MKQNAR----------KDPGGGGGGG-AFVIA-  8
Sbi_Sb10g013010.1|Sb10g013010.                  ----------MKPDAR----------KDAG--GGVAALG---  9
Zma_GRMZM2G136800_T02|GRMZM2G1                  ----------MKPDVR----------KDAG--CGVTAFG---  10
Sit_Si006346m|Si006346m|Si0063                  ----------MKPAAR----------KDAGAAAGVAAFG---  11
Bdi_Bradi1g42560.1|Bradi1g4256                  ----------MKPAVK----------KDAGG-VGGGAFG---  12
Osa_LOC_Os06g22330.1|LOC_Os06g                  ----------MKPAGGRS--------KDGGG-GGAALLGGGG  13
Sbi_Sb10g024570.1|Sb10g024570.                  ----------MRKDAA----------PAG----IAPG-SAPA  14
Zma_GRMZM2G121621_T01|GRMZM2G1                  ----------MRKDAA----------AAG----IAPG-SAPA  15
Sit_Si006264m|Si006264m|Si0062                  ----------MRKDAA----------GGG----IAPGAAAPA  16
Bdi_Bradi1g35710.1|Bradi1g3571                  ----------MRKEST----------GGGGG-IAAASPPPA   17
Osa_LOC_Os06g41910.1|LOC_Os06g                  ----------MRKDAA----------AGGGGGGGIAAGPAPA  18
Rco_30170.m014246|30170.m01424                  ----------MRKDCPP----------LSS-------VTGGKL-  19
Ptr_POPTR_0005s28030.1|POPTR_0                  ----------MRKDCQP----------PSSG------ATVSKL-  20
Mtr_CU179634_27.1|CU179634_27.                  ----------------------------------------MAKT-  21
Aco_AcoGoldSmith_v1.003723m|Ac  MFKTTPNSKTTLPLELAGKKPIMRKDSPPP----PLSGG-----TTPGKIS  22
Ath_AT2G33570.1|AT2G33570.1|AT                  ----------MRKEVLP----------------------PVLSTT  1
Smo_77156|77156|synonym_                        ----------MPITQS----------------------------  23
Rco_28563.m000093|28563.m00009                  ----------MGKER-EKERGGGGGGGDKKMFIGVVMN  24
Ptr_POPTR_0001s12330.1|POPTR_0                  ----------MGKERSEKERGGGEHNKRNLFVSCVVMN  25
Mtr_Medtr8g084210.1|Medtr8g084                  ----------MAIKD----------KEKKLFLWSSRN  26
Ath_AT4G20170.1|AT4G20170.1|AT                  ----------MAMVKEK--------EQNTKDKKLLVGVTWN  3
Ath_AT5G44670.1|AT5G44670.1|AT                  ----------MAKER----------DQNTKDKNLLICFTWN  2
Aco_AcoGoldSmith_v1.003774m|Ac                  ----------MVKEK----------EKDQKEKDNKLFVGVVWN  27
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.                  ----------MARQWLRKSELNEKLWKGDKGESLLDSS  28
Ppa_Pp1s112_159V6.1|Pp1s112_15                  ----------MARQWLRKSELNDKLWKGDKGDTVLDTS  29
Ppa_Pp1s32_144V6.1|Pp1s32_144V                  ----------MKMDKNHTVVKVEKETEKVNAKVWSREL  30
Ppa_Pp1s198_141V6.1|Pp1s198_14                  ----------MKMEKHQTVIKVENETGRVKDKVRSKEL  31
Ppa_Pp1s227_17V6.1|Pp1s227_17V                  ----------MEREHSVLKVEAGREKACSRP-----QL  32
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.                  ----------MDFKDTAAR---ESGKVKEKVGSKEV  33
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.                  ----------MKSQK-SLVENIPT----------SGLG  34
Ppa_Pp1s98_254V6.1|Pp1s98_254V                  ----------MKSSTGKILERIPA----------SGTC  35
Ppa_Pp1s25_268V6.1|Pp1s25_268V                  ----------MKSKTDWPQGRVG----------L      36
Ppa_Pp1s225_53V6.1|Pp1s225_53V                  ----------MVKPDFTQCRLRKFLGRFSLNSKTTGFSLKR  37
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.                  ----------MSYSVSQQFSNILSTRCTACLGQALI       38
Smo_406883|406883|406883|synon                  ----------MAAFHKAMICRIWIVLLVTLLWIGLWYSRIS  39
C_elegans_NP_504545.2                           ----------------------------------------   40
pigeon_ADC84389.1                               ----------------------------------------   41
```

FIG. 1

```
Sbi_Sb04g030080.1|Sb04g030080.  --CVDIKSFVASLAFLTLFVAFWQLQPYGS--LLTAARTTASSS--PCTL
Zma_GRMZM2G164912_T01|GRMZM2G1  --CVDIKSFVASLAFLTLFVVFWQLQPYGS--VLTAARTSASP---PCTL
Sit_Si017001m|Si017001m|Si0170  ---CVDVKSFVASLAFLTLFVAFWQLQPYGS--MLTAARTSASS---PCSL
Osa_LOC_Os02g48190.1|LOC_Os02g  --CVDIKSFVLSLAFLTVFVALWQLQPYGS--LLTAARSTASVSASPCSL
Bdi_Bradi3g53170.1|Bradi3g5317  --CVDIKLFVVSLAFLTILLGLWQLQPYGS---FMAAARSSANAP--PCFL
Sbi_Sb10g013010.1|Sb10g013010.  VSCFDIKSFVASLALLTLVMALWQLHPYQP--LLSASRPSTTCPLLPRPP
Zma_GRMZM2G136800_T02|GRMZM2G1  VSCFDIKSFVASLALLTLVMALWQLHPYQP--LLSASRPSSICPLLPRPT
Sit_Si006346m|Si006346m|Si0063  VSCFDIKSFVASLALLTLVMALWQLHPYQP--LLSTSR-SSSCPLLPSQP
Bdi_Bradi1g42560.1|Bradi1g4256  ISCLDIKSFAASLALLTLIMALWQLHPCQP--LLAITR-SPSCPILP---
Osa_LOC_Os06g22330.1|LOC_Os06g  VTCFDVKSFVASLALLTLIMALWQLHPYQPLVLLPAALSSSPCPLLPR--
Sbi_Sb10g024570.1|Sb10g024570.  LLCFDLKPFLAALTVLTLLAAAWQLR---AYHSVLASPFSAACPQPTAAG
Zma_GRMZM2G121621_T01|GRMZM2G1  FLCFDLKPFLAALTVLTLLAAAWQLR---PYQSLLASPFSAACPQPTAG-
Sit_Si006264m|Si006264m|Si0062  LLCFDVKPFLAALTVLTLLAAAWQLR---PYHSLLASPFPAACAQAAAG-
Bdi_Bradi1g35710.1|Bradi1g3571  LLCFDLKPFLAALTVVTLLTAAWQLRPYYQYHSILASSPFPACPDPPAS-
Osa_LOC_Os06g41910.1|LOC_Os06g  LLCFDLKPFLAALTVLTLITAAWQLR---SYSSVLP-SPFPVCADP-AA-
Rco_30170.m014246|30170.m01424  PSCLETKPLVATLLALTLVMLLWNLPPYYQNLLS-TTRP-CSAPAAAAAL
Ptr_POPTR_0005s28030.1|POPTR_0  PACFETKPLVATLLALTLVMLLWNLPPYYQNLLSTTTRPSCSAPETTVSI
Mtr_CU179634_27.1|CU179634_27.  PSN-----LVVTLLAFTLVMLIWNLQPYYDTILN--------PPAS----
Aco_AcoGoldSmith_v1.003723m|Ac  LLCFETKPIFAILALTLVMFIWNIQPYYDNLLS-TTRPCTATPEKPKPL
Ath_AT2G33570.1|AT2G33570.1|AT  TVCFEKKPIIATLLALSLVMIVWNLPPYYHNLIS-TARPCSAVTTTTTTT
Smo_77156|77156|syonym_         ---CENSSSNLSVSTNFGPAAAIDPIP-----------------------
Rco_28563.m000093|28563.m00009  -CAAELKLLILTALLICTIATLLQFLPSRFTISTSDLRFCISRITTTTTT
Ptr_POPTR_0001s12330.1|POPTR_0  SCAAELKLLILTALLVICSVATLFQILPPRFTISASDLRFCTSRISTSNTS
Mtr_Medtr8g084210.1|Medtr8g084  -YAAEHKLFLITTLLLLCTIATLFCFTIPSSFTISASDLRLCISRISQT---
Ath_AT4G20170.1|AT4G20170.1|AT  -FSAELKLTFMALLVLCTLATLLPFTIPSSFSLSTSDFRFCISRFSS-AVP
Ath_AT5G44670.1|AT5G44670.1|AT  -FSAELKLALMALLVLCTLATLLPFLPSSFSISASELRFCISRIAVNSTS
Aco_AcoGoldSmith_v1.003774m|Ac  -CATETKILLTALLFLSSLVIILQFLPSRFSLSSSDLQLCLSKLSSSTST
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  VAGVRTALSAESLAKILLLVVVNSVTLLAALHLLNSKREFDNRGDRSNCV
Ppa_Pp1s112_159V6.1|Pp1s112_15  MAGVRTALSAESLAKFILLVGINSVFLLAALHVLHSNREFENPGDRSNCI
Ppa_Pp1s32_144V6.1|Pp1s32_144V  GCHTAIKFQMRTFPDVRVNVVYTMIFLIGLILTAAQLWPLTDFD-PKHSC
Ppa_Pp1s198_141V6.1|Pp1s198_14  RCHTAIKFHMRAFPDIRVNVIYTMVLAIGLCLVVAQFSPLLGFD-PKVWC
Ppa_Pp1s227_17V6.1|Pp1s227_17V  GCSTPIKFST--------IRINIIYPMVMAIVLFLVVTQLLSLSGFG-IKFWC
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  DCQAAVKFHMSAFSDLRIYLVYVLIAACCFLILLAQLLPLRSLDKSQLPC
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  FNAVVVLVVINCLLLVFQISSSGSLYERAWWPTEAQP-----------
Ppa_Pp1s98_254V6.1|Pp1s98_254V  FSVVAVLVILNCLLLGFQISSNAGRYESTWWLASAPP-----------
Ppa_Pp1s25_268V6.1|Pp1s25_268V  LNVVVILLLTNVLLLLALLTTDIRISQDLWWPKPGSTGP----------
Ppa_Pp1s225_53V6.1|Pp1s225_53V  ATATLFLVILPLVLLVTLLHELSGLRTGIWLSTSRSQAQGPRESSETARK
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  FPCWFARLTIAGFLEYRSCKGFPSISDKLHAASAKDNYLGFGSIQGSQAK
Smo_406883|406883|406883|synon  PATSDLAAHRKDLHRFGGNLARNSSFSKFWEVEDGQLTADLIHEVVEELE
C_elegans_NP_504545.2           -----MPRITASKIVLLIALSFCITVTIYHFPIATRSSKEYDEYGNEYENV
pigeon_ADC84389.1               -MIHCGKKHICAFVTCILISASILMYSWKDPQLQNNITR-----------
```

FIG. 1 (Cont. 1)

```
Sbi_Sb04g030080.1|Sb04g030080.          LTTTAATDLASPDATSGTAANTNQ----PAAAKAAASDTAAASNAAPVRL
Zma_GRMZM2G164912_T01|GRMZM2G1          LPTTAAIDLASEDTTSSTAAKTDQ----PAAA-AAVSATAAGANAAPVWL
Sit_Si017001m|Si017001m|Si0170          LATTAAADLPSYNATAGAAADTKQ----PAAV----SGTTTAAANAAHVRL
Osa_LOC_Os02g48190.1|LOC_Os02g          IATPAAA-----AASAGVIRSENS----TADT--AKTAPAAVASAVPARL
Bdi_Bradi3g53170.1|Bradi3g5317          LSTTVAAVNNEVASSDSIPIKGTA----ATSDVPANIATVPARPARPARV
Sbi_Sb10g013010.1|Sb10g013010.          ISASSRAAATVAPFPSGNSSANAA----STKAASSAVPAVTTTKPAASVV
Zma_GRMZM2G136800_T02|GRMZM2G1          ISASSRAASTAAAFPSANSSTNAA----STKAASSAAPAVTTTKPAASVL
Sit_Si006346m|Si006346m|Si0063          ISASSRAATAAALPTANSTTADAA----DTKTAPSTVPAATATKPDAAVL
Bdi_Bradi1g42560.1|Bradi1g4256          LSHTTNATAALPSSTTKLPSSNFT----TDRAP----------------PTAAAA
Osa_LOC_Os06g22330.1|LOC_Os06g          -SPTSGIAVSFLSTAAATNSTDTA----TVPTTTAAARVAATTRPTLPAR
Sbi_Sb10g024570.1|Sb10g024570.          ASLPRALAVHANKASPSSSA----------------------AAATNSTASSS--
Zma_GRMZM2G121621_T01|GRMZM2G1          VSLPRALAVHAKKAS----------------------------SATNSTVSSS--
Sit_Si006264m|Si006264m|Si0062          -TLPRALAVHAKKSSSSS----------------------STPNYTASSSSL
Bdi_Bradi1g35710.1|Bradi1g3571          SSPPRLLAIDGKASKSNASSSSSP----------------LPKQDERR----
Osa_LOC_Os06g41910.1|LOC_Os06g          LSPPRALAVRTVASSGNASVSSDPGGPP----------ASLPEVGNKKPVA
Rco_30170.m014246|30170.m01424          AASN------ASSLPITSVSEQKYST----------------GVS-------
Ptr_POPTR_0005s28030.1|POPTR_0          SNT-----SSSFTSTSLSDQKYLS----------------SSSSS-------
Mtr_CU179634_27.1|CU179634_27.          ----------------SNITD----------------------------------
Aco_AcoGoldSmith_v1.003723m|Ac          PLQH----STTAQKISALTQKP----------------------------------
Ath_AT2G33570.1|AT2G33570.1|AT          LLSSSNFTSAENFTTSLSTTTAAA----------------SQKYD---------
Smo_77156|77156|syonym_                 ----------APEEQGTGIHS----------------------------------
Rco_28563.m000093|28563.m00009          TTTTTSIALPLNSSSVSPQSSPPSS----------LQQPEEQLPPNG-
Ptr_POPTR_0001s12330.1|POPTR_0          IASLN--STPTSTPTMPPSPSPPS----------IRKDQVADNG-
Mtr_Medtr8g084210.1|Medtr8g084          ---------PPTPTATPPSPSPPPP----------SSIVHEKLITNTN
Ath_AT4G20170.1|AT4G20170.1|AT          LNTTTTVEE----SSSSPSPEK----------NLDRVLDNG-
Ath_AT5G44670.1|AT5G44670.1|AT          VNFTTVVEKPVLDNAVKLTEKPVLDN----------GVTKQPLTEEKVLNNG-
Aco_AcoGoldSmith_v1.003774m|Ac          STSEVPQVSVLSSSSSPLIQK----------DQIIQQG-
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.          TNYRLNNVHPMSLSLDQNKLIQSS----------EFAPKNESN--
Ppa_Pp1s112_159V6.1|Pp1s112_15          TSDRLNNVHPISLSQDQNRLIKTS----------ESTTNNEL---
Ppa_Pp1s32_144V6.1|Pp1s32_144V          TQPANQENWERLDKSEPQQPQIFS----------ESFLHEK-VDASS
Ppa_Pp1s198_141V6.1|Pp1s198_14          AVMAKLEKLGKPEVSHP--PQLIS----------EALFHEK-INTSS
Ppa_Pp1s227_17V6.1|Pp1s227_17V          TETTKVVGLET------QLSPSLPQ----------GSFLDEN-IDTSS
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.          NDSVSIVSPETN---LEFPIDPLS----------KSFLHENIINKSS
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.          ----------MPCANITIAMDEN----------------------------
Ppa_Pp1s98_254V6.1|Pp1s98_254V          ----------MACTNASFALDEK----------------------------
Ppa_Pp1s25_268V6.1|Pp1s25_268V          ----------TLRLSDDITFSPSELI----------SESTQNQHRA
Ppa_Pp1s225_53V6.1|Pp1s225_53V          FLSRAHGQNPKRLSDETYPLQRPL----------NPKTVMSTVF
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.          YSEPKVVKEWQLPSIVADRAAEKLN----------------------------
Smo_406883|406883|406883|synon          KIRDGTPGGVVAANRKEPCDERRD----------------------------
C_elegans_NP_504545.2                   ASIESDIKNVRRLLDEVPDPSQNR----------------------------
pigeon_ADC84389.1                       ----------KIFQATSALPASQLCR----------------------------
```

FIG. 1 (Cont. 2)

```
Sbi_Sb04g030080.1|Sb04g030080.          AE-----AARLARPEDPNKRVLRPYGSAAALFVQMGAYRGGPRTFAIVGLA
Zma_GRMZM2G164912_T01|GRMZM2G1          AK-----AARPARPEDPNKRVFRPYGSAAALFVQMGAYRGGPRTFAIVGLA
Sit_Si017001m|Si017001m|Si0170          A----------RPARPEDPNKRVLRPYGSAAALFVQMGAYRGGPRTFAVVGLA
Osa_LOC_Os02g48190.1|LOC_Os02g          AR-----AARPARVEDPNKRELRPYGSAAALFVQMGAYRGGPRTFAIVGLA
Bdi_Bradi3g53170.1|Bradi3g5317          ENQNKRVLRPAPAVDPNKRALRPYGSAAALFVQFGAYRGGPRTFAIVGLA
Sbi_Sb10g013010.1|Sb10g013010.          PA----------ARPRDPNKREFRSYGSAAALFVQMGAYRGGPRTFAVIGLA
Zma_GRMZM2G136800_T02|GRMZM2G1          PA----------ARPRDPNKREFRSYGSAAALFVQMGAYRGGPRTFAVIGLA
Sit_Si006346m|Si006346m|Si00063         PA----------ARPRDFNKRDLRPYGSAAALFVQMGAYRGGPRTFAVVGLA
Bdi_Bradi1g42560.1|Bradi1g4256          RP----------APLRDPNKRELRPYGSAAALFVQMGAYRGGPRTFAVVGLA
Osa_LOC_Os06g22330.1|LOC_Os06g          ----------QRERDPNKRELRPYGTAAALFVQMGAYRGGPRTFAVVGLA
Sbi_Sb10g024570.1|Sb10g024570.          ----------SAPPPGPERREFRAVGSAAALFVQMGAYRGGPYTFAVVGLA
Zma_GRMZM2G121621_T01|GRMZM2G1          ----------PSPPPGPERREFHAVGSAAALFVQMGAYRGGPYTFAVVGLA
Sit_Si006264m|Si006264m|Si0062          ----------PPPPPGPERREFRAVGSAAALFVQMGAYRGGPYTFAVVGLA
Bdi_Bradi1g35710.1|Bradi1g3571          ----------AAPDPNRREFRAVGSAAALFVQMGAYRGGPYTFAVVGLA
Osa_LOC_Os06g41910.1|LOC_Os06g          ----------AAAAADPNRREFRAVGSAAALFVQMGAYRGGPYTFAVIGLA
Rco_30170.m014246|30170.m01424          ----------DPNKRIFEAYGNAAALFVKMGAYRGGPRTFAVVGLA
Ptr_POPTR_0005s28030.1|POPTR_0          ----------SSNADPNKRIFQAYGNAAALFVQMGAYRGGPTTFAVVGLA
Mtr_CU179634_27.1|CU179634_27.          ----------PNKRTFITYGNAASLFVQMGAYRGGPTTFAVVGLA
Aco_AcoGoldSmith_v1.003723m|Ac          ----------ADPNKRTFKTYGNAAALFVQMGAYRGGPNTFAIVGLA
Ath_AT2G33570.1|AT2G33570.1|AT          ----------STPSDPNKRVFQPFGNAAALFVLMGAYRGGPTTFSVIGLA
Smo_77156|77156|synonym_                ----------TVRHLHPFGIASYTFVLTGGYRTQRRSFAVVGLA
Rco_28563.m000093|28563.m00009          ----------TVKRAFYPHGAAAYNFIAMGAYRGGFDTFAIVGLS
Ptr_POPTR_0001s12330.1|POPTR_0          ----------VIKRVFNPYGSAAYNFITMGAYRGGLNTFAIIGLA
Mtr_Medtr8g084210.1|Medtr8g084          TT----------TIKRFFNPYGSAAYNFITMSAYRGGLNTFAITGLS
Ath_AT4G20170.1|AT4G20170.1|AT          ----------VIKRTFTGYGSAAYNFVSMSAYRGGVNSFAVIGLS
Ath_AT5G44670.1|AT5G44670.1|AT          ----------VIKRTFTGYGWAAYNFVLMNAYRGGVNTFAVIGLS
Aco_AcoGoldSmith_v1.003774m|Ac          ----------IIKRAFNSYGSAAYSFIQMGSYRGGLNTFAIIGLA
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.          ----------QENIRQFDTFGVAVHLFIKMSAYRGGSNSFAIIGLE
Ppa_Pp1s112_159V6.1|Pp1s112_15          ----------QENVRQFDTFGVAVYLFIKMSAYRDGPRSFTIVGLE
Ppa_Pp1s32_144V6.1|Pp1s32_144V          ----------IDTSHGELRVFKPHGLATHLYIEMSAYRGGPRQFSVVGLT
Ppa_Pp1s198_141V6.1|Pp1s198_14          ----------IDLTHGELRLFKPHGLATHLYIEMSAYRGGPRVFSIVGLA
Ppa_Pp1s227_17V6.1|Pp1s227_17V          ----------IDLRHGDVRLFKPHGLATYLYIDMSAYRGSPRHFSIVGLT
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.          ----------LDISKGELRVFKPHGLAMHLFIEMGAYRGGPRVFSVVGLI
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.          -----LDMSSIDLSRGEVRSFRAHGIATRLIVEVGSYRNGPRTFSSVVLT
Ppa_Pp1s98_254V6.1|Pp1s98_254V          ------LNMSSIDLSKGEVRSFRAHGVASRLIIEVGSYRNGPRTFSSVVMT
Ppa_Pp1s25_268V6.1|Pp1s25_268V          LYPVTLNTTTLDTSRGELRTFKPHGIASHLFIEFGAYRISPSQFSIVGLI
Ppa_Pp1s225_53V6.1|Pp1s225_53V          LRESNINASKFDLRRGEIRAYRPHGVAAHFFLQLGTYRSSPRAFSTAGSI
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.          ----------DSSYGLPRRFLPMGTATFLFVHFSTYRVAPKSFATIGLG
Smo_406883|406883|406883|synon          ----------LARERGFRRIGETLHAFVLSSVARFSLEEFTAVGLV
C_elegans_NP_504545.2                   ----------LQFLKLDEHAFAFSAYTDDRNGMMGYKYVRVL
pigeon_ADC84389.1                       ----------GKPAQNVITALEDNRTFIISPYFDDRESKVTRVIGIV
```

FIG. 1 (Cont. 3)

```
Sbi_Sb04g030080.1|Sb04g030080.  SK-PTHVFGTPYFKCEWLPNP---TAGEPSPR-PVRTKAYKMLPDWG---
Zma_GRMZM2G164912_T01|GRMZM2G1  SK-PTHVFGTPYFKCEWLPNP---TAGDPSPR-PVRTKAYKMLPDWG---
Sit_Si017001m|Si017001m.        SK-PTHVFGTPYFKCEWLPNP----TASDPSPR-PVRTKAYKMLPDWG----
Osa_LOC_Os02g48190.1|LOC_Os02g  SK-PTHVFSNPYFKCEWLPNP---TAGNPSPR-PVRTKAYKMLPDWG---
Bdi_Bradi3g53170.1|Bradi3g5317  SK-PTHVFGTPYFKCEWLPNP----SAADPAER-PVRTKAYKMLPDWG----
Sbi_Sb10g013010.1|Sb10g013010.  SK-PAHVYGTPYFKCEWVPNQ---DPSSPAPPRAVRTKAYKMLPDWG---
Zma_GRMZM2G136800_T02|GRMZM2G1  SK-PAHVYDTPYFKCEWVPNQ---EPSSPAPPRAVRTKAYKMLPDWG---
Sit_Si006346m|Si006346m.        SK-PAHVFGTPYFKCEWVPNL----DPSSPAPPRPVRTKAYKMLPDWG----
Bdi_Bradi1g42560.1|Bradi1g4256  SK-PAFVFNTPYFKCEWVPN----------PGAGEPVRTKAYKMLPDWG---
Osa_LOC_Os06g22330.1|LOC_Os06g  SK-PAHVFSNPYFKCEWLPN----------APAGAPPVRTKAYKMLPDWG---
Sbi_Sb10g024570.1|Sb10g024570.  SK-PTHVYGKPWFRCEWEP-----NGASSSPPMRAAKTYHMLPDWG---
Zma_GRMZM2G121621_T01|GRMZM2G1  SK-PTHVYGKPWFRCEWEP------SDAS---SPPMRAAKTYHMLPDWG----
Sit_Si006264m|Si006264m.        SK-PTHVYGKPWFRCEWEPTIPSNTNASSPPEPMRAAGTYHMLPDWG---
Bdi_Bradi1g35710.1|Bradi1g3571  SK-PTHVYGKPWFRCEWVPNN----------ASQPTRAGKAYHMLPDWG----
Osa_LOC_Os06g41910.1|LOC_Os06g  SK-PTHVYGKPWFRCEWVPNANG-SAAAAAARPMRAANTYHMLPDWG---
Rco_30170.m014246|30170.m01424  SK-PIHVFGRPWYKCEWISN----------NGSSMRAKAYKMLPDWG---
Ptr_POPTR_0005s28030.1|POPTR_0  SK-PIHVFRLPWYKCEWISN----------NGSSIRAKAYKMLPDWG---
Mtr_CU179634_27.1|CU179634_27.  SK-PLHVFGRPWYKCEWIPNT----------NTNSSKAKAYKILPDWG---
Aco_AcoGoldSmith_v1.003723m|Ac  SK-PTHVFGKAWYKCEWISN----------NGSSIKTKAYKMLPDWG---
Ath_AT2G33570.1|AT2G33570.1|AT  SK-PIHVYGKPWYKCEWISN----------NGTSIRAKAQKILPDWG---
Smo_77156|77156|synonym_        AK-SLHVFGKPNFACEWVPDS----------TTSDSPSIFGSARKILPDWG----
Rco_28563.m000093|28563.m00009  SK-PLHLYAKPTYQCEWVPHKN---SSGGGSDPITAVAYKILPDWG---
Ptr_POPTR_0001s12330.1|POPTR_0  SK-PLHVYSKPAYQCEWVPQ-----SSSASNSTFSTVSYKMLPDWG----
Mtr_Medtr8g084210.1|Medtr8g084  SK-PLHVYGNPTYECEWIPNTN----TNTNSSKNITTIGYKMLPDWG---
Ath_AT4G20170.1|AT4G20170.1|AT  SK-PLHVYGHPSYRCEWVSLD----------PTQDPISTTGFKILTDWG---
Ath_AT5G44670.1|AT5G44670.1|AT  SK-PLHVYSHPTYRCEWIPLN----------QSDNRILTDGTKILTDWG---
Aco_AcoGoldSmith_v1.003774m|Ac  SK-PLIVYGKPSYQCQWISDD---------DRNNNNNITTFGSKILPDWG---
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  AKNPGELYNDPPYECVWIPGP---------DSLVWAPLKGSAIKMLPDTGN--
Ppa_Pp1s112_159V6.1|Pp1s112_15  AKNPAELYNDPPYECVWIPGP---------DSVVWAPSKGSTFKMLPDTGN--
Ppa_Pp1s32_144V6.1|Pp1s32_144V  SK-PIESHHQPPYACEWVSNA----------TGEVVKGRPHKVLPDWD----
Ppa_Pp1s198_141V6.1|Pp1s198_14  SK-SLETHHNPPYACEWVSNS----------TGEVVQGRSHKVLPDWG---
Ppa_Pp1s227_17V6.1|Pp1s227_17V  AK-SIEKHHKPPYACEWVSSS----------NDKVVKGHAYKILPDWD----
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  AK-PIQTFHKPPYVCEWVTN----------VGRTIKGRASKILPDWN----
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  SK-RLNNLHDILYECEWA----------TADSPTLKVKARAIKPDWN---
Ppa_Pp1s98_254V6.1|Pp1s98_254V  SK-LLNELHDITHECEWV----------TSDSKALRVKAMAIKPDWN----
Ppa_Pp1s25_268V6.1|Pp1s25_268V  SK-RLHDLHDPPYNCTWR----------SGEVTVSALTTWPIKPDWE---
Ppa_Pp1s225_53V6.1|Pp1s225_53V  QR-SFEDWFKPRFQCEWIG----------SNGTAVRAKEAKKIRPAQS---
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  PK-ALFLYSNPAFTCSWHPAK----------EAAPVIKTNGTFLQPVPTH--
Smo_406883|406883|406883|synon  SIQLRGFAGNPLQSCAWVPKAG---------SSSVAAALAASNDSSVIWGNSW
C_elegans_NP_504545.2           MF-----ITSQDNFSCEINGRK----------STDVSLYEFSENHK---
pigeon_ADC84389.1               HH-----EDVKQLYCWFCCQP----------DGKIYVARAKIDVHSDR--
                                                *
```

FIG. 1 (Cont. 4)

```
Sbi_Sb04g030080.1|Sb04g030080.  ----------YGRVYTVVVVNCTFPS---NPNAGNAGGKLLVHAYYS----
Zma_GRMZM2G164912_T01|GRMZM2G1  ----------YGRVYTVVVVNCTFPS---NPNARNAGGKLLVHAYYS----
Sit_Si017001m|Si017001m|Si0170  ----------YGRVYTVVVVNCTFPS---NPNAGNAGGKLLVHAYYS----
Osa_LOC_Os02g48190.1|LOC_Os02g  ----------YGRVYTVVVVNCTFPS---NPNADNAGGKLLVHAYYS----
Bdi_Bradi3g53170.1|Bradi3g5317  ----------YGRVYTTVVVNCTFPS---NPNVDNAGGKLVHAYYS----
Sbi_Sb10g013010.1|Sb10g013010.  ----------YGRIYTVVVVNCTFPT---NPNADNRGGKLLIHAYYS----
Zma_GRMZM2G136800_T02|GRMZM2G1  ----------YGRIYTVVVVNCTFPT---NPNADNRGGKLFIQAYYS----
Sit_Si006346m|Si006346m|Si0063  ----------YGRIYTVVVVNCTFPT---NPNAGNRGGKLLVHAYYS----
Bdi_Bradi1g42560.1|Bradi1g4256  ----------YGRVYTVVVVNCTFPS---NPNAGNRGGKLLVHAYYS----
Osa_LOC_Os06g22330.1|LOC_Os06g  ----------YGRVYTVVVVNCTFPS---NPNADNLGGKLLVHAYYS----
Sbi_Sb10g024570.1|Sb10g024570.  ----------YGRVYTVVVVNCTFPR---VPNADNAGGRLILYAHYG----
Zma_GRMZM2G121621_T01|GRMZM2G1  ----------YGRVYTVVVVNCTFPR---VPNADSAGGRIVLYAHYG----
Sit_Si006264m|Si006264m|Si0062  ----------YGRVYTVVVVNCTFPR---VPNADNAGGRLVLYAHHGG---
Bdi_Bradi1g35710.1|Bradi1g3571  ----------YGRVYTVVVVNCTFSS---FAPNADNAGGKLLNAYYG----
Osa_LOC_Os06g41910.1|LOC_Os06g  ----------YGRVYTVVVVNCTFAR---VPNADNAGGKLVLNAYYG----
Rco_30170.m014246|30170.m01424  ----------YGRVYTVVVVNCTFPL---NPNRENAGGKLILNAYYG----
Ptr_POPTR_0005s28030.1|POPTR_0  ----------YGRVYTVVVVNCTFPV---NPNQDNAGGRLMLNAYYD----
Mtr_CU179634_27.1|CU179634_27.  ----------YGRVYTVVVVNCTFTV---NPNLDNNGGKLILNAYYS----
Aco_AcoGoldSmith_v1.003723m|Ac  ----------YGRVYTVVVVNCTFTV---NPNEDNSGGKLVVYAYYG----
Ath_AT2G33570.1|AT2G33570.1|AT  ----------YGRVYTVVVVNCTFNS---NFNSDNTGGKLILLNAYYN----
Smo_77156|77156|77156|synonym_  ----------YGRVYTVVVVHCYFKS---SVGEDGSGGRLILLHAGEK----
Rco_28563.m000093|28563.m00009  ----------YGEVYTVVVVNCTFA--Q-PINSDNSGGKLFLHASTSGG--
Ptr_POPTR_0001s12330.1|POPTR_0  ----------YGHVYTVVVVNCTFS---E-AVNSENSGGKLFLEASTSGG--
Mtr_Medtr8g084210.1|Medtr8g084  ----------YGHVYTVVIVNCTFN---E-SINVDNSGGKLMLYASTSGG--
Ath_AT4G20170.1|AT4G20170.1|AT  ----------YGRIYTTVVVNCTFSSIS-AVNPQNSGGTLILHATT-----
Ath_AT5G44670.1|AT5G44670.1|AT  ----------YGRVYTTVVVNCTFPSNT-VINPKNTGGTLLLHATT-----
Aco_AcoGoldSmith_v1.003774m|Ac  ----------YGEVYTVVVINCTFP--I-PVGINGSGGRLILHASDGGS--
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  ----------V-YSRLYSAVIINCTFSQ---DVGVDRKGGQLVLYASYG----
Ppa_Pp1s112_159V6.1|Pp1s112_15  ----------V-YSRLYSAVIINCTFSQ---DVGVDRKGGQLVLYASYG----
Ppa_Pp1s32_144V6.1|Pp1s32_144V  ----------YGKLYTVVVITCNFAK---DVGVDGEGGELILYASYG----
Ppa_Pp1s198_141V6.1|Pp1s198_14  ----------YGKLYTVVVITCNFAK---DVGVDGEDGELILLASYG----
Ppa_Pp1s227_17V6.1|Pp1s227_17V  ----------YGKLYTVVVITCVFAE---DVGTDGEGGELVLYASYG----
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  ----------YGRLYTVVVITCTFRV---DVGVQKEGGELVLQVSYG----
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  ----------MGELYGTMVIVCTFPK---DVGTKSEGGRLILT-AHYPG--
Ppa_Pp1s98_254V6.1|Pp1s98_254V  ----------MGRLYGTMVLVCNFPR---DVGTDAEGGQMILT-AGYSD--
Ppa_Pp1s25_268V6.1|Pp1s25_268V  ----------LGLMYGTMVVVCIFPE---DVGTAGEGGSLQLS-IGYAD--
Ppa_Pp1s225_53V6.1|Pp1s225_53V  ----------FKGFYDSLIIICHFDE---DVGTDGLGGKLYMKYWHQVD--
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  ----------VSYGKQYTGTAVYCNFDT---PIGNDGSGGRLEIQVTHG----
Smo_406883|406883|406883|synon  LGYVNESHTEKTFYDVGVVKCSFEEP---VGANGEGGFLVLNMTRNSSPP
C_elegans_NP_504545.2  ----------MKWQMFILNCKLPD---GIDFNNVSSVKVIRSTTK----
pigeon_ADC84389.1  ----------FGFPYGAADIVCLEPE-----NCNFTHVSIHQSPHAN----
                                    :  :*  .
```

FIG. 1 (Cont. 5)

```
Sbi_Sb04g030080.1|Sb04g030080.          ----TASRYERFVALEEAPGSYDESLFS----PPFQYDYLYCGSSLYG
Zma_GRMZM2G164912_T01|GRMZM2G1          -------TASRRHERFVALEEAPGSYDESRFS-------PPFQYDYLYCGSSLYG
Sit_Si001700lm|Si001700lm|Si0170        -------TASRR-----------------------------PPFQYDYLYCGSSLYG
Osa_LOC_Os02g48190.1|LOC_Os02g          -------TTSRRYERFVALEEAPGSYDESRFS----PPFPYDYLYCGSSLYG
Bdi_Bradi3g53170.1|Bradi3g5317          -------IHSRRYERFVALEEAPGSYDESRFS----PPFQYDYLYCGSSLYG
Sbi_Sb10g013010.1|Sb10g013010.          -------TASRRYERFVALEETRGSYDESRFR----PPFPYEYLYCGSSLYG
Zma_GRMZM2G136800_T02|GRMZM2G1          -------TASRRYERFVALEEVPGSYDESRFR-------PPFPYEYLYCGSSLYG
Sit_Si006346m|Si006346m|Si0063          -------TASRRYERFVALEEAPGSYDESRFR----PPFPYEYLYCGSSLYG
Bdi_Bradi1g42560.1|Bradi1g4256          -------TKSRRYERFLALEEAPGSYDESRFR-------PPFPYEYLYCGSSLYG
Osa_LOC_Os06g22330.1|LOC_Os06g          -------TASRRYERFVALEEAPGSYDDARFR----PPFAYDYLYCGSSLYG
Sbi_Sb10g024570.1|Sb10g024570.          --PSRSPASRHERIVALEESPGAYDEAAFRT-TPPQHRYDYLYCGSSLYG
Zma_GRMZM2G121621_T01|GRMZM2G1          --PSRSPASRRERIVALEESPGAYDEAAFRT-TPP-HRYDYLYCGSSLYG
Sit_Si006264m|Si006264m|Si0062          --PSRSPSSPHERIVALEEAPGAYDEAAFRP-GAP-HRYDYLYCGSSLYG
Bdi_Bradi1g35710.1|Bradi1g3571          -----PSP-ARYERIVALEEAPGAYDSAAF-------AQHRYDYLYCGSSLFG
Osa_LOC_Os06g41910.1|LOC_Os06g          -----ASP-ARYERIVAMEEAPGAYDAAEFR-----PPHRYDYLYCGSSLYG
Rco_30170.m014246|30170.m01424          ---------ESPRKYEKIVALEEAPGSYNDSNYH-------PPYQYEYLYCGSSLYG
Ptr_POPTR_0005s28030.1|POPTR_0          -----ESQRKYEKFMALEELFGSYNESKFR----PPYQYEYLYCGSSLYG
Mtr_CU179634_27.1|CU179634_27.          ---------ESPKRYEKITALEEAPGSYNQSKFS-------PPYTYEYLYCGSSLYG
Aco_AcoGoldSmith_v1.003723m|Ac          -------PSQKKYERIVALEEEAGSYNELKYK----PPYKYDYLYCGSSLYG
Ath_AT2G33570.1|AT2G33570.1|AT          -------ESPKLFERFTTLEESAGIYDESKYS----PPYQYDYLYCGSSLYG
Smo_77156|77156|synonym_                ---------GSPSPPSKITALVEAPGSYSPSQFLP-------KYDYLYCGSSLFG
Rco_28563.m000093|28563.m00009          --GDRDFNI-TDRFEVLQEPPGNLNISLFTT----KPKYDYLYCGSSLYG
Ptr_POPTR_0001s12330.1|POPTR_0          --GDRDFNI-TDRFEVLNESPGDINTTLFSS----KPKYDYLYCGSSLYG
Mtr_Medtr8g084210.1|Medtr8g084          --GDTKFNI-TDRMEVLVEQPKVLDITLFNS----KPKLDYFYCGSSLFG
Ath_AT4G20170.1|AT4G20170.1|AT          --GDPTLNL-TDSISVLTEPPKSVDFDLYNS--TKKTKKYDYLYCGSSLYG
Ath_AT5G44670.1|AT5G44670.1|AT          --GDTDRNI-TDSIPVLTETFNTVDFALYESNLRRREKYDYLYCGSSLYG
Aco_AcoGoldSmith_v1.003774m|Ac          ---GDTDVVSTEKIEALVEAPGSLNASVFSS-------PPVYDYFYCGSPLYG
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.          ----DQYFRNPERIVALTEQPDEFFGVEFYE--SSDMKYDYVYCGSPLFG
Ppa_Pp1s112_159V6.1|Pp1s112_15          ----DQFFRNPERIVALTEAPDEFLGMEYYE--SPEMKYDYVYCGSPLFG
Ppa_Pp1s32_144V6.1|Pp1s32_144V          -------DQYRQPERIVVLTESKGTYNSSIFDP----KNFPYDYVYCGSSVYG
Ppa_Pp1s198_141V6.1|Pp1s198_14          -------DQYRQPERIKVLSESKGEYNGSIFNP----QNFPYDYVYCGSSVYG
Ppa_Pp1s227_17V6.1|Pp1s227_17V          -------DRYRQPERVVVLTEVKGEYNSSVFDT----KSFPYDYVYCGSSVYG
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.          -----DQFRQPEKIVVLTERGQYNATMFDP----P-YPYDYVYCGSSLYG
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.          ---------AFRTPERFIALTEMQGEYNASRFQP-------PYPYDMTFCGSPLYG
Ppa_Pp1s98_254V6.1|Pp1s98_254V          -------SYRIPERFVALTEMRGEYSASHFQQ----PYSYDLAFCGSPLYG
Ppa_Pp1s25_268V6.1|Pp1s25_268V          ---------EFRQPENFVVLTEEAGKYNASLWMP-------PFPYDITFCVSRLYN
Ppa_Pp1s225_53V6.1|Pp1s225_53V          -------KYRTPETFLALVEAPSEYNASMFEP----PFAYDIVYCGSPIYG
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.          --TQDASGLEDVNFVAKEEGAGEFNASLYQE----PYQYDHVYCGAPIYG
Smo_406883|406883|406883|synon          PLVDHSGIIEQSYVPVFQELPEEISSIQFEG-------PFRFDYAYCSPLLTP
C_elegans_NP_504545.2                   ----------QFVDVPIRYRIQDEKIITP----DEYDYKMSICVPAILFG
pigeon_ADC84389.1                       ----------IDQLPSFKIKNRKSE-------TFSVDFTVCISAMFG
                                                                          * . :
```

FIG. 1 (Cont. 6)

```
Sbi_Sb04g030080.1|Sb04g030080.          -------NLSASRMREWVAYHAHFLGA-RSHFVLHDAGGIS-PEVKAVLD
Zma_GRMZM2G164912_T01|GRMZM2G1          ---------NLSASRVREWVAYHARFLGP-RSHFVLHDAGGIS-PEVKAVLD
Sit_Si017001m|Si017001m|Si0170          ---------NLSASRMREWVAYHAHFFGP-RSHFVLHDAGGIS-PEVKAVLD
Osa_LOC_Os02g48190.1|LOC_Os02g          ---------NLSASRMREWVAYHAHFFGP-RSHFVFHDAGGIS-PEVKAVLD
Bdi_Bradi3g53170.1|Bradi3g5317          -------NLSASRMREWMAYHAHFFGP-RSHFVFHDASGFS-PEVKAVLD
Sbi_Sb10g013010.1|Sb10g013010.          ---------NLSAARMREWLAYHAHFFGP-ASHFVLHDAGGVS-PEVRAVLD
Zma_GRMZM2G136800_T02|GRMZM2G1          -------NLSAARMREWLAYHAHFFGP-KSHFVLHDAGGVS-TEVRAVLD
Sit_Si006346m|Si006346m|Si0063          ---------NLSAARMREWLAYHAHFFGP-SSHFVLHDAGGVS-PEVWAVLD
Bdi_Bradi1g42560.1|Bradi1g4256          ---------NLSAFRLREWLAYHARFFGP-RSHFVLHDAGGVS-PEVRAVLD
Osa_LOC_Os06g22330.1|LOC_Os06g          ---------NLSSARMREWLAYHARFFGP-RSHFVLHDAGGVT-PEVRAVLD
Sbi_Sb10g024570.1|Sb10g024570.          ---------NLSAARVREWMAYHARFFGA-RSHFVFHDAGGVS-PAVRAALQ
Zma_GRMZM2G121621_T01|GRMZM2G1          -------NLSAARVREWMAYHARFFGA-RSHFVFHDAGGVG-PAVRAALE
Sit_Si006264m|Si006264m|Si0062          ---------DISAARVREWMAYHARFFGD-RSHFVFHDAGGVG-PAVRAALE
Bdi_Bradi1g35710.1|Bradi1g3571          -------ELSAARVREWMAYHARFFGP-RSHFVFHDAGGVTSPAVRAALE
Osa_LOC_Os06g41910.1|LOC_Os06g          ---------NLSAARVREWMAYHARFFGL-RSHFVFHDAGGVS-PAVRAALE
Rco_30170.m014246|30170.m01424          -------NLSAARMREWMAYHAWTFGP-SSHFVFHDAGGVS-PQVRAALE
Ptr_POPTR_0005s28030.1|POPTR_0          ---------NLSASRFREWMAYHAWFFGP-SSHFVFHDAGGVS-PEVRAALD
Mtr_CU179634_27.1|CU179634_27.          ---------NLSASRMREWMAYHAWFFGP-SSHFVFHDAGGVS-SEVRAVLE
Aco_AcoGoldSmith_v1.003723m|Ac          -------NLSAVRIREWMAYHAWFFGK-SSHFVFHDAGGVS-SEVRAVLE
Ath_AT2G33570.1|AT2G33570.1|AT          ---------NVSASRMREWMAYHAWFFGD-KSHFVFHDAGGVS-PEVRKVLE
Smo_77156|77156|77156|synonym_          -------DVNPQRLREWIAYHAWLFGP-RSHFVLHDAGGARSKGVRLVLE
Rco_28563.m000093|28563.m00009          ---------GLSPQRVREWMAYHVRLFGE-RSHFVIHDAGGVH-EEVLEVLK
Ptr_POPTR_0001s12330.1|POPTR_0          ---------GLSPQRVREWIAYHVRLFGK-RSHFVIHDAGGVH-EEVLEVLK
Mtr_Medtr8g084210.1|Medtr8g084          ---------NLNPQRVREWIAYHVRLFGP-NSHFVLHDAGGVH-EEVFEVLK
Ath_AT4G20170.1|AT4G20170.1|AT          -------NLSPQRVREWIAYHVRFFGE-RSHFVLHDAGGIH-EEVFEVLK
Ath_AT5G44670.1|AT5G44670.1|AT          -------NLSPQRIREWIAYHVRFFGE-RSHFVLHDAGGIT-EEVFEVLK
Aco_AcoGoldSmith_v1.003774m|Ac          ---------DLSPQRVREWIAYHVKLFGE-KSHFVIYDAGGIH-SGVLEVLK
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.          ---------NLSPQKIREWIAYHAHFFGP-RSHFFLYDAGGVH-DNVRRMIE
Ppa_Pp1s112_159V6.1|Pp1s112_15          ---------NLSPQKIREWIAYHAHTFGP-RSHFFLYDAGGIH-EEVRRVIE
Ppa_Pp1s32_144V6.1|Pp1s32_144V          -------DISPQRMREWMAYHAKLFGD-GSHFILHDSGGFH-DDVRKVIE
Ppa_Pp1s198_141V6.1|Pp1s198_14          ---------NISPQRMREWMAYHARFFGD-KAHFILHDSGGFH-EDVRKVLE
Ppa_Pp1s227_17V6.1|Pp1s227_17V          -------NISPQRMREWMAYHAKFFGE-KSHFILHDAGGFH-EDVRKVLE
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.          ---------DVSPQRMREWLAYHAHTFGE-RSHTMFHDAGGFH-PQVWKVLD
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.          -------NISPQRIREWIAYHAYFLGN-RTLFIFQDAGGIH-DDVYRVLK
Ppa_Pp1s98_254V6.1|Pp1s98_254V          -------NISPQRVREWIAYHAWFFGN-RTLFMFHDAGGIH-DDVYRVLK
Ppa_Pp1s25_268V6.1|Pp1s25_268V          ---------TVDAHRIREWLAYHAHLFGN-RTHFLFHNAGGLN-SDVYKVLK
Ppa_Pp1s225_53V6.1|Pp1s225_53V          ---------DISPQRVREWVAYHTYVFGN-RSHFILYDAGGFH-DDVQKVLE
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.          ----------SINAQRIREWIAYHVWFFGE-RTHFIYLYDAGGFD-EQVRAVLQ
Smo_406883|406883|406883|synon          SK-------NVSAKYVKEWLMYHQGLWSESHVHYFFYDAGGID-SRLLEVFQ
C_elegans_NP_504545.2                   -N---------GYDAKRIVEFIELNT---LQGIEKIYIYTNQKELD-GSMKKTLK
pigeon_ADC84389.1                       NYNNVLQFIQSVEMYKILGVQKVVIYKNNCSQLMEKVLKFYMEEGTVEII
                                                    .. . : : :  .                    :
```

FIG. 1 (Cont. 7)

```
Sbi_Sb04g030080.1|Sb04g030080.  PWVRAGRVTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Zma_GRMZM2G164912_T01|GRMZM2G1  PWVRAGRVTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Sit_Si017001m|Si017001m|Si0170  PWVRAGRVTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Osa_LOC_Os02g48190.1|LOC_Os02g  PWVRAGRLTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Bdi_Bradi3g53170.1|Bradi3g5317  PWVRAGRLTVQDVRAEAEYDS--YYYNQFLVVNDCLHRYRHAANWTFFFD
Sbi_Sb10g013010.1|Sb10g013010.  FWVRAGRVTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Zma_GRMZM2G136800_T02|GRMZM2G1  PWIRAGRVTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Sit_Si006346m|Si006346m|Si0063  PW----------------------------------------------
Bdi_Bradi1g42560.1|Bradi1g4256  PWVRAGRVTVQDIRAQAEYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Osa_LOC_Os06g22330.1|LOC_Os06g  PWVSAGRVTVQDIRAQEDYDG--YYYNQFLVVNDCLHRYRHAANWTFFFD
Sbi_Sb10g024570.1|Sb10g024570.  PWVRAGRATLQDVRAQAEYDG--WYYNQFLVVNDCLHRYRHAAKWTFFFD
Zma_GRMZM2G121621_T01|GRMZM2G1  PWVRAGRATLQDVRAQAEYDG--WYHNQFLVVNDCLHRYRHAAKWTFFFD
Sit_Si006264m|Si006264m|Si0062  PWVRAGRATLQDVRAQAEYDG--WYYNQFLVVNDCLHRYRHAAKWTFFFD
Bdi_Bradi1g35710.1|Bradi1g3571  PWVRAGRATLQDVRAQAEYDG--WYHNQFLVVNDCLHRYRHAAKWTFFFD
Osa_LOC_Os06g41910.1|LOC_Os06g  PWVRAGRATLQDVRAQAEYDG--WYYNQFLVVNDCLHRYRHAARWTFFFD
Rco_30170.m01424|30170.m01424   PWVRAGRATVQDIRGQAEFDG--YYYNQFLVVNDCLHRYRHAANWTFYFD
Ptr_POPTR_0005s28030.1|POPTR_0  PWVRAGRATVQDIRGQAEFDG--YYYNQFLVVNDCLHRYRYSANWTFYFD
Mtr_CU179634_27.1|CU179634_27.  PWVQAGRVTLQDIRDQSEFDG--YYYNQFLVVNDCLHRYRHSAYWTFYFD
Aco_AcoGoldSmith_v1.003723m|Ac  PWVKLGRVTIQDIRDQAEFDG--YYYNQFLVVNDCLHRYRNDANWTFYFD
Ath_AT2G33570.1|AT2G33570.1|AT  PWIRAGRVTVQNIRDQSQYTG--YYYNQFLIVNDCLHRYRYAANWTFFFD
Smo_77156|77156|77156|synonym_  PWVRSGRVTVQDFREQARFDG--YYHNQFLVVNDCLFRYKHAANWTFFFD
Rco_28563.m000093|28563.m00009  PWMELGYVTLQDIRDQERFDG--YYHNQFMVVNDCLHRYKFVAKWIFFFD
Ptr_POPTR_0001s12330.1|POPTR_0  PWMELGYVTLQDIKEQERFDG--YYHNQFMVVNDCLHRYKFMAKWMFFFD
Mtr_Medtr8g084210.1|Medtr8g084  FWIELGYVTLQDIRDQERFDG--YYHNQFMVLNDCLHRYKFMAKWMFFFD
Ath_AT4G20170.1|AT4G20170.1|AT  PWIELGRVTLHDIRDQERFDG--YYHNQFMIVNDCLHRYRFMTKWMFFFD
Ath_AT5G44670.1|AT5G44670.1|AT  PWIELGRVTVHDIREQERFDG--YYHNQFMVVNDCLHRYRFMAKWMFFFD
Aco_AcoGoldSmith_v1.003774m|Ac  PWIEKGYVTLQDVREQERFDG--YYHNQFLVVNDCLHRYRFMAKWMFFFD
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  PWIKAGRVTLDNIREQEKFDG--YYHNQFMVVNDCFHRARHLARWIFFFD
Ppa_Pp1s112_159V6.1|Pp1s112_15  PWIKAGRVTLDNIREQEKFDG--YYHNQFLVVNDCFHRARRLAHWLFFFD
Ppa_Pp1s32_144V6.1|Pp1s32_144V  PWIKQGRVTIQNIRQQEIFDG--YYHNQFLIVNDCLFRSRFLANWTFFFD
Ppa_Pp1s198_141V6.1|Pp1s198_14  PWIKQGKVTIQNIRQQEIYDG--YYHNQFLIVNDCLFRSRFMANWTFFFD
Ppa_Pp1s227_17V6.1|Pp1s227_17V  PWIRQGRVTLQNIRQQEIYDG--YYHNQFMVVNDCLFRSRFLANWTFFFD
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  PWIKKGRVSVQNIQQQEIYDG--YYHNQFLVVNDCLFRSRFMANWTFFFD
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  PWIDLGQVLVENLRQAEVYEG--YYHHQFTMLNDCLLRSQTLSNWTFFFD
Ppa_Pp1s98_254V6.1|Pp1s98_254V  PWIDLGRVQVQNLRQAEVYEG--YYHHQFTILNDCLMRSQTLASWTFFFD
Ppa_Pp1s25_268V6.1|Pp1s25_268V  PWIDLGRVSVQNLVMLEVYQG--RSHHQFTMLNDCMFRAQTLSNWTFFFD
Ppa_Pp1s225_53V6.1|Pp1s225_53V  PWVKLGRVTVQNIQQIEMYPA--YYHHQHTVVNDCLMKSQFLANWTFFFD
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  PWVKLGVVIIMNIRQEARFNS--HYHNQFVILNDCLFRAKSMAKWTWFFD
Smo_40688|40688|40688|synon     PLMEKKFLTVIDMRELVLFES--AGDGHQLVINDCLQRSRFLATWAFFWD
C_elegans_NP_504545.2           YYSDNHKITLIDYTLLPFREDGV-WYHGQLATVTDCLLRNTGITKYTFFND
pigeon_ADC84389.1               PWPINSHLKVSTKWHFSMDAKDIGYYGQITALNDCIYRNMQRSKFVVLND
```

FIG. 1 (Cont. 8)

```
Sbi_Sb04g030080.1|Sb04g030080.          VDEYLYLPSG---QKLDEVLGQL----------SGYSQFTIEQNPMSSKLCLQ
Zma_GRMZM2G164912_T01|GRMZM2G1          VDEYLYLPNG---QKLAEVIGQL----------SGYSQFTIEQNPMSSKLCLQ
Sit_Si017001m|Si017001m|Si0170          VDEYIYLPNG---QKLDEVIGKL----------SGYSQFTIEQNEMSSKLCVE
Osa_LOC_Os02g48190.1|LOC_Os02g          VDEYIYLPNG---QTLDQVLGKL----------SGYSQFTIEQNPMSSKLCVQ
Bdi_Bradi3g53170.1|Bradi3g5317          VDEYMYLPNG---RTLDQVLGNL----------SGYTQLTIKQNPMSSKLCLK
Sbi_Sb10g013010.1|Sb10g013010.          VDEYIYLPDG---RTLQEVLGQL----------EAYTQFTIEQNPMSSKLCVE
Zma_GRMZM2G136800_T02|GRMZM2G1          VDEYTYLPDG---RTLQEVLGQL----------EPYTQFTIEQNPMSSKLCVE
Sit_Si006346m|Si006346m|Si0063          ---YIYLPDG---RTLQEVLGQL----------ERYTQFTIEQNPMSSKLCVE
Bdi_Bradi1g42560.1|Bradi1g4256          VDEYMYLPDG---RTLEEVLGQL----------QRYTQFTIEQNPMSSKLCLE
Osa_LOC_Os06g22330.1|LOC_Os06g          VDEYIYLPDG---RALEDVLAQL----------QPYTQFTIEQNPMSSKLCID
Sbi_Sb10g024570.1|Sb10g024570.          VDEYIFLPDG---RSLEDVLAEL----------EPYTQFTIEQNPMSSKLCVD
Zma_GRMZM2G121621_T01|GRMZM2G1          VDEYMFLPDG---RALEDVLAEL----------EPYTQFTIEQNPMSSKLCVD
Sit_Si006264m|Si006264m|Si0062          VDEYIFLPDG---RKLEDVLAEL----------EPYTQFTIEQNPMSSRLCVD
Bdi_Bradi1g35710.1|Bradi1g3571          VDEYIFIPGG---RTLESVMAEL----------EPYTQFTIEQNPMSSRLCAR
Osa_LOC_Os06g41910.1|LOC_Os06g          VDEYIFLPDG---RSLEAVLAEL----------EPYTQFTIEQNPMSSRLCAR
Rco_30170.m014246|30170.m01424          VDEYIYLPDG---STLQSVLAEF----------SDYTQFTIEQNPMSSVLCLN
Ptr_POPTR_0005s28030.1|POPTR_0          VDEYTYLPEG---NTLESVLKDF----------SNYTQFTIEQNPMSSALCFN
Mtr_CU179634_27.1|CU179634_27.          VDEYIYLPDAN--TTLESVLRDF----------SNNTQFTIEQNAMSSILCLN
Aco_AcoGoldSmith_v1.003723m|Ac          VDEYIYLPDG---NTLESVLNEF----------NDYTQFTIEQNPMSSKLCLN
Ath_AT2G33570.1|AT2G33570.1|AT          VDEYIYLPHG---NTLESVLDEF----------SVNTQFTIEQNPMSSVLCIN
Smo_77156|77156|syncnym_                VDEYVHVPGP---STLDSVLNELG---------QKYSQIIFKQTPMGHNVCID
Rco_28563.m000093|28563.m00009          VDEYIYIPFK---NTIKSVLDSL----------SDYNQFTFEQMPMSSKLCLS
Ptr_POPTR_0001s12330.1|POPTR_0          VDEYIHVPQK---NTIKSVLDSL----------SDYTQFTIEQMPMNNKLCLS
Mtr_Medtr8g084210.1|Medtr8g084          VDEYIYVPPK---STIKTVLDSL----------SEYSQFTIEQMAMSVKVCLS
Ath_AT4G20170.1|AT4G20170.1|AT          VDEFLHVPVK---ETISSVMESL----------EEYSQFTIEQMPMSSRICYS
Ath_AT5G44670.1|AT5G44670.1|AT          VDEFIYVPAK---SSISSVMVSL----------EEYSQFTIEQMPMSSQLCYD
Aco_AcoGoldSmith_v1.003774m|Ac          VDEYIYVQPE---STIQSIVNEY----------SGYTQFTIEQMPMSNKLCLA
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.          VDEFIWAPPND--NSLPAILARY----------ENQSQIIIWQKPMSKSLCAQ
Ppa_Pp1s112_159V6.1|Pp1s112_15          VDEYIWAPPND--NSLASILARF----------EDQSQIIIWQKPMSKSLCAQ
Ppa_Pp1s32_144V6.1|Pp1s32_144V          IDEFIFVEPT---TTLSAVLNEN----------PNITQITIEQTPMAKDLCVA
Ppa_Pp1s198_141V6.1|Pp1s198_14          VDEYMYVEPT---TTLSQVLNEN----------PNITQITIEQVPMAKDICVA
Ppa_Pp1s227_17V6.1|Pp1s227_17V          IDEYMYVDPS---TTLSDVLNEK----------PDVSQIVIKQVPIESGLCVA
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.          VDEYLHVPPT---TTLDKVLNDN----------PNVTQITFEQVPISNNLCVA
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.          VDEFLFSSGD---KKP---VDLLAEN--ARNNVTQFLFKTVKMSDGLCLK
Ppa_Pp1s98_254V6.1|Pp1s98_254V          IDEYLYIPEA---HSPRAILNLLAEQ--ARNNVTQILFKNIKISDALCEK
Ppa_Pp1s25_268V6.1|Pp1s25_268V          VDEYLYIPPN---KIL---QQILDES--ERMNFTQIRMKTIKMNDALCTK
Ppa_Pp1s225_53V6.1|Pp1s225_53V          MDEYLFVEPP---HSI---TKLLEEK---TQQDVSVIRYTGKRMSTSMCSL
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.          VDEYIHRPAG---SNLNEVIKRIETENFWGKRLQRVSIKQMSMDYSHCLR
Smo_406883|406883|406883|synon          LDEYIQLVNS---TSMAALLRSY----------ESSPWISLANLVWSRTYCKE
C_elegans_NP_504545.2                   FDEFTVPVIKS--RTLFETISGL----------FEDPTIGSQRTALKYINA
pigeon_ADC84389.1                       ADEIILPLKHLDWKAMMSSLQEQN----------PGAGIFLFENHIFPKTVSTP
```

FIG. 1 (Cont. 9)

```
Sbi_Sb04g030080.1|Sb04g030080.  ----DPSRNYSRE----------WGFEKLVFR--NSIT-KVRRDRKYAIQARNAYS
Zma_GRMZM2G164912_T01|GRMZM2G1  ---DPSRNYSRE----------WGFEKLVFR--NSIT-KVRRDRKYVIQGRNVYS
Sit_Si017001m|Si017001m|Si0170  ----DPSRNYSRE----------WGFEKLVFR---NSIT--KVRRDRKYAIQARNAYS
Osa_LOC_Os02g48190.1|LOC_Os02g  ----DPSKDYSRE----------WGFEKLVFR--NSIT--KVRRDRKYAIQARNAYS
Bdi_Bradi3g53170.1|Bradi3g5317  ---NPSKDSSRE----------WGFEKFVFF--NAVV-KPRRDRKYAVQARNTYS
Sbi_Sb10g013010.1|Sb10g013010.  ----DPTMEYSRK----------WGFEKLVFR--NSIT-GVRRDRKYAIQARNAYS
Zma_GRMZM2G136800_T02|GRMZM2G1  ---DPTMEYSRK----------WGFEKLVFR--NSIT-RVRRDRKYAIRARNAYS
Sit_Si006346m|Si006346m|Si0063  ---DPNKEYSRE----------WGFEKLVFR--NSIT-GVRRDRKYAIQARNAYS
Bdi_Bradi1g42560.1|Bradi1g4256  ---DPAKEYSRK----------WGFEKLVFR--NSIT-GVRRDRKYAVQARNAYS
Osa_LOC_Os06g22330.1|LOC_Os06g  ----DPTEDYSRE----------WGFEKLVFR--NSIT-GVRRDRKYAIQARNAYS
Sbi_Sb10g024570.1|Sb10g024570.  ---DPEADYSNQ----------WGFEKLVFR--NSIT-GVRRDRKYAIQAKNAYA
Zma_GRMZM2G121621_T01|GRMZM2G1  ----DPAADYSNQ----------WGFEKLVFR---NSIT-GVRRDRKYAIQAKNAYA
Sit_Si006264m|Si006264m|Si0062  ----DPEADYSNQ----------WGFEKLVFR---NSIT-GVRRDRKYAIQAKNAYA
Bdi_Bradi1g35710.1|Bradi1g3571  DPGDPKADYSNQ----------WGFEKLVFR--NSIT-GVRRDRKYAIQAKNAYA
Osa_LOC_Os06g41910.1|LOC_Os06g  NPDEPETDYSNE----------WGFEKLVFR--NSIT-GVRRDRKYAIQAKNAYA
Rco_30170.m014246|30170.m01424  ----DSSHDYPRE----------WGFEKLLFR--ESRS-GIRRDRKYAIQAKNAFA
Ptr_POPTR_0005s28030.1|POPTR_0  ----DSTQDYPRQ----------WGFEKLLFR--ESRT-GIRRDRKYAIQAKNAYA
Mtr_CU179634_27.1|CU179634_27.  ---NSSQNYSRL----------WGFEKLLFR--DSRS-NIRRDRKYAIQAKNAYA
Aco_AcoGoldSmith_v1.003723m|Ac  ----ESTHDYSSQ----------WGFEKLVFR---DSRT-KIRRDRKYAIQARNAYS
Ath_AT2G33570.1|AT2G33570.1|AT  ---DSSQDYPRQ----------WGFEKLLFK-|DSRT-KIRRDRKYAIQAKNAFA
Smo_77156|77156|77156|synonym_  ---SNRTNLSRQ----------WGFEKLVFI--NVQK-SVRLDRKYVLRAPLADS
Rco_28563.m000093|28563.m00009  --AD-YGRYYRK----------WGFEKLVYR--DIKR-GIRRDRKYAIQPRNVFA
Ptr_POPTR_0001s12330.1|POPTR_0  --AD-YGRYYRK----------WGFEKLVYK--DVKR-GIRRDRKYAIQPRNVFA
Mtr_Medtr8g084210.1|Medtr8g084  --HD-YGKTYRK----------WGFEKLVYR--DAIT-GIRRDRKYAVQPRSLFA
Ath_AT4G20170.1|AT4G20170.1|AT  --GDGPARTYRK----------WGIEKLAYR--DVKK-VPRRDRKYAVQPENVFA
Ath_AT5G44670.1|AT5G44670.1|AT  --GDGPARTYRK----------WGFEKLAYR--DVKK-VPRRDRKYAVQPRNVFA
Aco_AcoGoldSmith_v1.003774m|Ac  --SD-AGNTERE----------WGFEKLVYR--NVQR-GIRRDRKYAIQPRNTFA
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  EQGEAGNDSSFSRLRSKWTFEKLVFK--SIKW-RESHDCKYAIQGQKAMG
Ppa_Pp1s112_159V6.1|Pp1s112_15  DQGAASNDSSFA---SKWTFEKLVFK--SIKW-RESHDCKYAIRGLKAMA
Ppa_Pp1s32_144V6.1|Pp1s32_144V  D-NTTSGEHESR----------WGFEKLIYR--KVLKRGIRYDRKFVVQARHAEA
Ppa_Pp1s198_141V6.1|Pp1s198_14  D-NATRGDLKSS----------WGFEKLVYR--KVLKPGLRYDRKYAVQARHAEA
Ppa_Pp1s227_17V6.1|Pp1s227_17V  D-NTR--DQHER----------WLTERLIYR--KVLKHGIHYDRKYAVQARHAEA
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  D-NTTTDGHARK----------WAIEKLMYR--KVLRSGVRYDRKYAVQARYAAA
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  ENKRESDSHFMAKISRKWAFEKLVYAKCDYGWTT-VESRKYAVQFQMVWA
Ppa_Pp1s98_254V6.1|Pp1s98_254V  SNKTTTEAQSMAKQSRKWAFEKLVYADCNYGWEP--VESRKYALQPQLVWA
Ppa_Pp1s25_268V6.1|Pp1s25_268V  PDPPLDDAARDAVNARKWAIEKLVYANCDYGWEP-VESRKYALKPRAAWA
Ppa_Pp1s225_53V6.1|Pp1s225_53V  PKANVDNATRDAINSKLWGFEKFLLH----WHWEHGPRDPKYAVQARHMWA
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  ----GSQPNSNE----------WGISKMVYR---KTDVGEYPDRKNFVLSDSCLA
Smo_406883|406883|406883|synon  T--SAEDEWAVERMRFRLRYPTCGDEQRAETCHGVEGHRKWIANPRKVLV
C_elegans_NP_504545.2           -------KIKSAP----------YSLKNIVSE-----KRIETRFTKCVVRPEMVFE
pigeon_ADC84389.1               VFNISSWNRVPG----------VNILQHVHR--EPDRKEVFNPKKMIIDPRQVVQ
                                                                  *
```

FIG. 1 (Cont. 10)

```
Sbi_Sb04g030080.1|Sb04g030080.      AGVHMSQNLY----------GRTTHKT-ESLIRYYHYHNSINVMG-EPCREFV
Zma_GRMZM2G164912_T01|GRMZM2G1      AGVHMSQNLY----------GRTTHKT--EGLIRYYHYHDSINVMG--EPCREFV
Sit_Si017001m|Si017001m|Si0170      TGVHMSQNVR----------GRTTHKT-ESLIRYYHYHNSINVMG-EPCREFV
Osa_LOC_Os02g48190.1|LOC_Os02g      AGVHMSQNVY----------GRTTHKT-ESLIRYYHYHNSINVMG-EPCREFV
Bdi_Bradi3g53170.1|Bradi3g5317      AGVHLSQNLY----------GRSTHDT-ENVIRYYHFHNSINVLG-EPCKKFV
Sbi_Sb10g013010.1|Sb10g013010.      TGVHMSQNVI----------GRTTHKT-ESLIRYYHYHNSINVMG-EPCREFV
Zma_GRMZM2G136800_T02|GRMZM2G1      TGVHMSQNVI----------GRTTHKT--ESLILYYHYHNSINVMG--EPCREFV
Sit_Si006346m|Si006346m|Si0063      TGVHMSQNVI----------GRTSHKT-ERLIRYYHYHNSININMG-EPCREFV
Bdi_Bradi1g42560.1|Bradi1g4256      AGVHMSQNVY----------GRTTHRT-ESLIRYYHYHNTINVMG-EPCREFV
Osa_LOC_Os06g22330.1|LOC_Os06g      TGVHMSQNVY----------GRTTHKT-ESLIRYYHYHNSINVMG-EPCRKFV
Sbi_Sb10g024570.1|Sb10g024570.      TGVHMSENVI----------GNTTHKT--EHLIRYYHYHNTINVLG--EVCREFV
Zma_GRMZM2G121621_T01|GRMZM2G1      TGVHMSENVI----------GNTTHKT-EHLIRYYHYHNTINVLG-EVCREFV
Sit_Si006264m|Si006264m|Si0062      TGVHMSENVI----------GNTTHKT--EHLIRYYHYHNTINVLD--EVCREFV
Bdi_Bradi1g35710.1|Bradi1g3571      TGVHMSENVI----------GNTTHKT-EHLIRYYHYHNTINVLG-EVCREFV
Osa_LOC_Os06g41910.1|LOC_Os06g      TGVHMSENVI----------GNTTHKT-EHLIRYYHYHNTINVIG-EVCREFV
Rco_30170.m014246|30170.m01424      TGVHMSENVV----------GKTLHKT--ETKIRYYHYHNSITVQG--ELCRQLL
Ptr_POPTR_0005s28030.1|POPTR_0      TGVHMSENVI----------GKTLHQT-ETKIRYYHYHNSIQVPG-ELCREFL
Mtr_CU179634_27.1|CU179634_27.      TGVHMSENVV----------GGTLHQT--ETKIRYYHYHNSITVHE--ELCREFL
Aco_AcoGoldSmith_v1.003723m|Ac      TGVHMSENVY----------GKTLHKT-ETKIRYYHYHNTITVDG-EPCREFL
Ath_AT2G33570.1|AT2G33570.1|AT      TGVHMSENIV----------GKTLHKT--ETKIRYYHYHNTITVHE--ELCREML
Smo_77156|77156|syntenym_           AGVHLSENVR----------GKSLYS---RGQISYFHFHNTITEHQ--ENCREFV
Rco_28563.m000093|28563.m00009      TGVHMSQNLA----------GKTNHKT-ESKIKYFHYHGTIAQRR-EPCRNLL
Ptr_POPTR_0001s12330.1|POPTR_0      TGVHMSQNVA----------GKTTHKT-EGMIRYFHYHGTAAQRR-EPCRNLL
Mtr_Medtr8g084210.1|Medtr8g084      NGIHMSENLD----------GKTTHNT-EGRIKYFHYHGAISQRR-ETCKLLV
Ath_AT4G20170.1|AT4G20170.1|AT      TGVHMSQNLQ----------GKTYHKA-ESKIRYFHYHGSISQRR-EPCRQLF
Ath_AT5G44670.1|AT5G44670.1|AT      TGVHMSQHLQ----------GKTYHRA-EGKIRYFHYHGSISQRR-EPCRHLY
Aco_AcoGoldSmith_v1.003774m|Ac      TGVHMSQNVA----------GRTLHRI-DGRIKYFHYHGTISERH-ELCRQLV
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.      TGIHRSGHLI----------GGNATTY-ADPLNYYHYHNTINKRD-EVCEDFI
Ppa_Pp1s112_159V6.1|Pp1s112_15      TGIHRSGYLI----------GSNATTY-ADPLNYYHYHNTINKRE-EVCEQFI
Ppa_Pp1s32_144V6.1|Pp1s32_144V      TGIHMSMDTRH---------GQNLYPK-GDKIRYYHYHGTINKRQ-EVCTRFV
Ppa_Pp1s198_141V6.1|Pp1s198_14      AGIHMSMNMLK---------GQNLYPK-GDKIRYYHYHGTINKRG-EVCKKFV
Ppa_Pp1s227_17V6.1|Pp1s227_17V      TGIHMSMNLRR---------GKTRVLQ-DERLRYYHYHGTITNRG-NLCNKFV
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.      TGVHMSMNMRK---------GKNMYPK-GYKIRYYHYHDTINKRE-ELCQAFV
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.      TGVHYSEHMKQ---------G-KRIDMDPDVLRSYHYHNTVNGRT-ELCKEFP
Ppa_Pp1s98_254V6.1|Pp1s98_254V      TGVHYSEHLKQ---------G-HYITMDPTVLRSYHYHNTLNGRW-ETCKDFP
Ppa_Pp1s25_268V6.1|Pp1s25_268V      TGVHYCEYLKE---------G-GLMDADENLLRLYHYHGTTPGPK-EVCRDFT
Ppa_Pp1s225_53V6.1|Pp1s225_53V      VEIHNAVLARR---------GQNRLLQYRDHIWFHHFHGFIINRA-EGCRDFV
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.      TGPHEPMAMSMPPGANQTHHVTKWEGELLQYYHFHGAIGKAEGEICVDLL
Smo_406883|406883|406883|synon      ASIHRTVEPEWN---------GEILRAS---IARLNHFHGLAVEGSANSSCWIV
C_elegans_NP_504545.2               QGIHHTSRVIQD---------NYKTVSHGGSLLRVYHYKDKKYCCE---DESLL
pigeon_ADC84389.1                   TSVHSVLRAYG----------NSVNVPADVALVYHCRVPLQEEL--PRESLI
                                                          *  ;                    ;
```

FIG. 1 (Cont. 11)

```
Sbi_Sb04g030080.1|Sb04g030080.  PMPAN--GSKTMFEGVPYYVYDDNMKRLADEIKRFEKVT----------
Zma_GRMZM2G164912_T01|GRMZM2G1  PMPAN--GSKTMFEGVPYYVYDDNMKRLAGEIKRFEKVT----------
Sit_Si017001m|Si017001m|Si0170  PMPVN--GSKIMFEKTPFVYDDSMKRVAGEIKRFEKET----------
Osa_LOC_Os02g48190.1|LOC_Os02g  PVPVN--GSKLMFEGIPYYVYDDNMKRLAGQIKRFEKEA----------
Bdi_Bradi3g53170.1|Bradi3g5317  AKPAN--GSQIMFEGAPFVYDYNLKRLAGEIKRFEKET----------
Sbi_Sb10g013010.1|Sb10g013010.  PTPTN--GSKVMFEGIPYYVYDDNMKRITGEIKRFEEET----------
Zma_GRMZM2G136800_T02|GRMZM2G1  PRPNN--GSKVMFEGIPYYVYDDNMERVAGEIKRFEEET----------
Sit_Si006346m|Si006346m|Si0063  QKPTN--GSKVMFEGIPYYVYDDNMKRLTVEIKRFEEET----------
Bdi_Bradi1g42560.1|Bradi1g4256  SKPTN--GSKVTFEGIPYYVYDDNMKRLAGDIKHFEEATGTVLGGSSGLAG
Osa_LOC_Os06g22330.1|LOC_Os06g  PKPAN--GSKVMFEGIPYYVYDDNMKRLAGEIRRFEKQT----------
Sbi_Sb10g024570.1|Sb10g024570.  SVPPK--GGLTWSEKTPWYYDNSMKRVADAVREFEERET----------
Zma_GRMZM2G136800_T02|GRMZM2G1  SVPPKKGGGLTWSEKTPWYYDGSMKRVAGAVREFEERET----------
Sit_Si006346m|Si006346m|Si0063  PIPPK--GGLTWSEKTPWYYDDSMKRVANAVREFERET----------
Bdi_Bradi1g42560.1|Bradi1g4256  PVPPK--GGVTWSEKTPWYYDDSMKRVANAVREFERNT----------
Osa_LOC_Os06g41910.1|LOC_Os06g  PIPPN--GGLIWSEKTPWYYDDSMKRIADTVREFERKT----------
Rco_30170.m014246|30170.m01424  PASAK--HNVTWYNKLPYVYDDNMKKLVTTIRDFERNT----------
Ptr_POPTR_0005s28030.1|POPTR_0  PLSAK--NNVTWYNGLPYVYDDNMKKLASTIKDFERNT----------
Mtr_CU179634_27.1|CU179634_27.  PMSAK--HNITWFDKVPYEYDDSMKKLTQTIKDFERNT----------
Aco_AcoGoldSmith_v1.003723m|Ac  PDSAK--RTVTMSDNKPTVYDDKMKKLADTIKQFEKET----------
Ath_AT2G33570.1|AT2G33570.1|AT  PNSAK--KKVTLYNKLPYVYDDNMKKLVKTIKEFEQKK----------
Smo_77156|77156|synonym_        KVPKNR-TSKVWLNKVPYLYDDKLSTLADEVKQFELQT----------
Rco_28563.m000093|28563.m00009  N----------VTEINFENNPYVLDTTMRDIAWSVKKFEQKM----------
Ptr_POPTR_0001s12330.1|POPTR_0  N----------VTEINFENNPYVLDTTMRDLAWSVKKFENQM----------
Mtr_Medtr8g084210.1|Medtr8g084  N----------STKITYEKTPYVMDTTLRDIAGSIKKFELKM----------
Ath_AT4G20170.1|AT4G20170.1|AT  N----------DSRVVFENTPYVLDTTICDVGLAVRTFELRT----------
Ath_AT5G44670.1|AT5G44670.1|AT  N----------GTRIVHENNPYVLDTTMRDIGLAVKTFEIRT----------
Aco_AcoGoldSmith_v1.003774m|Ac  N----------ESQIYIEKKPYILDTTLRDAAAAVKSFEIKS----------
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  DKSVV----YVNPRNRTDIYTHDDGLAVLADHIKEFERDV----------
Ppa_Pp1s112_159V6.1|Pp1s112_15  DSNSV----YVNPRNRTDKFAYDDGLAVLADHIKEFELDV----------
Ppa_Pp1s32_144V6.1|Pp1s32_144V  DAGNK--TAVQRHEKHYHRLDESMAKLAKDVKLFELNT----------
Ppa_Pp1s198_141V6.1|Pp1s198_14  DAGNK--TALQTHDNHYHRLEETIAKMAKDVKLYELNT----------
Ppa_Pp1s227_17V6.1|Pp1s227_17V  DSKNK--TAVKSVQ---LRLDETMMKMAKDVKSYELKM----------
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  SFQNK--TSLQRHQRQNYRLDETLAMTTQLIKDYERRT----------
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  KNPNG---EPNSLKVKNCTIDKSMADLGPAVRSYEFAM----------
Ppa_Pp1s98_254V6.1|Pp1s98_254V  QHPNG----GPNSVNVTNCTTDHAMLSFAPAVKSYELAT----------
Ppa_Pp1s25_268V6.1|Pp1s25_268V  NGTNG---EPNYLGVKNCTVDRTMTTLANPVKQFELQT----------
Ppa_Pp1s225_53V6.1|Pp1s225_53V  EASEN----KAIAVGKSQYSGDYTIARLAKDIKQFEQAM----------
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  DPDVN----VTSFNNVTHQFDDALKSLVDDVMAFEHRM----------
Smo_406883|406883|406883|synon  KKLDQ--VNESSLVDGWWLKDESFAVAAMEQHRKGL----------
C_elegans_NP_504545.2           KKBHG----------DQLREKFDSVVGLLDL----------
pigeon_ADC84389.1               R----------DTALWRYNSSLITNVNKVLHQTVL----------
                                                        :
```

FIG. 1 (Cont. 12)

```
Sbi_Sb04g030080.1|Sb04g030080.  ----------------------------------------LGSVQT----------
Zma_GRMZM2G164912_T01|GRMZM2G1  ----------------------------------------LGSART----------
Sit_Si017001m|Si017001m|Si0170  ----------------------------------------IGSVQT----------
Osa_LOC_Os02g48190.1|LOC_Os02g  ----------------------------------------IGSAHT----------
Bdi_Bradi3g53170.1|Bradi3g5317  ----------------------------------------IGSSHT----------
Sbi_Sb10g013010.1|Sb10g013010.  ----------------------------------------LGTIRR----------
Zma_GRMZM2G136800_T02|GRMZM2G1  ----------------------------------------LGTVRT----------
Sit_Si006346m|Si006346m|Si0063  ----------------------------------------IGAIHT----------
Bdi_Bradi1g42560.1|Bradi1g4256  GGLLGLVLGGGRLARSECHWFAAAVALITPLVFLTIGALSALQETRRRWC
Osa_LOC_Os06g22330.1|LOC_Os06g  ----------------------------------------IGDVHT----------
Sbi_Sb10g024570.1|Sb10g024570.  ----------------------------------------IGDVRL----------
Zma_GRMZM2G121621_T01|GRMZM2G1  ----------------------------------------IGDVRL----------
Sit_Si006264m|Si006264m|Si0062  ----------------------------------------IGDVRL----------
Bdi_Bradi1g35710.1|Bradi1g3571  ----------------------------------------IGNVRV----------
Osa_LOC_Os06g41910.1|LOC_Os06g  ----------------------------------------IGDVRV----------
Rco_30170.m014246|30170.m01424  ----------------------------------------IGNVRQYS--------
Ptr_POPTR_0005s28030.1|POPTR_0  ----------------------------------------IGNVQAYS--------
Mtr_CU179634_27.1|CU179634_27.  ----------------------------------------INININAAPDHTNNS--
Aco_AcoGoldSmith_v1.003723m|Ac  ----------------------------------------LGEAVHV---------
Ath_AT2G33570.1|AT2G33570.1|AT  ----------------------------------------LGTDVKNFS-------
Smo_77156|77156|77156|synonym_  ----------------------------------------IGPVTL----------
Rco_28563.m000093|28563.m00009  ----------------------------------------IGSRLQSTRQ------
Ptr_POPTR_0001s12330.1|POPTR_0  ----------------------------------------IGPKLKNTRQ------
Mtr_Medtr8g084210.1|Medtr8g084  ----------------------------------------IGTSVVQSDIAL----
Ath_AT4G20170.1|AT4G20170.1|AT  ----------------------------------------IGDRLLRTRQ------
Ath_AT5G44670.1|AT5G44670.1|AT  ----------------------------------------IGDRLLRTRQ------
Aco_AcoGoldSmith_v1.003774m|Ac  ----------------------------------------IGSQLQRTRQ------
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  ----------------------------------------IGPQSFIL--------
Ppa_Pp1s112_159V6.1|Pp1s112_15  ----------------------------------------IGPQPFIL--------
Ppa_Pp1s32_144V6.1|Pp1s32_144V  ----------------------------------------VGAQPFIV--------
Ppa_Pp1s198_141V6.1|Pp1s198_14  ----------------------------------------VGTQPFIV--------
Ppa_Pp1s227_17V6.1|Pp1s227_17V  ----------------------------------------IGKQPFVL--------
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  ----------------------------------------IGSLPFIV--------
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  ----------------------------------------VGEQPFIL--------
Ppa_Pp1s98_254V6.1|Pp1s98_254V  ----------------------------------------IGEQPFLL--------
Ppa_Pp1s25_268V6.1|Pp1s25_268V  ----------------------------------------IGKQSTLE--------
Ppa_Pp1s225_53V6.1|Pp1s225_53V  ----------------------------------------VGKLPFIL--------
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  ----------------------------------------INQATHTS--------
Smo_406883|406883|406883|synon  ----------------------------------------------------------
C_elegans_NP_504545.2           ----------------------------------------------------------
pigeon_ADC84389.1               ----------------------------------------------------------
```

FIG. 1 (Cont. 13)

FIG. 1 (Cont. 14)

```
Sbi_Sb04g030080.1|Sb04g030080.  ----------
Zma_GRMZM2G164912_T01|GRMZM2G1  ----------
Sit_Si017001m|Si017001m|Si0170  ----------
Osa_LOC_Os02g48190.1|LOC_Os02g  ----------
Bdi_Bradi3g53170.1|Bradi3g5317  ----------
Sbi_Sb10g013010.1|Sb10g013010.  ----------
Zma_GRMZM2G136800_T02|GRMZM2G1  ----------
Sit_Si006346m|Si006346m|Si0063  ----------
Bdi_Bradi1g42560.1|Bradi1g4256  AFGTVGNSWSACPDHGGRGLSPSMRRFRDRPPASVAVAYQQEKVTLGVLI
Osa_LOC_Os06g22330.1|LOC_Os06g  ----------
Sbi_Sb10g024570.1|Sb10g024570.  ----------
Zma_GRMZM2G121621_T01|GRMZM2G1  ----------
Sit_Si006264m|Si006264m|Si0062  ----------
Bdi_Bradi1g35710.1|Bradi1g3571  ----------
Osa_LOC_Os06g41910.1|LOC_Os06g  ----------
Rco_30170.m014246|30170.m01424  ----------
Ptr_POPTR_0005s28030.1|POPTR_0  ----------
Mtr_CU179634_27.1|CU179634_27.  ----------
Aco_AcoGoldSmith_v1.003723m|Ac  ----------
Ath_AT2G33570.1|AT2G33570.1|AT  ----------
Smo_77156|77156|77156|synonym_  ----------
Rco_28563.m000093|28563.m00009  ----------
Ptr_POPTR_0001s12330.1|POPTR_0  ----------
Mtr_Medtr8g084210.1|Medtr8g084  ----------
Ath_AT4G20170.1|AT4G20170.1|AT  ----------
Ath_AT5G44670.1|AT5G44670.1|AT  ----------
Aco_AcoGoldSmith_v1.003774m|Ac  ----------
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  ----------
Ppa_Pp1s112_159V6.1|Pp1s112_15  ----------
Ppa_Pp1s32_144V6.1|Pp1s32_144V  ----------
Ppa_Pp1s198_141V6.1|Pp1s198_14  ----------
Ppa_Pp1s227_17V6.1|Pp1s227_17V  ----------
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  ----------
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  ----------
Ppa_Pp1s98_254V6.1|Pp1s98_254V  ----------
Ppa_Pp1s25_268V6.1|Pp1s25_268V  ----------
Ppa_Pp1s225_53V6.1|Pp1s225_53V  ----------
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  ----------
Smo_406883|406883|406883|synon  ----------
C_elegans_NP_504545.2           ----------
pigeon_ADC84389.1               ----------
```

FIG. 1 (Cont. 15)

```
Sbi_Sb04g030080.1|Sb04g030080.  ----------
Zma_GRMZM2G164912_T01|GRMZM2G1  ----------
Sit_Si017001m|Si017001m|Si0170  ----------
Osa_LOC_Os02g48190.1|LOC_Os02g  ----------
Bdi_Bradi3g53170.1|Bradi3g5317  ----------
Sbi_Sb10g013010.1|Sb10g013010.  ----------
Zma_GRMZM2G136800_T02|GRMZM2G1  ----------
Sit_Si006346m|Si006346m|Si0063  ----------
Bdi_Bradi1g42560.1|Bradi1g4256  SSLGEEEKQEMIEKSLSVSLEKKQETSSGCCVTMQL
Osa_LOC_Os06g22330.1|LOC_Os06g  ----------
Sbi_Sb10g024570.1|Sb10g024570.  ----------
Zma_GRMZM2G121621_T01|GRMZM2G1  ----------
Sit_Si006264m|Si006264m|Si0062  ----------
Bdi_Bradi1g35710.1|Bradi1g3571  ----------
Osa_LOC_Os06g41910.1|LOC_Os06g  ----------
Rco_30170.m014246|30170.m01424  ----------
Ptr_POPTR_0005s28030.1|POPTR_0  ----------
Mtr_CU179634_27.1|CU179634_27.  ----------
Aco_AcoGoldSmith_v1.003723m|Ac  ----------
Ath_AT2G33570.1|AT2G33570.1|AT  ----------
Smo_77156|77156|synonym_        ----------
Rco_28563.m000093|28563.m00009  ----------
Ptr_POPTR_0001s12330.1|POPTR_0  ----------
Mtr_Medtr8g084210.1|Medtr8g084  ----------
Ath_AT4G20170.1|AT4G20170.1|AT  ----------
Ath_AT5G44670.1|AT5G44670.1|AT  ----------
Aco_AcoGoldSmith_v1.003774m|Ac  ----------
Ppa_Pp1s35_57V6.1|Pp1s35_57V6.  ----------
Ppa_Pp1s112_159V6.1|Pp1s112_15  ----------
Ppa_Pp1s32_144V6.1|Pp1s32_144V  ----------
Ppa_Pp1s198_141V6.1|Pp1s198_14  ----------
Ppa_Pp1s227_17V6.1|Pp1s227_17V  ----------
Ppa_Pp1s35_63V6.1|Pp1s35_63V6.  ----------
Ppa_Pp1s237_6V6.1|Pp1s237_6V6.  ----------
Ppa_Pp1s98_254V6.1|Pp1s98_254V  ----------
Ppa_Pp1s25_268V6.1|Pp1s25_268V  ----------
Ppa_Pp1s225_53V6.1|Pp1s225_53V  ----------
Ppa_Pp1s556_1V6.1|Pp1s556_1V6.  ----------
Smo_406883|406883|406883|synon  ----------
C_elegans_NP_504545.2           ----------
pigeon_ADC84389.1               ----------
```

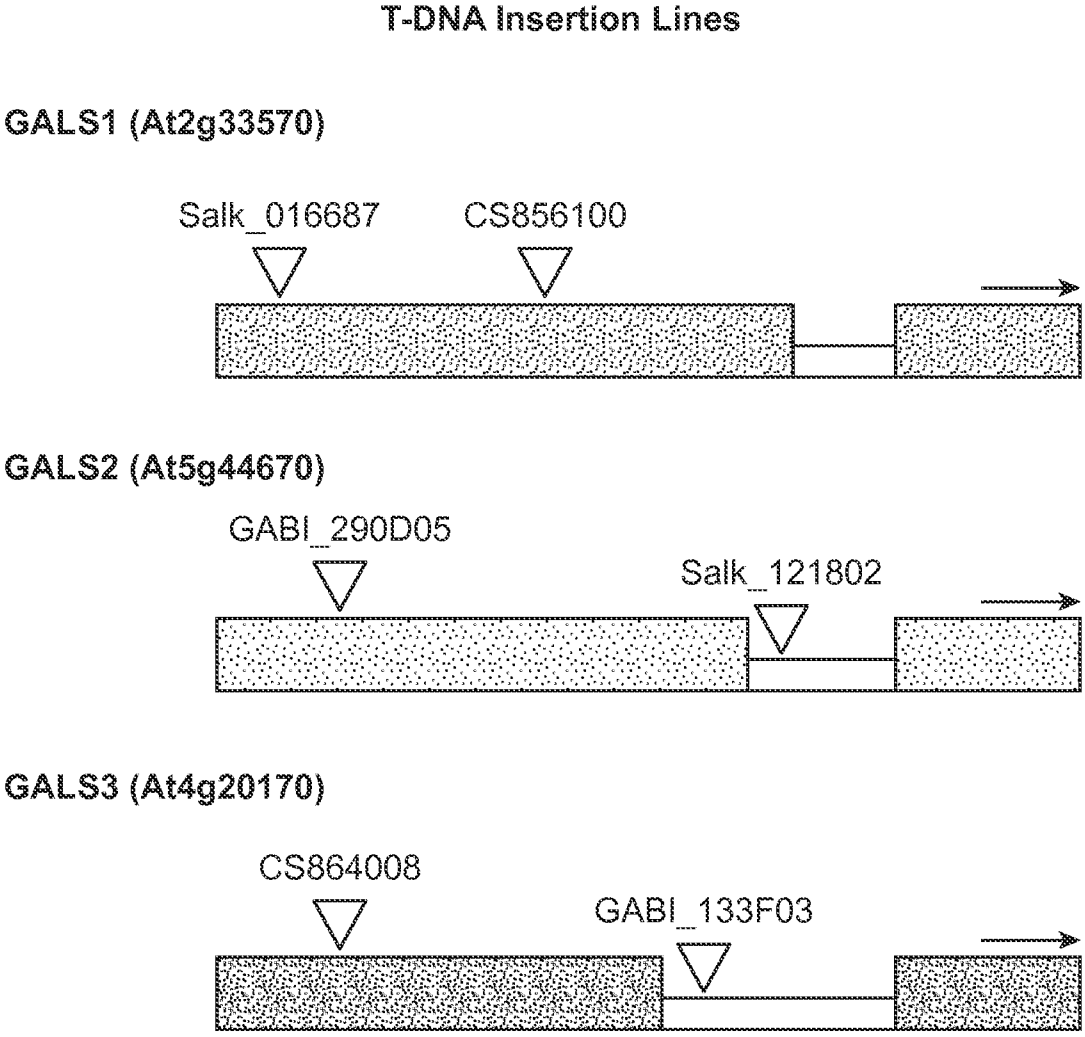
T-DNA Insertion Lines
GALS1 (At2g33570)
Salk_016687
CS856100
GALS2 (At5g44670)
GABI_290D05
Salk_121802
GALS3 (At4g20170)
CS864008
GABI_133F03

FIG. 2

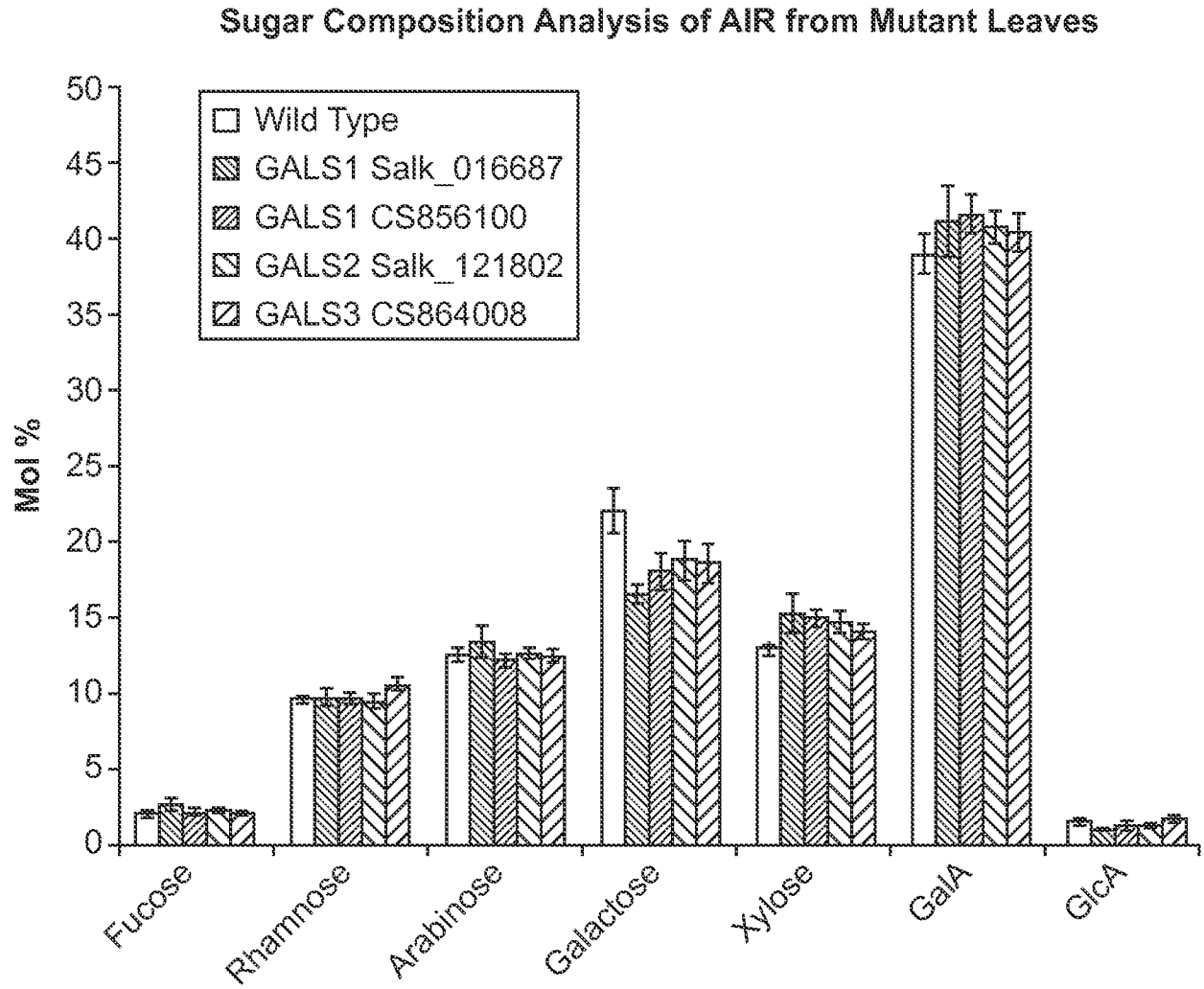
Sugar Composition Analysis of AIR from Mutant Leaves
Wild Type
GALS1 Salk_016687
GALS1 CS856100
GALS2 Salk_121802
GALS3 CS864008
Mol %
50
45
40
35
30
25
20
15
10
5
0
Fucose
Rhamnose
Arabinose
Galactose
Xylose
GalA
GlcA
All mutants shows a specific and significant deficiency in total cell wall galactose content.

FIG. 3

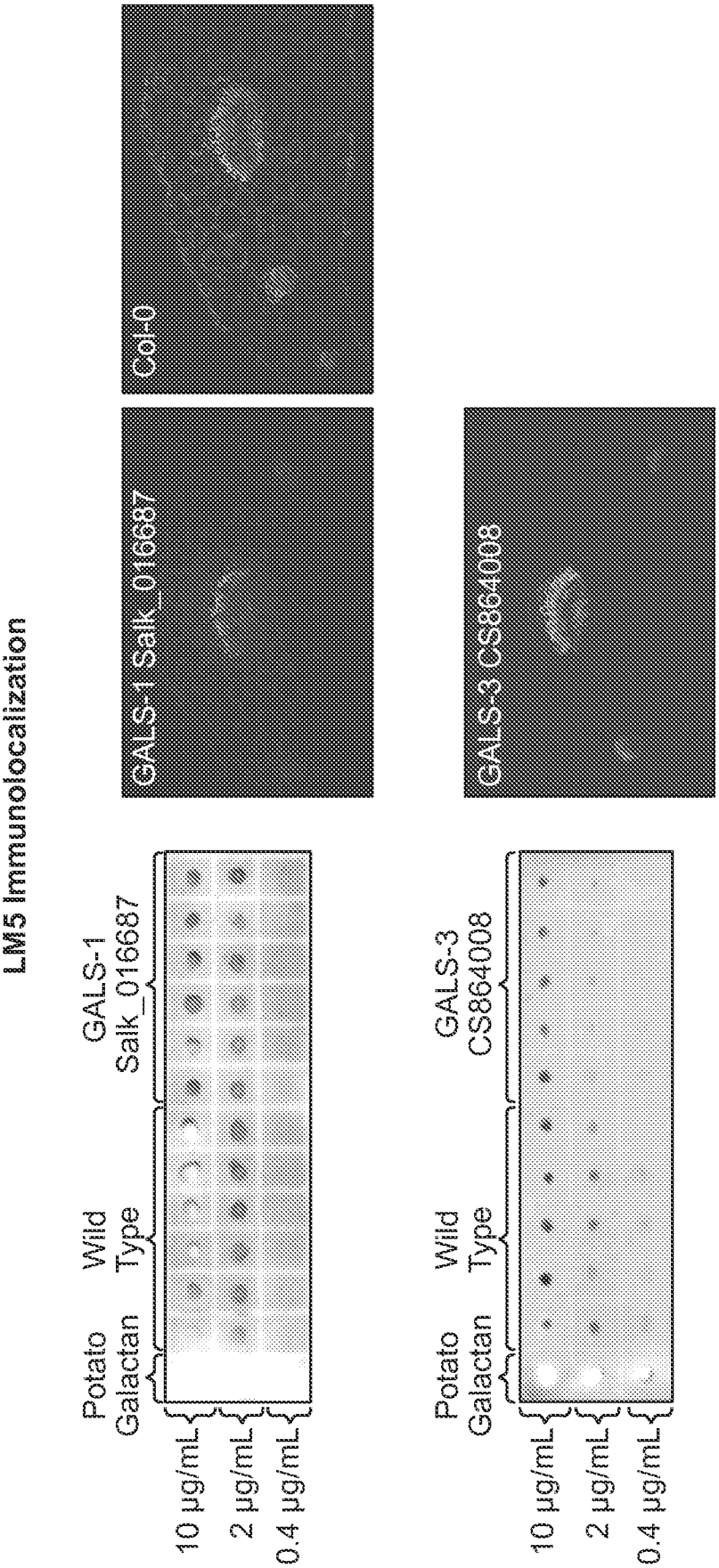
LM5 Immunolocalization
Col-0
GALS-1 Salk_016687
GALS-3 CS864008
Potato Galactan
Wild Type
GALS-1 Salk_016687
GALS-3 CS864008
10 µg/mL
2 µg/mL
0.4 µg/mL

FIG. 4

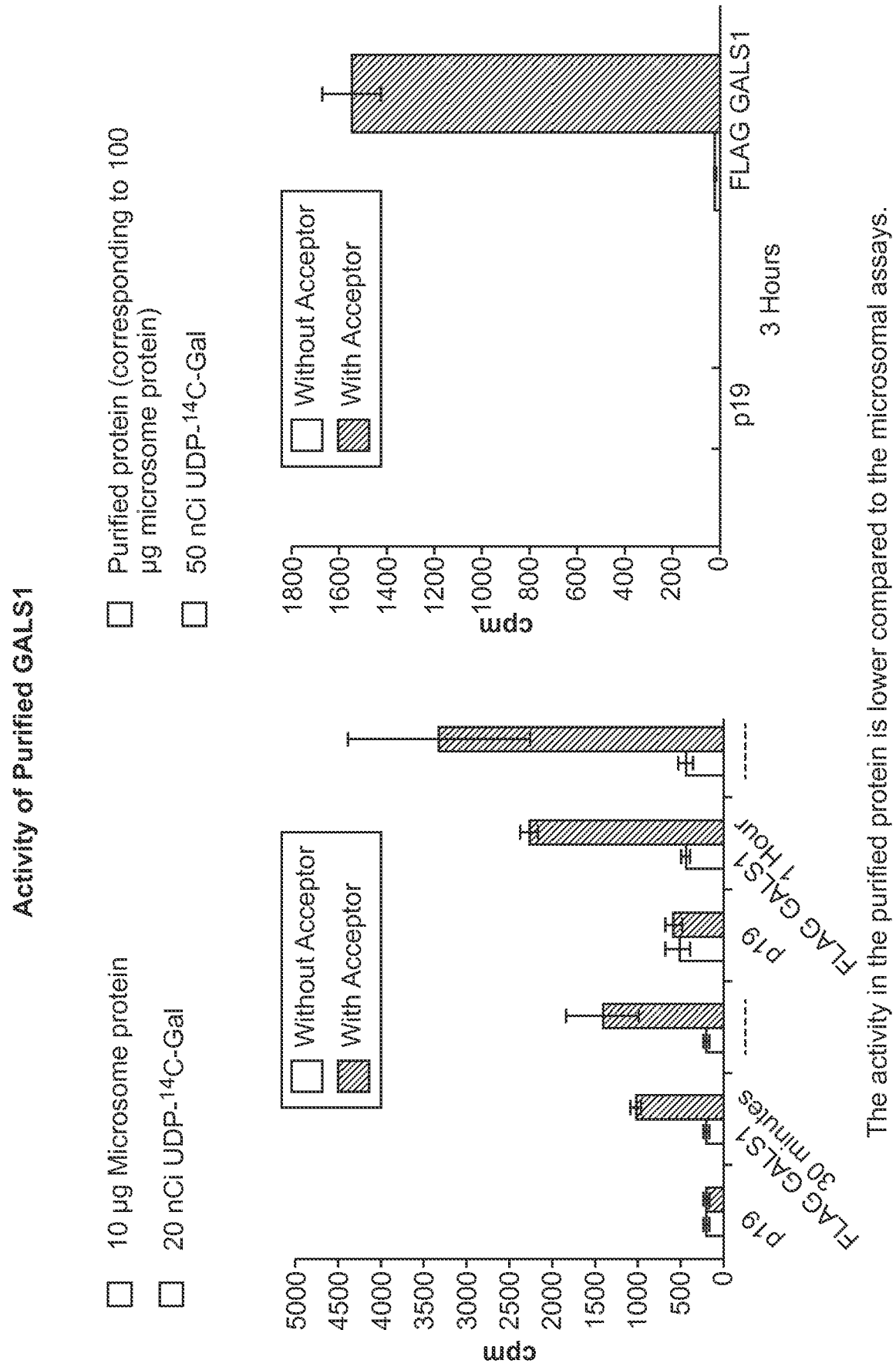
Activity of Purified GALS1
10 µg Microsome protein
20 nCi UDP-14C-Gal
Purified protein (corresponding to 100 µg microsome protein)
50 nCi UDP-14C-Gal
Without Acceptor
With Acceptor
cpm
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
p19
FLAG GALS1
30 minutes
1 Hour
3 Hours
1800
1600
1400
1200
1000
800
600
400
200
0
The activity in the purified protein is lower compared to the microsomal assays.

FIG. 5

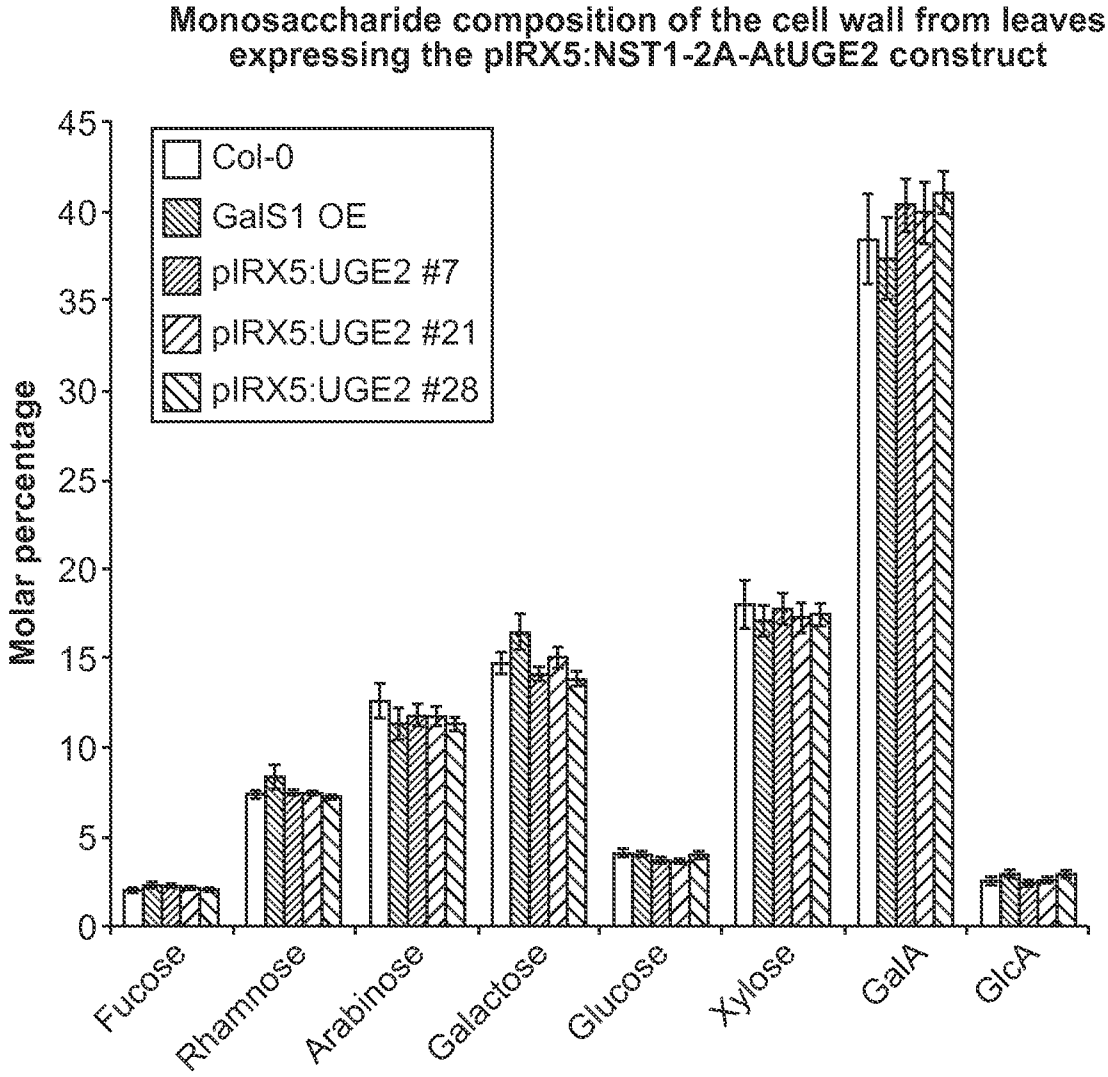
Monosaccharide composition of the cell wall from leaves expressing the pIRX5:NST1-2A-AtUGE2 construct
Molar percentage
0
5
10
15
20
25
30
35
40
45
Col-0
GalS1 OE
pIRX5:UGE2 #7
pIRX5:UGE2 #21
pIRX5:UGE2 #28
Fucose
Rhamnose
Arabinose
Galactose
Glucose
Xylose
GalA
GlcA

FIG. 6

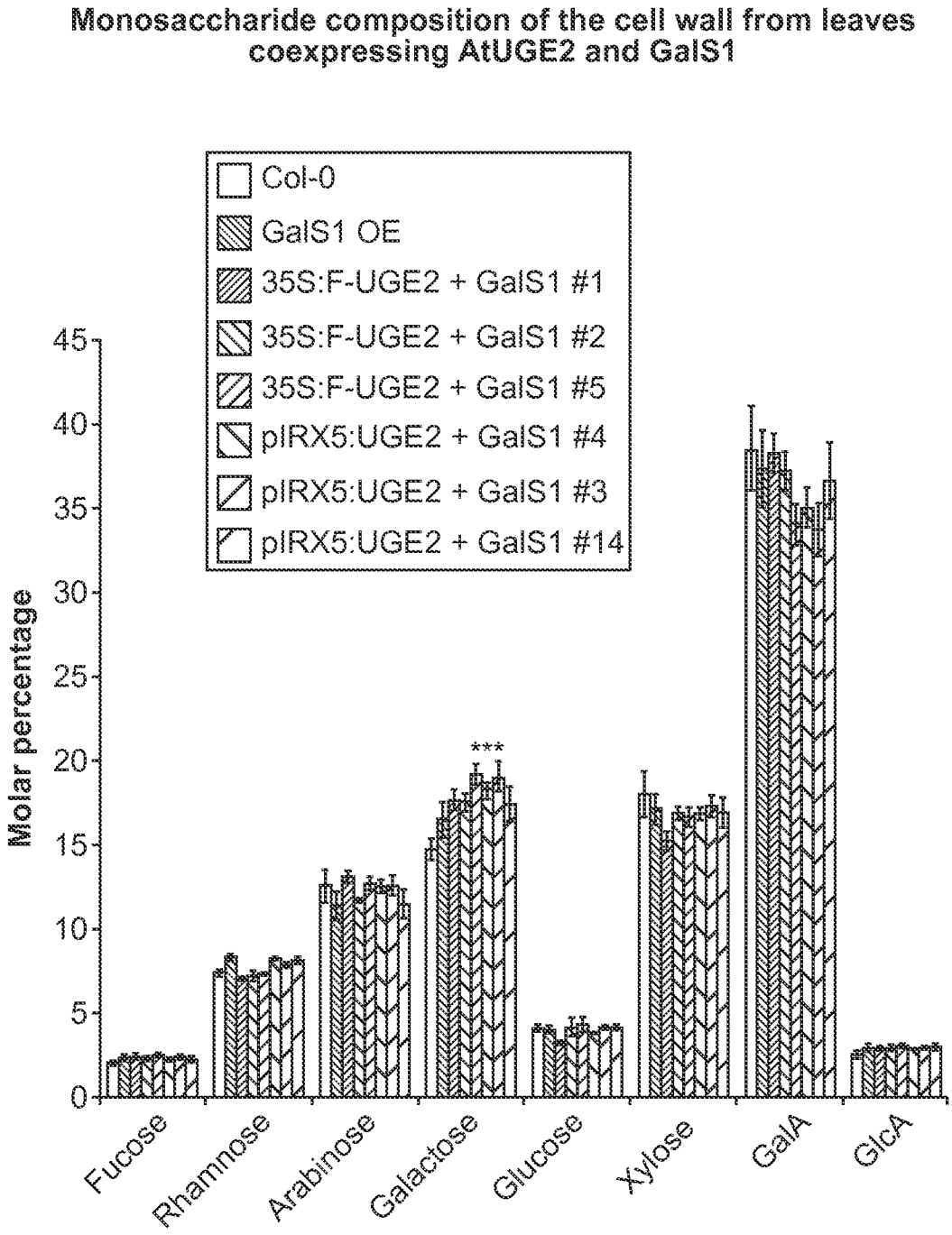
Monosaccharide composition of the cell wall from leaves coexpressing AtUGE2 and GalS1
Col-0
GalS1 OE
35S:F-UGE2 + GalS1 #1
35S:F-UGE2 + GalS1 #2
35S:F-UGE2 + GalS1 #5
pIRX5:UGE2 + GalS1 #4
pIRX5:UGE2 + GalS1 #3
pIRX5:UGE2 + GalS1 #14
Molar percentage
0
5
10
15
20
25
30
35
40
45
***
Fucose
Rhamnose
Arabinose
Galactose
Glucose
Xylose
GalA
GlcA

FIG. 7

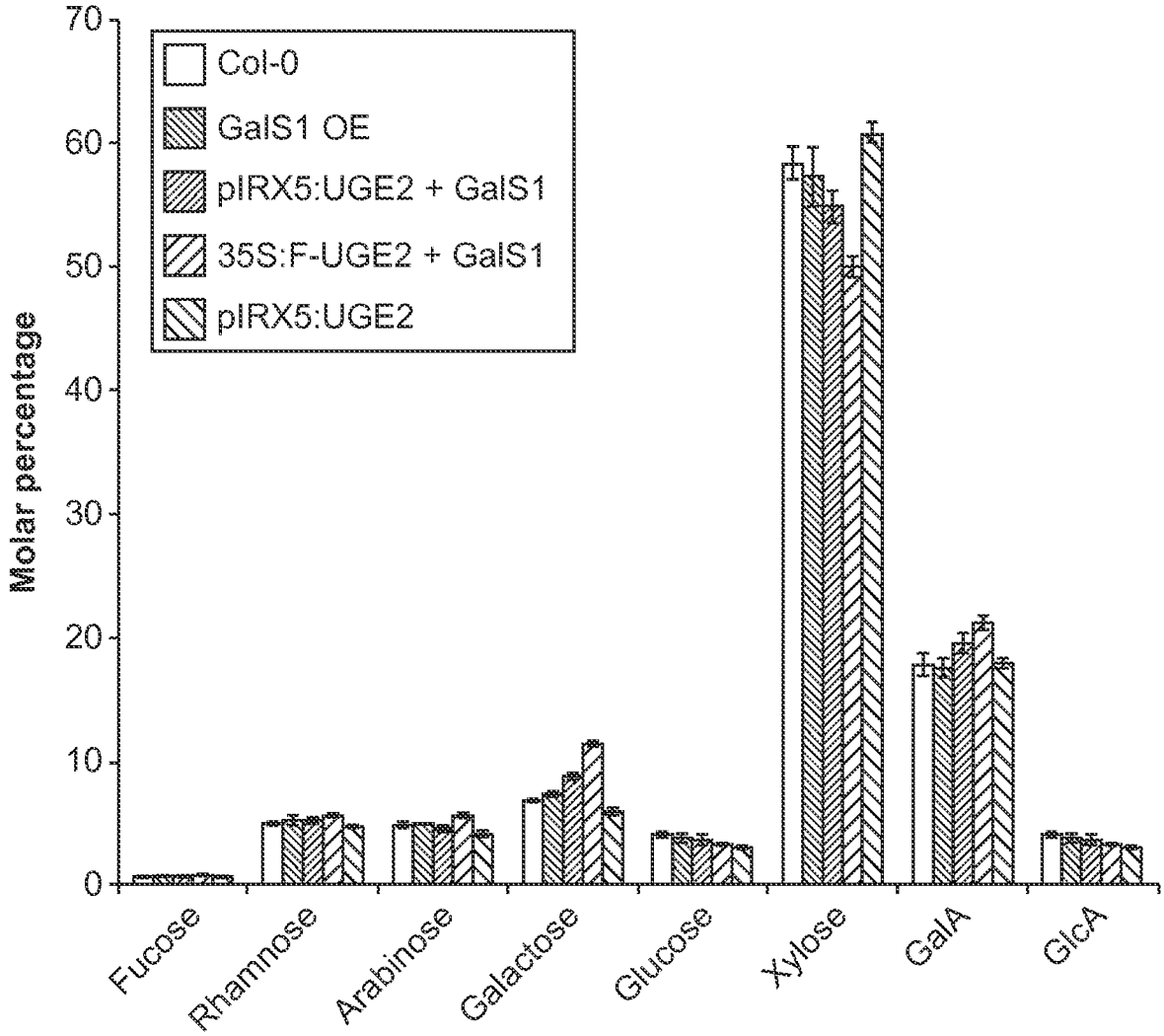
70
60
50
40
30
20
10
0
Molar percentage
Col-0
GalS1 OE
pIRX5:UGE2 + GalS1
35S:F-UGE2 + GalS1
pIRX5:UGE2
Fucose
Rhamnose
Arabinose
Galactose
Glucose
Xylose
GalA
GlcA

FIG. 8

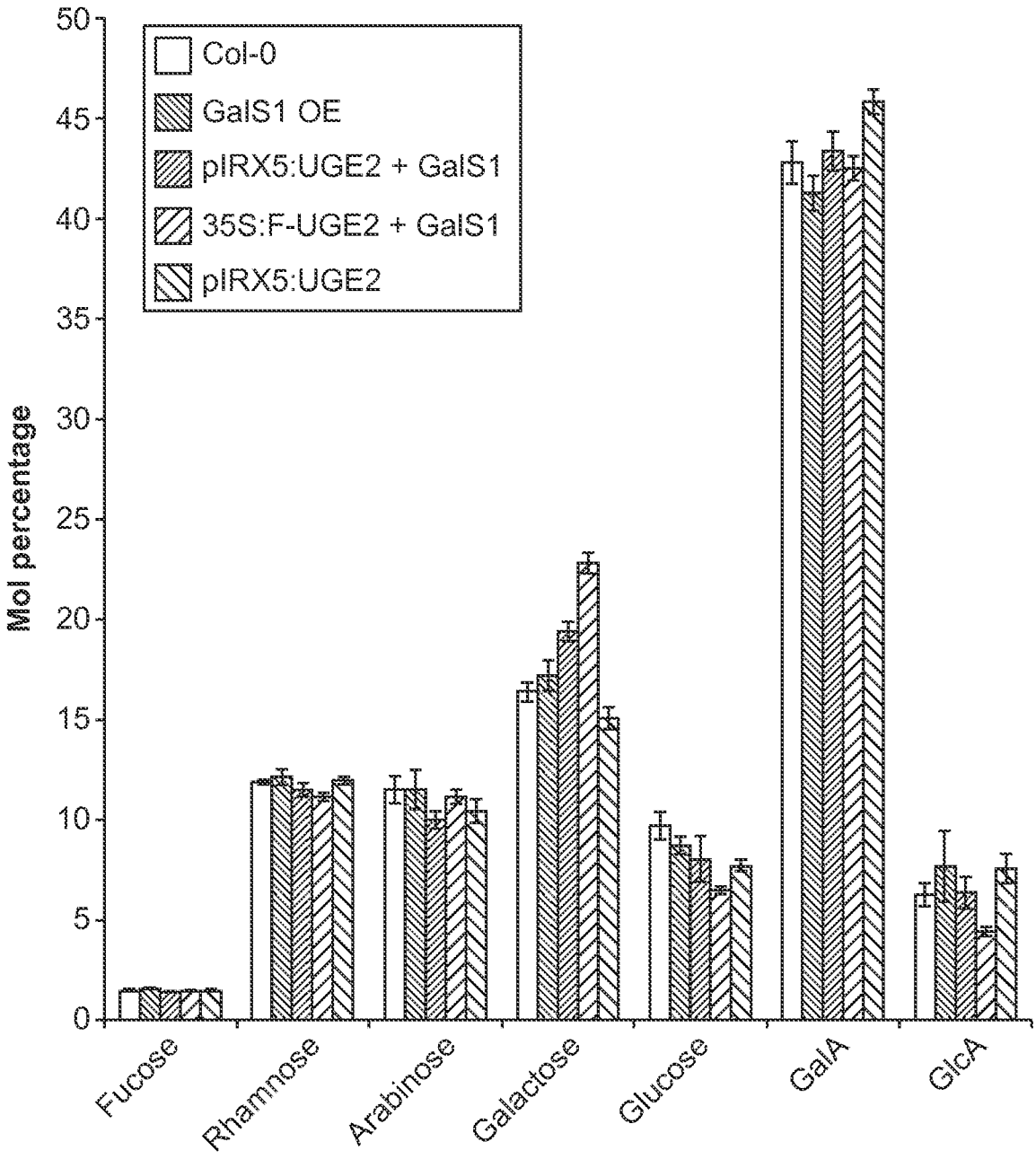
Col-0
GalS1 OE
pIRX5:UGE2 + GalS1
35S:F-UGE2 + GalS1
pIRX5:UGE2
Mol percentage
0
5
10
15
20
25
30
35
40
45
50
Fucose
Rhamnose
Arabinose
Galactose
Glucose
GalA
GlcA

FIG. 9