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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 February 2008 (21.02.2008)

PCT

(10) International Publication Number
WO 2008/021021 A2

(51) International Patent Classification:
A01H 5/00 (2006.01)

(21) International Application Number:
PCT/US2007/017321

(22) International Filing Date: 3 August 2007 (03.08.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/836,243 7 August 2006 (07.08.2006) US

(71) Applicant *(for all designated States except US)*: **MENDEL BIOTECHNOLOGY, INC.** [US/US]; 21375 Cabot Boulevard, Hayward, CA 94545 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **RATCLIFFE, Oliver, J.** [GB/US]; 814 East 21st Street, Oakland, CA 94606 (US). **REPETTI, Peter, P.** [US/US]; 1200 65th Street, Apt. 231, Berkeley, CA 94709 (US). **GUTTERSON, Neal, I.** [US/US]; 5169 Golden Gate Avenue, Oakland, CA 94618 (US). **CREELMAN, Robert, A.** [US/US]; 17931 Lamson Road, Castro Valley, CA 94546 (US).

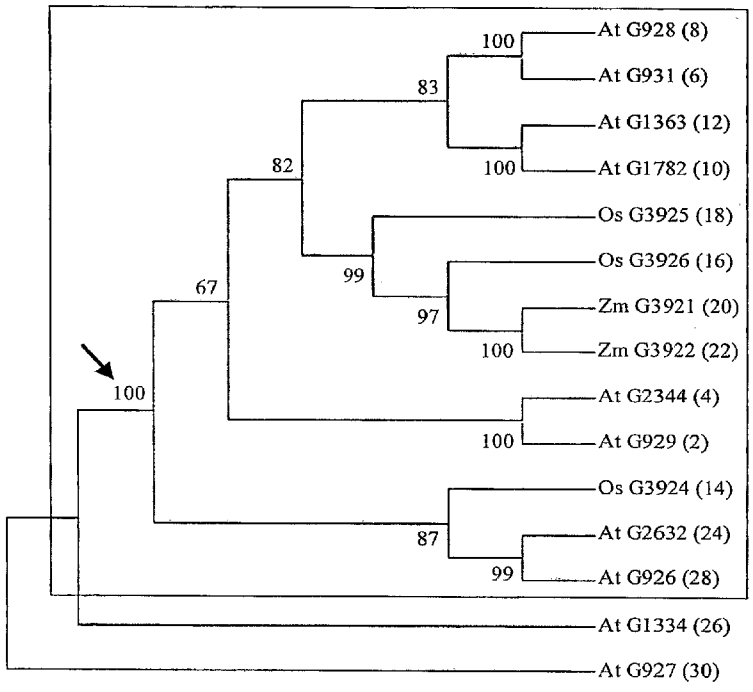
(74) Agents: **WARD, Michael, R.** et al.; Morrison & Foerster LLP, 425 Market Street, San Francisco, CA 94105-2482 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: PLANTS WITH ENHANCED SIZE AND GROWTH RATE



(57) Abstract: Polynucleotides and polypeptides incorporated into expression vectors have been introduced into plants and were ectopically expressed. The polypeptides of the invention regulate transcription in these plants and have been shown to confer at least one regulatory activity that results in increased size, biomass, growth rate, and/or yield as compared to a control plan.

WO 2008/021021 A2

Published:

— *without international search report and to be republished upon receipt of that report*

— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

PLANTS WITH ENHANCED SIZE AND GROWTH RATE

RELATED APPLICATION

This application claims priority to U.S. Patent Application 60/836,243, filed on August 7, 2006.

JOINT RESEARCH AGREEMENT

The claimed invention, in the field of functional genomics and the characterization of plant genes for the improvement of plants, was made by or on behalf of Mendel Biotechnology, Inc. and Monsanto Corporation as a result of activities undertaken within the scope of a joint research agreement in effect on or before the date the claimed invention was made.

FIELD OF THE INVENTION

The present invention relates to plant genomics and plant improvement.

BACKGROUND OF THE INVENTION

Increasing the size or growth rate of a commercially valuable plant provides a number of important practical applications, and may contribute to an increase in yield. For example, increasing the size of a cultivar may generate higher yield of the edible vegetative portion a crop plant. Increasing the size and/or growth rate of a plant may also provide a competitive advantage in the field. Many weeds outgrow slow-growing young crops or out-compete them for nutrients, and thus it is usually desirable to use plants that establish themselves quickly. Seedlings and young plants are also particularly susceptible to stress conditions such as salinity or disease. Increasing seedling growth rate and shortening the time to emergence from soil contributes to seedling vigor, aids seedlings in coping with these stresses, and may allow these crops to be planted earlier in the season. Early planting helps add days to a critical seed or grain-filling period and increases yield. Modification of the biomass of other tissues, such as root tissue, may be useful to improve a plant's ability to grow under harsh environmental conditions, including drought, high salt or nutrient deprivation, because larger roots may better reach or take up water or nutrients.

For many plants, including fruit-bearing trees, plants that are used for biofuels, trees that are used for lumber production, or trees and shrubs that serve as view or wind screens, increased stature provides improved benefits in the forms of greater yield or improved screening.

Increased leaf size may also be of particular interest. Increasing leaf biomass can be used to increase production of plant-derived pharmaceutical or industrial products. An increase in total plant photosynthesis is typically achieved by increasing leaf area of the plant. Additional photosynthetic capacity may be used to increase yield derived from particular plant tissue, including leaves, roots, fruits or seed, or permit better growth of a plant under both decreased and high light intensity.

However, increasing the size or growth rate of a plant may require controlling a number of regulatory and synthetic pathways. Transcription factors are proteins that influence the expression of a particular gene or sets of genes. Altering the expression of one or more transcription factors may provide the necessary control to manipulate complex biochemical or morphological traits in a plant, and thus multiple cellular processes. This application demonstrates that transformed plants that comprise cells having altered levels of at least one of the closely-related transcription factors of the invention exhibit increased size and/or growth rate relative to control plants.

SUMMARY OF THE INVENTION

An object of this invention is to provide plants that can express genes to increase the yield of commercially significant plants by increasing the growth rate, yield, and/or mass of the plants. A plant of the invention is transformed with an expression vector that encodes a CCAAT family transcription factor polypeptide of the invention, and the polypeptide is then overexpressed in the plant. Due to the function of these polynucleotides and their encoded polypeptides, the transgenic plant will have greater yield and/or increased size and/or growth rate at one or more stages of growth as compared to a control plant.

Methods for producing transgenic plants having increased size, yield and/or growth rate are also encompassed by the invention. These method steps include first providing an expression vector comprising a recombinant polynucleotide of the invention. The expression vector may also include at least one regulatory element flanking the polynucleotide sequence. Generally, the regulatory element(s) control expression of the recombinant polynucleotide in a target plant. The expression vector is then introduced into plant cells. The plant cells overexpress a polypeptide encoded by the recombinant polynucleotide, resulting in increased size and/or growth rate of the plant. Those plants that have increased yield, size and/or growth rate may be identified and possibly selected on the basis of the extent to which yield, size and/or growth rate is increased.

The recombinant polynucleotides, expression vectors and transgenic plants of the invention may comprise any of the following sequences:

(a) the nucleotide sequences found in the sequence listing;

(b) nucleotide sequences encoding polypeptides found in the sequence listing;

(c) sequence variants that are at least 35% sequence identical to any of the nucleotide sequences of (a) or (b);

(d) polypeptide sequences that are at least 35%, at least 36%, at least 37%, at least 38%, at least 39%, at least 40%, at least 41%, at least 42%, at least 45%, at least 46%, at least 47%, at least 49%, at least 50%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 61%, at least 62%, at least 63%, at least 72%, at least 78%, at least 79%, or at least 86% identical in their amino acid sequence to any of SEQ ID NOs: 2n, where n=1 to 42, or SEQ ID NOs: 198, 202, 210 or 213;

(e) orthologous and paralogous nucleotide sequences that are at least 35% identical to any of the nucleotide sequences of (a) or (b);

(f) nucleotide sequence that hybridize to any of the nucleotide sequences of (a) or (b) under stringent conditions, which may include, for example, hybridization with wash steps of 6x SSC and 65° C for ten to thirty minutes per wash step; and

(g) polypeptides, and the nucleotide sequences that encode them, having a conserved CCAAT family domain required for the function of regulating transcription and increasing size or biomass in a transgenic plant, the conserved domain being at least 34%, at least 37%, at least 39%, at least 44%, at least 47%, at least 52%, at least 55%, at least 61%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 70%, at least 71%, at least 72%, at least 73%, at least 75%, at least 76%, at least 78%, at least 80%, at least 81%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 89%, at least 90%, at least 91%, at least 92%, at least 96%, or 100% identical to any of the phylogenetically-related conserved domains of SEQ ID NO: 85-126, or SEQ ID NOs: 199, 203, 211 or 214. The polypeptides of the invention, SEQ ID NO: 2n, where n=1 to 42, or 198, 202, 210 or 213 are listed in Tables 1 - 4. Each polypeptide of the invention comprises a conserved domain required for the function of regulating transcription and altering a trait in a transgenic plant, said trait selected from the group consisting of increased size (for example, seedling size or size of the mature plant), increased growth rate,

increased yield, increased biomass, and increased height, as compared to the control plant.

The expression vectors, and hence the transgenic plants of the invention, comprise putative transcription factor polynucleotides sequences and, in particular, CCAAT family HAP2-like (NF-YA) and HAP5-like (NF-YC) sequences comprising conserved domains that are required for subunit association and/or DNA binding, and hence the regulatory activity of the CCAAT-box transcription factor complex. When any of the polypeptides of the invention is overexpressed in a plant, the polypeptide confers at least one transcriptional regulatory activity to the plant, which in turn is manifested in a trait selected from the group consisting of increased growth rate, increased size, increased biomass, increased yield, and increased height as compared to the control plant.

The invention is also directed to transgenic seed produced by any of the transgenic plants of the invention, and to methods for making transgenic seed.

Brief Description of the Sequence Listing and Drawings

The Sequence Listing provides exemplary polynucleotide and polypeptide sequences of the invention. The traits associated with the use of the sequences are included in the Examples.

CD-ROMs Copy 1 and Copy 2, as well as Copy 3, the latter being a CRF copy of the Sequence Listing under CFR Section 1.821(e), are read-only memory computer-readable compact discs. Each contains a copy of the Sequence Listing in ASCII text format. The Sequence Listing is named "MBI-0072P.ST25.txt", the electronic file of the Sequence Listing contained on each of these CD-ROMs was created on August 4, 2006, and the file is 331 kilobytes in size. The copies of the Sequence Listing on the CD-ROM discs are hereby incorporated by reference in their entirety.

Figure 1 shows a conservative estimate of phylogenetic relationships among the orders of flowering plants (modified from Angiosperm Phylogeny Group (1998) *Ann. Missouri Bot. Gard.* 84: 1-49). Those plants with a single cotyledon (monocots) are a monophyletic clade nested within at least two major lineages of dicots; the eudicots are further divided into rosids and asterids. *Arabidopsis* is a rosid eudicot classified within the order Brassicales; rice is a member of the monocot order Poales. Figure 1 was adapted from Daly et al., 2001).

In Figures 2 and 3, phylogenetic trees and multiple sequence alignments of related transcription factors in the HAP2 and HAP5 CCAAT binding families, respectively, were

constructed using ClustalW (CLUSTAL W Multiple Sequence Alignment Program version 1.83, 2003). ClustalW multiple alignment parameters were:

Gap Opening Penalty :10.00

Gap Extension Penalty :0.20

Delay divergent sequences :30 %

DNA Transitions Weight :0.50

Protein weight matrix :Gonnet series

DNA weight matrix :IUB

Use negative matrix :OFF

A FastA formatted alignment was then used to generate phylogenetic trees in MEGA2 software (MEGA2 (http://www.megasoftware.net) using the neighbor joining algorithm and a p-distance model. A test of phylogeny was done via bootstrap with 1000 replications and Random Seed set to default. Cut off values of the bootstrap tree were set to 50%. Closely-related homologs of G929 (SEQ ID NO: 2) or G3911 (SEQ ID NO: 36) are considered as being those proteins descending from ancestral sequences indicated by strong nodes of the trees. In Figures 2 and 3, two ancestral nodes are indicated by arrows have bootstrap values of 100 and 84, respectively. Sequences of closely related homologs that descended from these ancestral nodes are shown within the large boxes in these figures. As indicated in the experiments found in the Examples, many of these sequences have been overexpressed in plants and have been shown to retain the function of increasing size and/or growth rate. SEQ ID NOs. appear in parentheses. Abbreviations: At - *Arabidopsis thaliana*; Dc - *Daucus carota*; Ga- *Gossypium arboreum*; Gm - Glycine max; Gr - *Gossypium raimondii*; Le - *Lycopersicon esculentum*; Mt - *Medicago truncatula*; Nb - *Nicotiana benthamiana*; Os - *Oryza sativa*; Pp - *Physcomitrella patens*; Sb - *Sorghum bicolor*; St - *Solanum tuberosum*; Zm - *Zea mays*.

Figures 4A-4G show a Clustal W alignment of HAP2 transcription factors. SEQ ID NOs: appear in parentheses after each Gene IDentifier (GID). GIDs representing HAP2 polypeptides that are closely related to G929 and G3926 appear in the boxes along the left margin in Figures 4A-4G. Highly conserved domains comprising the contiguous subunit association domains and DNA binding domains (Edwards et al., 1998) are identified in Figures 4D-4E by the large boxes surrounding the residues within these domains.

Figures 5A-5G show a Clustal W alignment of HAP5 transcription factors. SEQ ID NOs: appear in parentheses after each Gene IDentifier (GID). GIDs representing HAP5 polypeptides that are closely related to G3911 and G3543 appear in the boxes along the left margin in Figures

5A-5G. The highly conserved "core sequence" domains first described in related sequences by Edwards et al. (1998) are identified in Figures 5B-5D by the large boxes surrounding the residues within these domains.

Figure 6 shows a field of transgenic tomato plants overexpressing a number of different promoter and transcription factor combinations. Of particular note is a transgenic plant in the center of this photograph, indicated by the arrow, overexpressing G929 under the regulatory control of the cruciferin promoter. This plant was transformed with a two component expression system consisting of SEQ ID NO: 205 (a driver vector comprising the cruciferin promoter, a LexA DNA binding domain, and a GAL4 transactivation (TA) domain) and SEQ ID NO: 206 (comprising a LexA operator (opLexA) and the G929 transcription factor sequence). The transgenic plant was much larger than virtually all of its neighboring plants, including wild-type and empty vector control plants, and was particularly noted for its high vigor, upright stems, and no noticeable loss in fruit production.

DETAILED DESCRIPTION

The present invention relates to polynucleotides and polypeptides for modifying phenotypes of plants, particularly those associated with increased yield with respect to a control plant (for example, a wild-type plant). Throughout this disclosure, various information sources are referred to and/or are specifically incorporated. The information sources include scientific journal articles, patent documents, textbooks, and World Wide Web browser-inactive page addresses. While the reference to these information sources clearly indicates that they can be used by one of skill in the art, each and every one of the information sources cited herein are specifically incorporated in their entirety, whether or not a specific mention of "incorporation by reference" is noted. The contents and teachings of each and every one of the information sources can be relied on and used to make and use embodiments of the invention.

As used herein and in the appended claims, the singular forms "a", "an", and "the" include the plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a host cell" includes a plurality of such host cells, and a reference to "a stress" is a reference to one or more stresses and equivalents thereof known to those skilled in the art, and so forth.

DEFINITIONS

"Polynucleotide" is a nucleic acid molecule comprising a plurality of polymerized nucleotides, e.g., at least about 15 consecutive polymerized nucleotides. A polynucleotide may be a nucleic acid, oligonucleotide, nucleotide, or any fragment thereof. In many instances, a polynucleotide comprises a nucleotide sequence encoding a polypeptide (or protein) or a domain or fragment thereof. Additionally, the polynucleotide may comprise a promoter, an intron, an enhancer region, a polyadenylation site, a translation initiation site, 5' or 3' untranslated regions, a reporter gene, a selectable marker, or the like. The polynucleotide can be single-stranded or double-stranded DNA or RNA. The polynucleotide optionally comprises modified bases or a modified backbone. The polynucleotide can be, e.g., genomic DNA or RNA, a transcript (such as an mRNA), a cDNA, a PCR product, a cloned DNA, a synthetic DNA or RNA, or the like. The polynucleotide can be combined with carbohydrate, lipids, protein, or other materials to perform a particular activity such as transformation or form a useful composition such as a peptide nucleic acid (PNA). The polynucleotide can comprise a sequence in either sense or antisense orientations. "Oligonucleotide" is substantially equivalent to the terms amplimer, primer, oligomer, element, target, and probe and is preferably single-stranded.

A "recombinant polynucleotide" is a polynucleotide that is not in its native state, e.g., the polynucleotide comprises a nucleotide sequence not found in nature, or the polynucleotide is in a context other than that in which it is naturally found, e.g., separated from nucleotide sequences with which it typically is in proximity in nature, or adjacent (or contiguous with) nucleotide sequences with which it typically is not in proximity. For example, the sequence at issue can be cloned into a vector, or otherwise recombined with one or more additional nucleic acid.

An "isolated polynucleotide" is a polynucleotide, whether naturally occurring or recombinant, that is present outside the cell in which it is typically found in nature, whether purified or not. Optionally, an isolated polynucleotide is subject to one or more enrichment or purification procedures, e.g., cell lysis, extraction, centrifugation, precipitation, or the like.

"Gene" or "gene sequence" refers to the partial or complete coding sequence of a gene, its complement, and its 5' or 3' untranslated regions. A gene is also a functional unit of inheritance, and in physical terms is a particular segment or sequence of nucleotides along a molecule of DNA (or RNA, in the case of RNA viruses) involved in producing a polypeptide chain. The latter may be subjected to subsequent processing such as chemical modification or folding to obtain a functional protein or polypeptide. A gene may be isolated, partially isolated, or found with an organism's genome. By way of example, a transcription factor gene encodes a

transcription factor polypeptide, which may be functional or require processing to function as an initiator of transcription.

Operationally, genes may be defined by the cis-trans test, a genetic test that determines whether two mutations occur in the same gene and that may be used to determine the limits of the genetically active unit (Rieger et al., 1976). A gene generally includes regions preceding ("leaders"; upstream) and following ("trailers"; downstream) the coding region. A gene may also include intervening, non-coding sequences, referred to as "introns", located between individual coding segments, referred to as "exons". Most genes have an associated promoter region, a regulatory sequence 5' of the transcription initiation codon (there are some genes that do not have an identifiable promoter). The function of a gene may also be regulated by enhancers, operators, and other regulatory elements.

A "polypeptide" is an amino acid sequence comprising a plurality of consecutive polymerized amino acid residues e.g., at least about 15 consecutive polymerized amino acid residues. In many instances, a polypeptide comprises a polymerized amino acid residue sequence that is a transcription factor or a domain or portion or fragment thereof. Additionally, the polypeptide may comprise: (i) a localization domain; (ii) an activation domain; (iii) a repression domain; (iv) an oligomerization domain; (v) a protein-protein interaction domain; (vi) a DNA-binding domain; or the like. The polypeptide optionally comprises modified amino acid residues, naturally occurring amino acid residues not encoded by a codon, non-naturally occurring amino acid residues.

"Protein" refers to an amino acid sequence, oligopeptide, peptide, polypeptide, or portions thereof whether naturally occurring or synthetic.

"Portion", as used herein, refers to any part of a protein used for any purpose, but especially for the screening of a library of molecules which specifically bind to that portion or for the production of antibodies.

A "recombinant polypeptide" is a polypeptide produced by translation of a recombinant polynucleotide. A "synthetic polypeptide" is a polypeptide created by consecutive polymerization of isolated amino acid residues using methods well known in the art. An "isolated polypeptide," whether a naturally occurring or a recombinant polypeptide, is more enriched in (or out of) a cell than the polypeptide in its natural state in a wild-type cell, e.g., more than about 5% enriched, more than about 10% enriched, or more than about 20%, or more than about 50%, or more, enriched, i.e., alternatively denoted: 105%, 110%, 120%, 150% or more, enriched relative to wild type standardized at 100%. Such an enrichment is not the result

of a natural response of a wild-type plant. Alternatively, or additionally, the isolated polypeptide is separated from other cellular components with which it is typically associated, e.g., by any of the various protein purification methods herein.

"Homology" refers to sequence similarity between a reference sequence and at least a fragment of a newly sequenced clone insert or its encoded amino acid sequence.

"Identity" or "similarity" refers to sequence similarity between two polynucleotide sequences or between two polypeptide sequences, with identity being a more strict comparison. The phrases "percent identity" and "% identity" refer to the percentage of sequence similarity found in a comparison of two or more polynucleotide sequences or two or more polypeptide sequences. "Sequence similarity" refers to the percent similarity in base pair sequence (as determined by any suitable method) between two or more polynucleotide sequences. Two or more sequences can be anywhere from 0-100% similar, or any integer value therebetween. Identity or similarity can be determined by comparing a position in each sequence that may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same nucleotide base or amino acid, then the molecules are identical at that position. A degree of similarity or identity between polynucleotide sequences is a function of the number of identical, matching or corresponding nucleotides at positions shared by the polynucleotide sequences. A degree of identity of polypeptide sequences is a function of the number of identical amino acids at corresponding positions shared by the polypeptide sequences. A degree of homology or similarity of polypeptide sequences is a function of the number of amino acids at corresponding positions shared by the polypeptide sequences.

"Alignment" refers to a number of nucleotide bases or amino acid residue sequences aligned by lengthwise comparison so that components in common (i.e., nucleotide bases or amino acid residues at corresponding positions) may be visually and readily identified. The fraction or percentage of components in common is related to the homology or identity between the sequences. Alignments such as those of Figures 4A-4G and specifically 4D-4E may be used to identify conserved domains and relatedness within these domains. An alignment may suitably be determined by means of computer programs known in the art, such as MACVECTOR software (1999) (Accelrys, Inc., San Diego, CA).

A "conserved domain" or "conserved region" as used herein refers to a region in heterologous polynucleotide or polypeptide sequences where there is at least one similar conserved function and a relatively high degree of sequence identity between the distinct sequences. Subunit association domains and DNA binding domains such as are found in a

polypeptide member of HAP2 transcription factors, or HAP5 core sequences from HAP2 transcription factors (Edwards, 1998) are examples of a conserved domain. With respect to polynucleotides encoding presently disclosed polypeptides, a conserved domain is preferably at least nine base pairs (bp) in length. A conserved domain with respect to presently disclosed polypeptides refers to a domain within a polypeptide family that exhibits similar function and a higher degree of sequence homology, such as at least about 34%, at least about 37%, at least about 39%, at least about 44%, at least about 47%, at least about 52%, at least about 55%, at least about 61%, at least about 63%, at least about 64%, at least about 65%, at least about 66%, at least about 67%, at least about 70%, at least about 71%, at least about 72%, at least about 73%, at least about 75%, at least about 76%, at least about 78%, at least about 80%, at least about 81%, at least about 83%, at least about 84%, at least about 85%, at least about 86%, at least about 87%, at least about 89%, at least about 90%, at least about 91%, at least about 92%, at least about 96%, or 100% amino acid sequence identity to the similar conserved domains of SEQ ID NO: 85-126, 199, 203, 211, or 214. Sequences that possess or encode for conserved domains that meet these criteria of percentage identity, and that have comparable biological activity to the present polypeptide sequences, thus being members of the HAP2 and HAP5 polypeptides, are encompassed by the invention. A fragment or domain can be referred to as outside a conserved domain, outside a consensus sequence, or outside a consensus DNA-binding site that is known to exist or that exists for a particular polypeptide class, family, or sub-family. In this case, the fragment or domain will not include the exact amino acids of a consensus sequence or consensus DNA-binding site of a transcription factor class, family or sub-family, or the exact amino acids of a particular transcription factor consensus sequence or consensus DNA-binding site. Furthermore, a particular fragment, region, or domain of a polypeptide, or a polynucleotide encoding a polypeptide, can be "outside a conserved domain" if all the amino acids of the fragment, region, or domain fall outside of a defined conserved domain(s) for a polypeptide or protein. Sequences having lesser degrees of identity but comparable biological activity are considered to be equivalents.

As one of ordinary skill in the art recognizes, conserved domains may be identified as regions or domains of identity to a specific consensus sequence (see, for example, Riechmann et al., 2000a, 2000b). Edwards (1998) defined conserved domains of HAP2 transcription factor sequences, and identified these contiguous domains as comprising subunit association and DNA binding activity. Edwards (1998) also defined conserved "core sequence" domains of HAP5 transcription factors that comprise a predicted histone fold triple helix for dimerization and are

required for formation of the CCAAT-box transcription factor complex. Thus, by using alignment methods well known in the art, the conserved domains of the CCAAT binding transcription factor proteins may be determined. The conserved domains for many of the polypeptide sequences of the invention are listed in Tables 1 - 4. Also, the polypeptides of Tables 1 - 4 have conserved domains specifically indicated by amino acid coordinates of the full length polypeptides. It is expected that these conserved domains are required for the functions of subunit association and/or DNA binding (Edwards, 1998).

"Complementary" refers to the natural hydrogen bonding by base pairing between purines and pyrimidines. For example, the sequence A-C-G-T (5' -> 3') forms hydrogen bonds with its complements A-C-G-T (5' -> 3') or A-C-G-U (5' -> 3'). Two single-stranded molecules may be considered partially complementary, if only some of the nucleotides bond, or "completely complementary" if all of the nucleotides bond. The degree of complementarity between nucleic acid strands affects the efficiency and strength of hybridization and amplification reactions. "Fully complementary" refers to the case where bonding occurs between every base pair and its complement in a pair of sequences, and the two sequences have the same number of nucleotides.

The terms "highly stringent" or "highly stringent condition" refer to conditions that permit hybridization of DNA strands whose sequences are highly complementary, wherein these same conditions exclude hybridization of significantly mismatched DNAs. Polynucleotide sequences capable of hybridizing under stringent conditions with the polynucleotides of the present invention may be, for example, variants of the disclosed polynucleotide sequences, including allelic or splice variants, or sequences that encode orthologs or paralogs of presently disclosed polypeptides. Nucleic acid hybridization methods are disclosed in detail by Kashima et al., 1985, Sambrook et al., 1989, and by Haymes et al., 1985), which references are incorporated herein by reference.

In general, stringency is determined by the temperature, ionic strength, and concentration of denaturing agents (e.g., formamide) used in a hybridization and washing procedure (for a more detailed description of establishing and determining stringency, see the section "Identifying Polynucleotides or Nucleic Acids by Hybridization", below). The degree to which two nucleic acids hybridize under various conditions of stringency is correlated with the extent of their similarity. Thus, similar nucleic acid sequences from a variety of sources, such as within a plant's genome (as in the case of paralogs) or from another plant (as in the case of orthologs) that may perform similar functions can be isolated on the basis of their ability to hybridize with

known related polynucleotide sequences. Numerous variations are possible in the conditions and means by which nucleic acid hybridization can be performed to isolate related polynucleotide sequences having similarity to sequences known in the art and are not limited to those explicitly disclosed herein. Such an approach may be used to isolate polynucleotide sequences having various degrees of similarity with disclosed polynucleotide sequences, such as, for example, encoded transcription factors having 34% or greater identity with a conserved domain of disclosed sequences as provided in SEQ ID NOs: 85-126, 199, 203, 211, or 214.

The terms "paralog" and "ortholog" are defined below in the section entitled "Orthologs and Paralogs". In brief, orthologs and paralogs are evolutionarily related genes that have similar sequences and functions. Orthologs are structurally related genes in different species that are derived by a speciation event. Paralogs are structurally related genes within a single species that are derived by a duplication event.

The term "equivalog" describes members of a set of homologous proteins that are conserved with respect to function since their last common ancestor. Related proteins are grouped into equivalog families, and otherwise into protein families with other hierarchically defined homology types. This definition is provided at the Institute for Genomic Research (TIGR) World Wide Web (www) website, under "http://www.tigr.org/TIGRFAMs/Explanations.shtml" for the heading "Terms associated with TIGRFAMs".

In general, the term "variant" refers to molecules with some differences, generated synthetically or naturally, in their base or amino acid sequences as compared to a reference (native) polynucleotide or polypeptide, respectively. These differences include substitutions, insertions, deletions or any desired combinations of such changes in a native polynucleotide of amino acid sequence.

With regard to polynucleotide variants, differences between presently disclosed polynucleotides and polynucleotide variants are limited so that the nucleotide sequences of the former and the latter are closely similar overall and, in many regions, identical. Due to the degeneracy of the genetic code, differences between the former and latter nucleotide sequences may be silent (i.e., the amino acids encoded by the polynucleotide are the same, and the variant polynucleotide sequence encodes the same amino acid sequence as the presently disclosed polynucleotide. Variant nucleotide sequences may encode different amino acid sequences, in which case such nucleotide differences will result in amino acid substitutions, additions, deletions, insertions, truncations or fusions with respect to the similar disclosed polynucleotide

sequences. These variations may result in polynucleotide variants encoding polypeptides that share at least one functional characteristic. The degeneracy of the genetic code also dictates that many different variant polynucleotides can encode identical and/or substantially similar polypeptides in addition to those sequences illustrated in the Sequence Listing.

Also within the scope of the invention is a variant of a nucleic acid listed in the Sequence Listing, that is, one having a sequence that differs from the one of the polynucleotide sequences in the Sequence Listing, or a complementary sequence, that encodes a functionally equivalent polypeptide (i.e., a polypeptide having some degree of equivalent or similar biological activity) but differs in sequence from the sequence in the Sequence Listing, due to degeneracy in the genetic code. Included within this definition are polymorphisms that may or may not be readily detectable using a particular oligonucleotide probe of the polynucleotide encoding polypeptide, and improper or unexpected hybridization to allelic variants, with a locus other than the normal chromosomal locus for the polynucleotide sequence encoding polypeptide.

"Allelic variant" or "polynucleotide allelic variant" refers to any of two or more alternative forms of a gene occupying the same chromosomal locus. Allelic variation arises naturally through mutation, and may result in phenotypic polymorphism within populations. Gene mutations may be "silent" or may encode polypeptides having altered amino acid sequence. "Allelic variant" and "polypeptide allelic variant" may also be used with respect to polypeptides, and in this case the terms refer to a polypeptide encoded by an allelic variant of a gene.

"Splice variant" or "polynucleotide splice variant" as used herein refers to alternative forms of RNA transcribed from a gene. Splice variation naturally occurs as a result of alternative sites being spliced within a single transcribed RNA molecule or between separately transcribed RNA molecules, and may result in several different forms of mRNA transcribed from the same gene. Thus, splice variants may encode polypeptides having different amino acid sequences, which may or may not have similar functions in the organism. "Splice variant" or "polypeptide splice variant" may also refer to a polypeptide encoded by a splice variant of a transcribed mRNA.

As used herein, "polynucleotide variants" may also refer to polynucleotide sequences that encode paralogs and orthologs of the presently disclosed polypeptide sequences. "Polypeptide variants" may refer to polypeptide sequences that are paralogs and orthologs of the presently disclosed polypeptide sequences.

Differences between presently disclosed polypeptides and polypeptide variants are limited so that the sequences of the former and the latter are closely similar overall and, in many regions, identical. Presently disclosed polypeptide sequences and similar polypeptide variants may differ in amino acid sequence by one or more substitutions, additions, deletions, fusions and truncations, which may be present in any combination. These differences may produce silent changes and result in a functionally equivalent polypeptides. Thus, it will be readily appreciated by those of skill in the art, that any of a variety of polynucleotide sequences is capable of encoding the polypeptides and homolog polypeptides of the invention. A polypeptide sequence variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties. Deliberate amino acid substitutions may thus be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues, as long as a significant amount of the functional or biological activity of the polypeptide is retained. For example, negatively charged amino acids may include aspartic acid and glutamic acid, positively charged amino acids may include lysine and arginine, and amino acids with uncharged polar head groups having similar hydrophilicity values may include leucine, isoleucine, and valine; glycine and alanine; asparagine and glutamine; serine and threonine; and phenylalanine and tyrosine. More rarely, a variant may have "non-conservative" changes, e.g., replacement of a glycine with a tryptophan. Similar minor variations may also include amino acid deletions or insertions, or both. Related polypeptides may comprise, for example, additions and/or deletions of one or more N-linked or O-linked glycosylation sites, or an addition and/or a deletion of one or more cysteine residues. Guidance in determining which and how many amino acid residues may be substituted, inserted or deleted without abolishing functional or biological activity may be found using computer programs well known in the art, for example, DNASTAR software (see USPN 5,840,544).

"Fragment", with respect to a polynucleotide, refers to a clone or any part of a polynucleotide molecule that retains a usable, functional characteristic. Useful fragments include oligonucleotides and polynucleotides that may be used in hybridization or amplification technologies or in the regulation of replication, transcription or translation. A "polynucleotide fragment" refers to any subsequence of a polynucleotide, typically, of at least about 9 consecutive nucleotides, preferably at least about 30 nucleotides, more preferably at least about 50 nucleotides, of any of the sequences provided herein. Exemplary polynucleotide fragments are the first sixty consecutive nucleotides of the polynucleotides listed in the Sequence Listing. Exemplary fragments also include fragments that comprise a region that encodes an conserved

domain of a polypeptide. Exemplary fragments also include fragments that comprise a conserved domain of a polypeptide. Exemplary fragments include fragments that comprise an conserved domain of a polypeptide, for example, amino acid residues 98-157 of G929 (SEQ ID NO: 2), amino acid residues 164-222 of G3926 (SEQ ID NO: 18), amino acid residues 83-148 of G3911 (SEQ ID NO: 36) or amino acid residues 70-135 of G3543 (SEQ ID NO: 68).

Fragments may also include subsequences of polypeptides and protein molecules, or a subsequence of the polypeptide. Fragments may have uses in that they may have antigenic potential. In some cases, the fragment or domain is a subsequence of the polypeptide that performs at least one biological function of the intact polypeptide in substantially the same manner, or to a similar extent, as does the intact polypeptide. For example, a polypeptide fragment can comprise a recognizable structural motif or functional domain such as a DNA-binding site or domain that binds to a DNA promoter region, an activation domain, or a domain for protein-protein interactions, and may initiate transcription. Fragments can vary in size from as few as 3 amino acid residues to the full length of the intact polypeptide, but are preferably at least about 30 amino acid residues in length and more preferably at least about 60 amino acid residues in length.

The invention also encompasses production of DNA sequences that encode polypeptides and derivatives, or fragments thereof, entirely by synthetic chemistry. After production, the synthetic sequence may be inserted into any of the many available expression vectors and cell systems using reagents well known in the art. Moreover, synthetic chemistry may be used to introduce mutations into a sequence encoding polypeptides or any fragment thereof.

"Derivative" refers to the chemical modification of a nucleic acid molecule or amino acid sequence. Chemical modifications can include replacement of hydrogen by an alkyl, acyl, or amino group or glycosylation, pegylation, or any similar process that retains or enhances biological activity or lifespan of the molecule or sequence.

The term "plant" includes whole plants, shoot vegetative organs/structures (for example, leaves, stems and tubers), roots, flowers and floral organs/structures (for example, bracts, sepals, petals, stamens, carpels, anthers and ovules), seed (including embryo, endosperm, and seed coat) and fruit (the mature ovary), plant tissue (for example, vascular tissue, ground tissue, and the like) and cells (for example, guard cells, egg cells, and the like), and progeny of same. The class of plants that can be used in the method 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, horsetails, psilophytes,

lycophytes, bryophytes, and multicellular algae (see for example, Figure 1, adapted from Daly et al., 2001, and see also Tudge, 2000).

A "control plant" as used in the present invention refers to a plant cell, seed, plant component, plant tissue, plant organ or whole plant used to compare against transgenic or genetically modified plant for the purpose of identifying an enhanced phenotype in the transgenic or genetically modified plant. A control plant may in some cases be a transgenic plant line that comprises an empty vector or marker gene, that is, a vector that does not contain the recombinant polynucleotide of the present invention that is expressed in the transgenic or genetically modified plant being evaluated. In general, a control plant is a plant of the same line or variety as the transgenic or genetically modified plant being tested. A suitable control plant would include a genetically unaltered or non-transgenic plant of the parental line used to generate a transgenic plant herein.

A "transgenic plant" refers to a plant that contains genetic material not found in a wild-type plant of the same species, variety or cultivar. The genetic material may include a transgene, an insertional mutagenesis event (such as by transposon or T-DNA insertional mutagenesis), an activation tagging sequence, a mutated sequence, a homologous recombination event or a sequence modified by chimeraplasty. Typically, the foreign genetic material has been introduced into the plant by human manipulation, but any method can be used as one of skill in the art recognizes.

A transgenic plant of the invention generally contains an expression vector or cassette. The expression cassette typically comprises a polypeptide-encoding sequence operably linked (i.e., under regulatory control of) to appropriate inducible or constitutive regulatory sequences that allow for the controlled expression of polypeptide. The expression cassette can be introduced into a plant by transformation or by breeding after transformation of a parent plant. A plant refers to a whole plant as well as to a plant part, such as seed, fruit, leaf, or root, plant tissue, plant cells or any other plant material, e.g., a plant explant, as well as to progeny thereof, and to *in vitro* systems that mimic biochemical or cellular components or processes in a cell.

"Wild type" or "wild-type", as used herein, refers to a plant cell, seed, plant component, plant tissue, plant organ or whole plant that has not been genetically modified or treated in an experimental sense. Wild-type cells, seed, components, tissue, organs or whole plants may be used as controls to compare levels of expression and the extent and nature of trait modification with cells, tissue or plants of the same species in which a polypeptide's expression is altered, e.g., in that it has been knocked out, overexpressed, or ectopically expressed.

A "trait" refers to a physiological, morphological, biochemical, or physical characteristic of a plant or particular plant material or cell. In some instances, this characteristic is visible to the human eye, such as seed or plant size, or can be measured by biochemical techniques, such as detecting the protein, starch, or oil content of seed or leaves, or by observation of a metabolic or physiological process, e.g. by measuring tolerance to water deprivation or particular salt or sugar concentrations, or by the observation of the expression level of a gene or genes, e.g., by employing Northern analysis, RT-PCR, microarray gene expression assays, or reporter gene expression systems, or by agricultural observations such as hyperosmotic stress tolerance or yield. Any technique can be used to measure the amount of, comparative level of, or difference in any selected chemical compound or macromolecule in the transgenic plants.

"Trait modification" refers to a detectable difference in a characteristic in a plant ectopically expressing a polynucleotide or polypeptide of the present invention relative to a plant not doing so, such as a wild-type plant. In some cases, the trait modification can be evaluated quantitatively. For example, the trait modification can entail at least about a 2% increase or decrease, or an even greater difference, in an observed trait as compared with a control or wild-type plant. It is known that there can be a natural variation in the modified trait. Therefore, the trait modification observed entails a change of the normal distribution and magnitude of the trait in the plants as compared to control or wild-type plants.

When two or more plants have "similar morphologies", "substantially similar morphologies", "a morphology that is substantially similar", or are "morphologically similar", the plants have comparable forms or appearances, including analogous features such as overall dimensions, height, width, mass, root mass, shape, glossiness, color, stem diameter, leaf size, leaf dimension, leaf density, internode distance, branching, root branching, number and form of inflorescences, and other macroscopic characteristics, and the individual plants are not readily distinguishable based on morphological characteristics alone.

"Modulates" refers to a change in activity (biological, chemical, or immunological) or lifespan resulting from specific binding between a molecule and either a nucleic acid molecule or a protein.

The term "transcript profile" refers to the expression levels of a set of genes in a cell in a particular state, particularly by comparison with the expression levels of that same set of genes in a cell of the same type in a reference state. For example, the transcript profile of a particular polypeptide in a suspension cell is the expression levels of a set of genes in a cell knocking out or overexpressing that polypeptide compared with the expression levels of that same set of genes

in a suspension cell that has normal levels of that polypeptide. The transcript profile can be presented as a list of those genes whose expression level is significantly different between the two treatments, and the difference ratios. Differences and similarities between expression levels may also be evaluated and calculated using statistical and clustering methods.

With regard to gene knockouts as used herein, the term "knockout" refers to a plant or plant cell having a disruption in at least one gene in the plant or cell, where the disruption results in a reduced expression or activity of the polypeptide encoded by that gene compared to a control cell. The knockout can be the result of, for example, genomic disruptions, including transposons, tilling, and homologous recombination, antisense constructs, sense constructs, RNA silencing constructs, or RNA interference. A T-DNA insertion within a gene is an example of a genotypic alteration that may abolish expression of that gene.

"Ectopic expression or altered expression" in reference to a polynucleotide indicates that the pattern of expression in, e.g., a transgenic plant or plant tissue, is different from the expression pattern in a wild-type plant or a reference plant of the same species. The pattern of expression may also be compared with a reference expression pattern in a wild-type plant of the same species. For example, the polynucleotide or polypeptide is expressed in a cell or tissue type other than a cell or tissue type in which the sequence is expressed in the wild-type plant, or by expression at a time other than at the time the sequence is expressed in the wild-type plant, or by a response to different inducible agents, such as hormones or environmental signals, or at different expression levels (either higher or lower) compared with those found in a wild-type plant. The term also refers to altered expression patterns that are produced by lowering the levels of expression to below the detection level or completely abolishing expression. The resulting expression pattern can be transient or stable, constitutive or inducible. In reference to a polypeptide, the term "ectopic expression or altered expression" further may relate to altered activity levels resulting from the interactions of the polypeptides with exogenous or endogenous modulators or from interactions with factors or as a result of the chemical modification of the polypeptides.

The term "overexpression" as used herein refers to a greater expression level of a gene in a plant, plant cell or plant tissue, compared to expression in a wild-type plant, cell or tissue, at any developmental or temporal stage for the gene. Overexpression can occur when, for example, the genes encoding one or more polypeptides are under the control of a strong promoter (e.g., the cauliflower mosaic virus 35S transcription initiation region). Overexpression may also under the control of an inducible or tissue specific promoter. Thus, overexpression may occur

throughout a plant, in specific tissues of the plant, or in the presence or absence of particular environmental signals, depending on the promoter used.

Overexpression may take place in plant cells normally lacking expression of polypeptides functionally equivalent or identical to the present polypeptides. Overexpression may also occur in plant cells where endogenous expression of the present polypeptides or functionally equivalent molecules normally occurs, but such normal expression is at a lower level. Overexpression thus results in a greater than normal production, or "overproduction" of the polypeptide in the plant, cell or tissue.

The term "transcription regulating region" refers to a DNA regulatory sequence that regulates expression of one or more genes in a plant when a transcription factor having one or more specific binding domains binds to the DNA regulatory sequence. Transcription factors possess a conserved domain. The transcription factors also comprise an amino acid subsequence that forms a transcription activation domain that regulates expression of one or more yield-related genes in a plant when the transcription factor binds to the regulating region.

"Yield" or "plant yield" refers to increased plant growth, increased crop growth, increased biomass, and/or increased plant product production, and is dependent to some extent on temperature, plant size, organ size, planting density, light, water and nutrient availability, and how the plant copes with various stresses, such as through temperature acclimation and water or nutrient use efficiency. Relative indicators of yield may include volume per land area (e.g. bushels per acre) or weight per land area (e.g., kilograms per hectare) measurements.

"Planting density" refers to the number of plants that can be grown per acre. For crop species, planting or population density varies from a crop to a crop, from one growing region to another, and from year to year. Using corn as an example, the average prevailing density in 2000 was in the range of 20,000 - 25,000 plants per acre in Missouri, USA. A desirable higher population density (a measure of yield) would be at least 22,000 plants per acre, and a more desirable higher population density would be at least 28,000 plants per acre, more preferably at least 34,000 plants per acre, and most preferably at least 40,000 plants per acre. The average prevailing densities per acre of a few other examples of crop plants in the USA in the year 2000 were: wheat 1,000,000-1,500,000; rice 650,000-900,000; soybean 150,000-200,000, canola 260,000-350,000, sunflower 17,000-23,000 and cotton 28,000-55,000 plants per acre (Cheikh et al., 2003) U.S. Patent Application No. 20030101479). A desirable higher population density for each of these examples, as well as other valuable species of plants, would be at least 10% higher than the average prevailing density or yield.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Transcription Factors Modify Expression of Endogenous Genes

A transcription factor may include, but is not limited to, any polypeptide that can activate or repress transcription of a single gene or a number of genes. As one of ordinary skill in the art recognizes, transcription factors can be identified by the presence of a region or domain of structural similarity or identity to a specific consensus sequence or the presence of a specific consensus DNA-binding motif (see, for example, Riechmann et al., 2000a). The plant transcription factors of the present invention belong to the CCAAT binding HAP2 or HAP5 families.

Generally, transcription factors are involved in cell differentiation and proliferation and the regulation of growth. Accordingly, one skilled in the art would recognize that by expressing the present sequences in a plant, by, for example, introducing into the plant a polynucleotide sequence encoding a transcription factor of the invention, one may change the expression of autologous genes or induce the expression of introduced genes and thus alter the plant's phenotype to one with improved traits related to size, growth rate and/or yield. Plants may then be selected for those that produce the most desirable degree of over- or under-expression of target genes of interest and coincident trait improvement.

The sequences of the present invention may be derived from any species, particularly plant species, in a naturally occurring form or from any source whether natural, synthetic, semi-synthetic or recombinant. The sequences of the invention may also include functional fragments of the present amino acid sequences. Where "amino acid sequence" is recited to refer to an amino acid sequence of a naturally occurring protein molecule, "amino acid sequence" and like terms are not meant to limit the amino acid sequence to the complete native amino acid sequence associated with the recited protein molecule.

In addition to methods for modifying a plant phenotype by employing one or more polynucleotides and polypeptides of the invention described herein, the polynucleotides and polypeptides of the invention have a variety of additional uses. These uses include their use in the recombinant production (i.e., expression) of proteins; as regulators of plant gene expression, as diagnostic probes for the presence of complementary or partially complementary nucleic acids (including for detection of natural coding nucleic acids); as substrates for further reactions, e.g., mutation reactions, PCR reactions, or the like; as substrates for cloning e.g., including digestion or ligation reactions; and for identifying exogenous or endogenous modulators of the

transcription factors. The polynucleotide can be, e.g., genomic DNA or RNA, a transcript (such as an mRNA), a cDNA, a PCR product, a cloned DNA, a synthetic DNA or RNA, or the like. The polynucleotide can comprise a sequence in either sense or antisense orientations.

Expression of genes that encode polypeptides that modify expression of endogenous genes, polynucleotides, and proteins are well known in the art. In addition, transgenic plants comprising isolated polynucleotides encoding transcription factors may also modify expression of endogenous genes, polynucleotides, and proteins. Examples include Peng et al. (1997) and Peng et al. (1999). In addition, many others have demonstrated that an *Arabidopsis* transcription factor expressed in an exogenous plant species elicits the same or very similar phenotypic response (see, for example, Fu et al., 2001; Nandi et al., 2000; Coupland, 1995; and Weigel and Nilsson, 1995).

In another example, Mandel et al. (1992b) and Suzuki et al. (2001) teach that a transcription factor expressed in another plant species elicits the same or very similar phenotypic response of the endogenous sequence, as often predicted in earlier studies of *Arabidopsis* transcription factors in *Arabidopsis* (see Mandel et al., 1992a; and Suzuki et al., 2001). Other examples include Müller et al. (2001); Kim et al., (2001); Kyozuka and Shimamoto (2002); Boss and Thomas (2002); He et al. (2000); and Robson et al. (2001).

In yet another example, Gilmour et al. (1998) teach that an *Arabidopsis* AP2 transcription factor, CBF1, which, when overexpressed in transgenic plants, increases plant freezing tolerance. Jaglo et al. (2001) further identified sequences in *Brassica napus* which encode CBF-like genes and that transcripts for these genes accumulated rapidly in response to low temperature. Transcripts encoding CBF-like proteins were also found to accumulate rapidly in response to low temperature in wheat, as well as in tomato. An alignment of the CBF proteins from *Arabidopsis*, *B. napus*, wheat, rye, and tomato revealed the presence of conserved consecutive amino acid residues, PKK/RPAGRxKFxETRHP and DSAWR, which bracket the AP2/EREBP DNA binding domains of the proteins and distinguish them from other members of the AP2/EREBP protein family. (Jaglo et al., 2001)

Transcription factors mediate cellular responses and control traits through altered expression of genes containing cis-acting nucleotide sequences that are targets of the introduced transcription factor. It is well appreciated in the art that the effect of a transcription factor on cellular responses or a cellular trait is determined by the particular genes whose expression is either directly or indirectly (e.g., by a cascade of transcription factor binding events and transcriptional changes) altered by transcription factor binding. In a global analysis of

transcription comparing a standard condition with one in which a transcription factor is overexpressed, the resulting transcript profile associated with transcription factor overexpression is related to the trait or cellular process controlled by that transcription factor. For example, the PAP2 gene and other genes in the MYB family have been shown to control anthocyanin biosynthesis through regulation of the expression of genes known to be involved in the anthocyanin biosynthetic pathway (Bruce et al., 2000; and Borevitz et al., 2000). Further, global transcript profiles have been used successfully as diagnostic tools for specific cellular states (e.g., cancerous vs. non-cancerous; Bhattacharjee et al., 2001; and Xu et al., 2001). Consequently, it is evident to one skilled in the art that similarity of transcript profile upon overexpression of different transcription factors would indicate similarity of transcription factor function.

Polypeptides and Polynucleotides of the Invention

The present invention includes putative transcription factors (TFs), and isolated or recombinant polynucleotides encoding the polypeptides, or novel sequence variant polypeptides or polynucleotides encoding novel variants of polypeptides derived from the specific sequences provided in the Sequence Listing; the recombinant polynucleotides of the invention may be incorporated in expression vectors for the purpose of producing transformed plants. Also provided are methods for modifying yield from a plant by modifying the mass, size or number of plant organs or seed of a plant by controlling a number of cellular processes. These methods are based on the ability to alter the expression of transcription factors, critical regulatory molecules that may be conserved between diverse plant species. Related conserved regulatory molecules may be originally discovered in a model system such as *Arabidopsis* and homologous, functional molecules may then be discovered in other plant species. The latter may then be used to confer increased yield in diverse plant species.

Exemplary polynucleotides encoding the polypeptides of the invention were identified in the *Arabidopsis thaliana* GenBank database using publicly available sequence analysis programs and parameters. Sequences initially identified were then further characterized to identify sequences comprising specified sequence strings corresponding to sequence motifs present in families of known polypeptides. In addition, further exemplary polynucleotides encoding the polypeptides of the invention were identified in the plant GenBank database using publicly available sequence analysis programs and parameters. Sequences initially identified were then further characterized to identify sequences comprising specified sequence strings corresponding to sequence motifs present in families of known polypeptides.

Additional polynucleotides of the invention were identified by screening *Arabidopsis thaliana* and/or other plant cDNA libraries with probes corresponding to known polypeptides under low stringency hybridization conditions. Additional sequences, including full length coding sequences, were subsequently recovered by the rapid amplification of cDNA ends (RACE) procedure using a commercially available kit according to the manufacturer's instructions. Where necessary, multiple rounds of RACE are performed to isolate 5' and 3' ends. The full-length cDNA was then recovered by a routine end-to-end polymerase chain reaction (PCR) using primers specific to the isolated 5' and 3' ends. Exemplary sequences are provided in the Sequence Listing.

Many of the sequences in the Sequence Listing, derived from diverse plant species, have been ectopically expressed in transgenic plants. Therefore, the present polynucleotides and polypeptides can be used to change expression levels of genes, polynucleotides, and/or proteins of plants or plant cells. The changes in the characteristic(s) or trait(s) of the plants were then observed and found to confer increased growth rate and/or size.

Background Information for HAP2 and HAP5 related sequences; the role of the CCAAT-box element and CCAAT-box binding proteins

Transcriptional regulation of most eukaryotic genes occurs through the binding of transcription factors to sequence specific binding sites in their promoter regions. Many of these protein binding sites have been conserved through evolution and are found in the promoters of diverse eukaryotic organisms. One element that shows a high degree of conservation is the CCAAT-box (Gelinas et al., 1985). This cis-acting regulatory element is found in all eukaryotic species and is present in the promoter and enhancer regions of approximately 30% of genes (Bucher and Trifonov, 1988; Bucher, 1990). The CCAAT-box can function in either orientation, and operates alone, or in possible cooperation with other cis regulatory elements (Tasanen et al., 1992).

Proteins that bind the CCAAT-box element were first identified in yeast, and function as a hetero-tetrameric complex called the HAP complex (heme activator protein complex) or the CCAAT binding factor (Forsburg and Guarente, 1988). The yeast HAP complex is composed of at least four subunits, HAP2, HAP3, HAP4, and HAP5, each of which is encoded by a single gene. The yeast HAP4 polypeptide does not bind to DNA but associates with the HAP2,3,5 complex and activates transcription through an acidic domain.(Forsburg and Guarente, 1989). The yeast HAP complex has a key role in the regulation of energy metabolism. In particular, the

HAP complex is required for growth on non-fermentable carbon sources and is involved in the activation of genes involved in mitochondrial biogenesis (Mazon et al., 1982; Dang et al., 1996; Gancedo, 1998).

CCAAT binding factors of the HAP2-like, HAP3-like and HAP5-like classes are found in plant proteomes, and as in mammals, HAP4-like factors are absent (Edwards et al., 1998). In vertebrates, the three sequences of the CCAAT-binding factor are known as NF-YA, NF-YB, and NF-YC, respectively, and are homologous to HAP2, HAP3 and HAP5 subunits, respectively. In plants, the HAP2-like, HAP3-like and HAP5-like proteins are each encoded by small gene families and likely play a more complex role in regulating gene transcription than in yeast. We have identified 36 CCAAT family genes in the *Arabidopsis* genome, and these are approximately equally divided into each of the three subfamilies. In *Arabidopsis* there are 10 members of the HAP2 (NF-YA) subfamily, 12 members of the HAP3 (NF-YB) subfamily, and 11 members of the HAP5 (NF-YC) subfamily. Three additional *Arabidopsis* proteins were also identified that did not clearly fit into any of the three sub-groups, but that have some similarity to HAPs; we have designated these as HAP-like factors.

The three types of subunits in plants have the same kind of structural organization as their counterparts from mammals. For example, G481 (found in PCT patent publication WO2004076638) encodes a 141 amino acid protein of the HAP3 (NF-YB) class. In the case of the HAP3 class, the central conserved region, which confers the DNA binding and subunit interaction properties, is termed the B domain. The more variable N and C terminal regions are called the A and C domains, respectively (Li et al., 1992).

Like their mammalian counterparts, plant CCAAT binding factors most likely bind DNA as heterotrimers composed of HAP2-like, HAP3-like and HAP5-like subunits. All subunits contain regions that are required for DNA binding and subunit association. However, regions that might have an activation domain function are less apparent than in the mammalian proteins, where Q-rich regions within the HAP2 and HAP5 subunits are thought to fulfill such a role. Nonetheless, some of the HAP2 and HAP5 class proteins that we have identified do have Q-rich regions within the N and C-termini. However, these regions have not been confirmed yet as having such activation domain properties.

There is some support for the notion that HAP subunits might function in close association with other transcription factors on target promoters as part of a larger complex. This is evidenced by that fact that the CCAAT box is generally found in close proximity to other promoter elements. In particular, a HAP3-like protein from rice, OsNF-YB1, interacts with a

MADS-box protein OsMADS18 in vitro as part of a ternary complex (Masiero et al., 2002). It was also shown that the in vitro interaction between these two types of transcription factors requires that OsNF-YB1 dimerizes with a HAP5-like protein, and that OsMADS18 forms a heterodimer with another MADS-box protein. Interestingly, the OsNF-YB1/HAP5 protein dimer is incapable of interacting with HAP2-like subunits and therefore cannot bind the CCAAT element. The authors therefore speculated that there is a select set of HAP3-like proteins in plants that act on non-CCAAT promoter elements by virtue of their interaction with other non-CCAAT transcription factors (Masiero et al., 2002). In support of this, HAP3/HAP5 subunit dimers have been shown to be able to interact with TFIID in the absence of HAP2 subunits (Romier et al., 2003).

A number of phylogenetically-related sequences from diverse plant species are listed in Tables 1 - 4 for HAP2 (Tables 1 and 2) and HAP5 (Tables 3 and 4) proteins, respectively. These tables include the SEQ ID NO: (Column 1 of each table), the species from which the sequence was derived and the Gene Identifier ("GID"; Column 2 of each table), the percent identity of each polypeptide to the full length polypeptide of G929, SEQ ID NO: 2 (Table 1, Column 3), G3926, SEQ ID NO: 18 (Table 2, column 3), G3911, SEQ ID NO: 36 (Table 3, Column 3) and G3543, SEQ ID NO: 68 (Table 4, Column 3), as determined by a BLASTp analysis with a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix Henikoff & Henikoff (1989). The numbers in parentheses in Column 3 in each of these tables indicate the number of identical residues over the number of residues in the length of sequence compared in the BLAST analysis. Tables 1 - 4 also list the amino acid residue coordinates for the conserved domains, in coordinates beginning at the N-terminus of each of the sequences (Column 4 of each table) , the conserved domain sequences of the respective polypeptides (Column 5 of each table); the SEQ ID NO: of each of the conserved domains (Column 6 of each table), the percentage identity of each conserved domain in each Column 5 to the conserved domain of G929, SEQ ID NO: 85 (Table 1, Column 7), G3926, SEQ ID NO: 93 (Table 2, Column 7), G3911, SEQ ID NO: 102 (Table 3, Column 7), or G3543, SEQ ID NO: 118 (Table 4, Column 7), and in the assays performed thus far, whether a transgenic plant overexpressing the CCAAT-binding transcription factor was larger, had greater biomass, and/or a faster growth rate relative to a control plant at the seedling stage or adult stage (Column 8 of each table). Positive results are reported when more than one line (except in Tables 3 and 4 for G3886 and G3894 as noted[1]) had larger size, biomass and/or faster growth rate than wild type controls or control plants harboring an empty vector. Transgenic plants generated with the sequences in Tables 1 - 4 overexpressed the

transcription factor under the regulatory control of the constitutive CaMV 35S promoter, unless otherwise noted for certain tissue-specific promoters. "OE" refers to a transgenic plant overexpressing a CCAAT-binding transcription factor of Column 1. Species abbreviations used in these tables included: At - *Arabidopsis thaliana*; Gm - *Glycine max*; Gr - *Gossypium raimondii*; Le - *Lycopersicon esculentum*; Os - *Oryza sativa*; Zm - *Zea mays*.

At the time of evaluation, plants were given one of the following scores:

(+) Enhanced size, biomass and/or growth rate compared to controls. The response was consistent but was only moderately above the normal levels of variability observed for that assay.

(-) No detectable difference from wild-type controls, or impaired size, biomass and/or growth rate compared to controls.

(n/d) Experiment failed, data not obtained, or assay not performed.

Table 1. Sequential and functional similarity of G929-related HAP2 polypeptides and conserved domains

Col. 1 Polypeptide SEQ ID NO:	Col. 2 Species/ GID No.	Col. 3 Percent identity of polypeptide in Column 1 to G929	Col. 4 Conserved domain in amino acid coordinates	Col. 5 HAP2 conserved domain	Col. 6 SEQ ID NO: of conserved domain	Col. 7 Percent identity of conserved domain in Column 5 to conserved domain of G929	Col. 8 OE had greater size, biomass or faster growth rate
2	At/G929	100% (198/198)	98-157	EPVFVNAKQ YHGILRRRQS RAKLEARNR AIKAKKPYM HESRHLHAIR RPRGCGGRF LNAK	85	100% (60/60)	+
4	At/G2344	63% (126/197)	100-159	EPVFVNAKQ YHGILRRRQS RARLESQNK VIKSRKPYLH ESRHLHAIRR PRGCGGRFL NAK	86	86% (52/60)	+
6	At/G931	47%	172-231	EPVFVNAKQ	87	76%	+

		(73/155)		FHAIMRRRQ QRAKLEAQN KLIKARKPYL HESRHVHAL KRPRGSGGR FLNTK		(46/60)	
8	Gm/G3920	50% (69/136)	149-208	EPVYVNAKQ YHGILRRRQS RAKAEIEKK VIKNRKPYL HESRHLHAM RRARGNGGR FLNTK	88	76% (46/60)	+
10	At/G928	45% (68/151)	179-238	DPVFVNAKQ YHAIMRRRQ QRAKLEAQN KLIRARKPYL HESRHVHAL KRPRGSGGR FLNTK	89	75% (45/60)	+
12	At/G1782	58% (56/96)	178-237	EPIFVNAKQY HAILRRRKH RAKLEAQNK LIKCRKPYLH ESRHLHALK RARGSGGRF LNTK	90	75% (45/60)	+
213	Zm/G4261	55% (64/116)	175-231	EPVYVNAKQ YHGILRRRQS RAKAELEKK VVKARKPYL HESRHQHAM RRARGNGGR FL	214	75% (45/60)	n/d
14	At/G1363	42% (66/156)	171-230	EPIFVNAKQY QAILRRRERR AKLEAQNKL IKVRKPYLHE SRHLHALKR VRGSGGRFL NTK	91	73% (44/60)	+
16	Os/G3924	42% (74/174)	163-222	EPVYVNAKQ YHGILRRRQS RAKAELEKK VVKSRKPYL HESRHQHAM RRARGTGGR FLNTK	92	73% (44/60)	+
28	At/G2632	50%	166-223	EPVFVNAKQ	98	73%	-

		(67/134)		YQAILRRRQ ARAKAELEK KLIKSRKPYL HESRHQHAM RRPRGTGGR FAK		(41/56)	
18	Os/G3926	42% (79/184)	164-222	EPIFVNAKQY NAILRRRQTR AKLEAQNKA VKGRKPYLH ESRHHHAMK RARGSGGRF LTK	93	71% (43/60)	+
20	Os/G3925	42% (67/158)	138-197	EPIYVNAKQ YHAILRRRQI RAKLEAENK LVKNRKPYL HESRHQHAM KRARGTGGR FLNTK	94	71% (43/60)	+
22	Zm/G3921	41% (71/170)	148-207	EPIYVNAKQ YHAILRRRQT RAKLEAQNK MVKGRKPYL HESRHRHAM KRARGSGGR FLNTK	95	71% (43/60)	n/d
24	Zm/G3922	35% (69/193)	171-230	EPIYVNAKQ YHAILRRRQT RAKLEAQNK MVKNRKPYL HESRHRHAM KRARGSGGR FLNTK	96	71% (43/60)	n/d
26	Zm/G4264	38% (74/193)	155-214	EPIYVNAKQ YHAILRRRQT RAKLEAQNK MVKNRKPYL HESRHRHAM KRARGSGGR FLNTK	97	71% (43/60)	+
30	At/G1334	40% (60/149)	133-190	DGTIYVNSK QYHGIIRRRQ SRAKAEKLS RCRKPYMHH SRHLHAMRR PRGSGGRFL NTK	99	70% (41/58)	+
32	At/G926	44%	171-228	EPVYVNAKQ	100	66%	+[1]

		(58/131)		YEGILRRRKA RAKAELERK VIRDRKPYLH ESRHKHAMR RARASGGRF AK		(37/56)	
34	At/G927	37% (59/156)	136-199	STIYVNSKQY HGIIRRRQSR AKAAAVLDQ KKLSSRCRKP YMHHSRHLH ALRRPRGSG GRFLNTK	101	64% (40/62)	-

Table 2. Sequential and functional similarity of G3926-related HAP2 polypeptides and conserved domains

Col. 1 Polypeptide SEQ ID NO:	Col. 2 Species / GID No.	Col. 3 Percent identity of polypeptide in Column 1 to G3926	Col. 4 Conserved domain in amino acid coordinates	Col. 5 HAP2 conserved domain	Col. 6 SEQ ID NO: of conserved domain	Col. 7 Percent identity of conserved domain in Column 5 to conserved domain of G3926	Col. 8 OE had greater size, biomass or faster growth rate
18	Os/G3926	100% (317/317)	164-222	EPIFVNAKQY NAILRRRQTR AKLEAQNKA VKGRKPYLH ESRHHHAMK RARGSGGRF LTK	93	100% (59/59)	+
22	Zm/G3921	47% (143/304)	148-207	EPIYVNAKQ YHAILRRRQT RAKLEAQNK MVKGRKPYL HESRHRHAM KRARGSGGR FLNTK	95	92% (53/57)	n/d
24	Zm/G3922	47% (140/295)	171-230	EPIYVNAKQ YHAILRRRQT RAKLEAQNK MVKNRKPYL HESRHRHAM KRARGSGGR FLNTK	96	91% (52/57)	n/d

26	Zm/G4264	46% (146/311)	155-214	EPIYVNAKQYHAILRRRQTRAKLEAQNKMVKNRKPYLHESRHRHAMKRARGSGGRFLNTK	97	91% (52/57)	+
12	At/G1782	37% (89/236)	178-237	EPIFVNAKQYHAILRRRKHRAKLEAQNKLIKCRKPYLHESRHLHALKRARGSGGRFLNTK	90	85% (49/57)	+
20	Os/G3925	50% (104/204)	138-197	EPIYVNAKQYHAILRRRQLRAKLEAENKLVKNRKPYLHESRHQHAMKRARGTGGRFLNTK	94	85% (49/57)	+
14	At/G1363	37% (98/259)	171-230	EPIFVNAKQYQAILRRRERRAKLEAQNKLIKVRKPYLHESRHLHALKRVRGSGGRFLNTK	91	84% (48/57)	+
6	At/G931	37% (104/278)	172-231	EPVFVNAKQFHAIMRRRQQRAKLEAQNKLIKARKPYLHESRHVHALKRPRGSGGRFLNTK	87	80% (46/57)	+
10	At/G928	35% (94/262)	179-238	DPVFVNAKQYHAIMRRRQQRAKLEAQNKLIRARKPYLHESRHVHALKRPRGSGGRFLNTK	89	78% (45/57)	+
4	At/G2344	49% (66/134)	100-159	EPVFVNAKQYHGILRRRQSRARLESQNKVIKSRKPYLHESRHLHAIRRPRGCGGRFLNAK	86	75% (43/57)	+

16	Os/G3924	45% (76/167)	163-222	EPVYVNAKQYHGILRRRQSRAKAELEKKVVKSRKPYLHESRHQHAMRRARGTGGRFLNTK	92	75% (43/57)	+
213	Zm/G4261	47% (87/183)	175-231	EPVYVNAKQYHGILRRRQSRAKAELEKKVVKARKPYLHESRHQHAMRRARGNGGRFL	214	75% (43/57)	n/d
2	At/G929	42% (79/184)	98-157	EPVFVNAKQYHGILRRRQSRAKLEARNRAIKAKKPYMHESRHLHAIRRPRGCGGRFLNAK	85	73% (42/57)	+
28	At/G2632	40% (87/217)	166-223	EPVFVNAKQYQAILRRRQARAKAELEKKLIKSRKPYLHESRHQHAMRRPRGTGGRFAK	98	72% (43/59)	-
8	Gm/G3920	36% (76/209)	149-208	EPVYVNAKQYHGILRRRQSRAKAEIEKKVIKNRKPYLHESRHLHAMRRARGNGGRFLNTK	88	73% (42/57)	+
32	At/G926	39% (75/192)	171-228	EPVYVNAKQYEGILRRRKARAKAELERKVIRDRKPYLHESRHKHAMRRARASGGRFAK	100	67% (40/59)	+[1]
30	At/G1334	39% (54/136)	133-190	DGTIYVNSKQYHGIIRRRQSRAKAEKLSRCRKPYMHHSRHLHAMRRPRGSGGRFLNTK	99	65% (36/55)	+

34	At/G927	34% (73/213)	136-199	STIYVNSKQY HGIIRRRQSR AKAAAVLDQ KKLSSRCRKI YMHHSRHLH ALRRPRGSG GRFLNTK	101	57% (34/59)	-

Specific notes for Tables 1 and 2:

[1] Assays with 35S::G926 *Arabidopsis* plants have not yet been performed. However, 35S::G926 overexpressing tomato plants produced increased average fruit weight in the top 5% and cruciferin::G926 tomato plants produced increased average fruit weight in the top 10% of 3,217 tomato lines tested that overexpressed many different *Arabidopsis* transcription factors. *Arabidopsis* plants overexpressing G926-YFP fusion proteins (YFP or "yellow fluorescent protein" is a red-shifted spectral variant of green fluorescent protein (GFP)) were not larger and did not appear to have a faster growth rate than controls.

Table 3. Sequential and functional similarity of G3911-related HAP5 polypeptides and conserved domains

Col. 1 Polypeptide SEQ ID NO:	Col. 2 Species/ GID No.	Col. 3 Percent identity of polypeptide in Column 1 to G3911	Col. 4 Conserved domain in amino acid coordinates	Col. 5 Conserved domain	Col. 6 SEQ ID NO: of conserved domain	Col. 7 Percent identity of conserved domain in Column 5 to conserved domain of G3911	Col. 8 OE had greater size, biomass or faster growth rate
36	Zm/G3911	100% (200/200)	83-148	LPLARIKKIM KADEDVRMI AAEAPVVFA RACEMFILEL THRGWAHA EENKRRTLQ KSDIAAAIAR T	102	100% (66/66)	+
38	Os/G3546	79% (167/211)	91-156	LPLARIKKIM KADEDVRMI AAEAPVVFA RACEMFILEL THRGWAHA EENKRRTLQ	103	100% (66/66)	+

				KSDIAAAIAR T			
40	Zm/G3909	78% (159/203)	86-151	LPLARIKKIM KADEDVRMI AAEAPVVFS RACEMFILEL THRGWAHA EENKRRTLQ KSDIAAAVA RT	104	96% (64/66)	+
202	Le/G3894	54% (119/220)	103-168	LPLARIKKIM KADEDVRMI SAEAPVVFA RACEMFILEL TLRAWNHTE ENKRRTLQK NDIAAAITRT	203	89% (59/66)	+[1]
46	Gm/G3547	55% (125/226)	102-167	LPLARIKKIM KADEDVRMI SAEAPVIFAR ACEMFILELT LRSWNHTEE NKRRTLQKN DIAAAITRT	107	87% (58/66)	+
48	At/G714	72% (96/132)	71-136	LPLARIKKIM KADEDVRMIS AEAPVVFARA CEMFILELTLR SWNHTEENK RRTLQKNDIA AAVTRT	108	87% (58/66)	+
52	At/G489	58% (107/182)	81-146	LPLARIKKIM KADEDVRMIS AEAPVVFARA CEMFILELTLR SWNHTEENK RRTLQKNDIA AAVTRT	110	87% (58/66)	+[2]
42	Zm/G3552	55% (121/218)	100-165	LPLARIKKIM KADEDVRMI SAEAPVVFA KACEIFILELT LRSWMHTEE NKRRTLQKN DIAAAITRT	105	86% (57/66)	+
44	At/G483	65% (95/144)	77-142	LPLARIKKIM KADEDVRMI SAEAPVIFAK ACEMFILELT	106	86% (57/66)	-

				LRAWIHTEE NKRRTLQKN DIAAAISRT			
50	Os/G3542	54% (124/228)	106-171	LPLARIKKIM KADEDVRMIS AEAPVVFAKA CEVFILELTLR SWMHTEENK RRTLQKNDIA AAITRT	109	86% (57/66)	+
56	Gm/G3550	56% (108/191)	107-172	LPLARIKKIM KADEDVRMIS AEAPVIFAKA CEMFILELTLR SWIHTEENKR RTLQKNDIAA AISRN	112	86% (56/65)	+
58	Gm/G3548	56% (112/198)	90-155	LPLARIKKIM KADEDVRMIS AEAPVIFAKA CEMFILELTLR SWIHTEENKR RTLQKNDIAA AISRN	113	86% (56/65)	+
54	Os/G3544	56% (111/198)	102-167	LPLARIKKIM KADEDVRMIS AEAPVIFAKA CEIFILELTLRS WMHTEENKR RTLQKNDIAA AITRT	111	84% (56/66)	+
60	At/G715	54% (117/215)	66-131	LPLARIKKIM KADEDVRMIS AEAPILFAKA CELFILELTIRS WLHAEENKR RTLQKNDIAA AITRT	114	84% (56/66)	+
62	Gm/G3886	61% (114/186)	72-137	LPLARIKKIM KADEDVRMIS AEAPILFAKA CELFILELTIRS WLHAEENKR RTLQKNDIAA AITRT	115	84% (56/66)	+[1]
64	Zm/G3889	55% (114/206)	69-134	LPLARIKKIM KADEDVRMIS AEAPVLFAKA CELFILELTIRS	116	84% (56/66)	+

				WLHAEENKR RTLQRNDVA AAIART			
66	At/G1646	53% (111/206)	79-144	LPLARIKKIM KADEDVRMIS AEAPILFAKA CELFILELTIRS WLHAEENKR RTLQKNDIAA AITRT	117	84% (56/66)	+
198	Gr/G3883	61% (104/168)	67-132	LPLARIKKIM KADEDVRMIS AEAPILFAKA CELFILELTIRS WLHAEENKR RTLQKNDIAA AITRT	199	84% (56/66)	-
210	Zm/4259	57% (115/201)	70-135	LPLARIKKIM KADEDVRMIS AEAPVLFAKA CELFILELTIRS WLHAEENKR RTLQRNDVA AAIART	211	84% (56/66)	n/d
68	Os/G3543	55% (108/193)	70-135	LPLAGIKKIM KADEDVRMIS AEAPVLFAKA CELFILELTIRS WLHAEENKR RTLQRKDVA AAIART	118	83% (55/66)	+
70	At/G1820	52% (76/145)	55-120	LPLARIKKIM KADPDVHMV SAEAPIIFAKA CEMFIVDLTM RSWLKAEEN KRHTLQKSDI SNAVASS	119	73% (47/64)	-
72	At/G1836	53% (65/122)	37-102	LPITRIKKIMK YDPDVTMIAS EAPILLSKACE MFIMDLTMRS WLHAQESKR VTLQKSNVDA AVAQT	120	63% (42/66)	+
74	At/G1819	37% (64/169)	64-135	FPLTRIKKIMK SNPEVNMVTA EAPVLISKACE MLILDLTMRS	121	52% (37/71)	+

				WLHTVEGGR QTLKRSDTLT RSDISAATTRS			
76	At/G1818	47% (57/119)	38-102	PISRIKRIMKF DPDVSMIAAE APNLLSKACE MFVMDLTMR SWLHAQESNR LTIRKSDVDA VVSQT	122	55% (36/65)	+
78	At/G490	41% (41/99)	68-133	LPLSRVRKILK SDPEVKKISC DVPALFSKAC EYFILEVTLRA WMHTQSCTR ETIRRCDIFQA VKNS	123	44% (28/63)	+
80	At/G3074	38% (30/77)	9-73	FPAARIKKIM QADEDVGKIA LAVPVLVSKS LELFLQDLCD RTYEITLERG AKTVSSLHLK HCVER	124	37% (24/64)	+
82	At/G1249	35% (27/77)	12-76	FPIGRVKKIM KLDKDINKIN SEALHVITYST ELFLHFLAEK SAVVTAEKKR KTVNLDHLRI AVKR	125	34% (22/64)	+
84	At/G3075	25% (19/76)	110-173	FPMNRIRRIM RSDNSAPQIM QDAVFLVNK ATEMFIERFSE EAYDSSVKDK KKFIHYKHLS SVVS	126	22% (14/63)	-

Table 4. Sequential and functional similarity of G3543-related HAP5 polypeptides and conserved domains

Col. 1 Polypeptide SEQ ID NO:	Col. 2 Species / GID No.	Col. 3 Percent identity of polypeptide in Column 1 to G3543	Col. 4 Conserved domain in amino acid coordinates	Col. 5 Conserved domain	Col. 6 SEQ ID NO: of conserved domain	Col. 7 Percent identity of conserved in Column 5 to conserved domain of G3543	Col. 8 OE had greater size, biomass or faster growth rate
68	Os/G3543	100% (246/246)	70-135	LPLAGIKKIM KADEDVRMI SAEAPVLFA KACELFILEL TIRSWLHAEE NKRRTLQRK DVAAAIART	118	100% (66/66)	+
210	Zm/4259	87% (219/251)	70-135	LPLARIKKIM KADEDVRMI SAEAPVLFA KACELFILEL TIRSWLHAEE NKRRTLQRN DVAAAIART	211	96% (64/66	n/d
64	Zm/G3889	86% (218/251)	69-134	LPLARIKKIM KADEDVRMI SAEAPVLFA KACELFILEL TIRSWLHAEE NKRRTLQRN DVAAAIART	116	96% (64/66)	+
60	At/G715	58% (144/248)	66-131	LPLARIKKIM KADEDVRMI SAEAPILFAK ACELFILELTI RSWLHAEEN KRRTLQKND IAAAITRT	114	90% (60/66)	+
62	Gm/G3886	58% (142/243)	72-137	LPLARIKKIM KADEDVRMI SAEAPILFAK ACELFILELTI RSWLHAEEN KRRTLQKND IAAAITRT	115	90% (60/66)	+[1]
66	At/G1646	56% (143/253)	79-144	LPLARIKKIM KADEDVRMI SAEAPILFAK	117	90% (60/66)	+

				ACELFILELTI RSWLHAEEN KRRTLQKND IAAAITRT			
198	Gr/G3883	58% (142/244)	67-132	LPLARIKKIM KADEDVRMI SAEAPILFAK ACELFILELTI RSWLHAEEN KRRTLQKND IAAAITRT	199	90% (60/66)	-
56	Gm/G3550	63% (98/154)	107-172	LPLARIKKIM KADEDVRMI SAEAPVIFAK ACEMFILELT LRSWIHTEEN KRRTLQKND IAAAISRN	112	84% (55/65)	+
58	Gm/G3548	62% (97/154)	90-155	LPLARIKKIM KADEDVRMI SAEAPVIFAK ACEMFILELT LRSWIHTEEN KRRTLQKND IAAAISRN	113	84% (55/65)	+
50	Os/G3542	55% (113/205)	106-171	LPLARIKKIM KADEDVRMI SAEAPVVFA KACEVFILEL TLRSWMHTE ENKRRTLQK NDIAAAITRT	109	84% (56/66)	+
42	Zm/G3552	56% (110/194)	100-165	LPLARIKKI MKADEDVR MISAEAPVV FAKACEIFIL ELTLRSWM HTEENKRRT LQKNDIAAA ITRT	105	84% (56/66)	+
54	Os/G3544	54% (108/198)	102-167	LPLARIKKIM KADEDVRMI SAEAPVIFAK ACEIFILELTL RSWMHTEEN KRRTLQKND IAAAITRT	111	84% (56/66)	+
44	At/G483	65% (105/161)	77-142	LPLARIKKI MKADEDVR	106	83% (55/66)	-

				MISAEAPVI FAKACEMFI LELTLRAWI HTEENKRRT LQKNDIAAA ISRT			
46	Gm/G3547	53% (117/220)	102-167	LPLARIKKI MKADEDVR MISAEAPVI FARACEMFI LELTLRSWN HTEENKRRT LQKNDIAAA ITRT	107	83% (55/66)	+
36	Zm/G3911	55% (108/193)	83-148 -	LPLARIKKI MKADEDVR MIAAEAPVV FARACEMFI LELTHRGW AHAEENKR RTLQKSDIA AAIART	102	83% (55/66)	+
38	Os/G3546	56% (110/194)	91-156	LPLARIKKI MKADEDVR MIAAEAPVV FARACEMFI LELTHRGW AHAEENKR RTLQKSDIA AAIART	103	83% (55/66)	+
48	At/G714	57% (106/185)	71-136	LPLARIKKIM KADEDVRMI SAEAPVVFA RACEMFILEL TLRSWNHTE ENKRRTLQK NDIAAAVTR T	108	81% (54/66)	+
52	At/G489	53% (117/220)	81-146	LPLARIKKIM KADEDVRMI SAEAPVVFA RACEMFILEL TLRSWNHTE ENKRRTLQK NDIAAAVTR T	110	81% (54/66)	+[2]
202	Le/G3894	59% (108/182)	103-168	LPLARIKKI MKADEDVR MISAEAPVV	203	81% (54/64)	+[1]

				FARACEMFI LELTLRAW NHTEENKR RTLQKNDIA AAITRT			
40	Zm/G3909	55% (108/193)	86-151	LPLARIKKI MKADEDVR MIAAEAPVV FSRACEMFI LELTHRGW AHAEENKR RTLQKSDIA AAVART	104	80% (53/66)	+
70	At/G1820	43% (85/195)	55-120	LPLARIKKIM KADPDVHM VSAEAPIIFA KACEMFIVD LTMRSWLKA EENKRHTLQ KSDISNAVAS S	119	71% (46/64)	-
72	At/G1836	46% (72/154)	37-102	LPITRIKKIM KYDPDVTMI ASEAPILLSK ACEMFIMDL TMRSWLHAQ ESKRVTLQK SNVDAAVAQ T	120	63% (42/66)	+
74	At/G1819	38% (67/174)	64-135	FPLTRIKKIM KSNPEVNMV TAEAPVLISK ACEMLILDLT MRSWLHTVE GGRQTLKRS DTLTRSDISA ATTRS	121	61% (35/57)	+
76	At/G1818	35% (70/195)	38-102	PISRIKRIMKF DPDVSMIAA EAPNLLSKA CEMFVMDLT MRSWLHAQE SNRLTIRKSD VDAVVSQT	122	55% (36/65)	+
78	At/G490	40% (41/101)	68-133	LPLSRVRKIL KSDPEVKKIS CDVPALFSK ACEYFILEVT LRAWMHTQS	123	47% (30/63)	+

				CTRETIRRCD IFQAVKNS			
80	At/G3074	39% (31/79)	9-73	FPAARIKKIM QADEDVGKI ALAVPVLVS KSLELFLQDL CDRTYEITLE RGAKTVSSL HLKHCVER	124	39% (25/64)	+
82	At/G1249	34% (27/79)	12-76	FPIGRVKKIM KLDKDINKIN SEALHVITYS TELFLHFLAE KSAVVTAEK KRKTVNLDH LRIAVKR	125	34% (22/64)	+
84	At/G3075	26% (27/101)	110-173	FPMNRIRRIM RSDNSAPQIM QDAVFLVNK ATEMFIERFS EEAYDSSVK DKKKFIHYK HLSSVVS	126	23% (15/63)	-

Specific notes for Tables 3 and 4:

[1] One of ten lines had larger seedlings

[2] Numerous plants overexpressing G489-YFP fusion proteins had larger rosettes than controls; YFP or "yellow fluorescent protein is a red-shifted spectral variant of green fluorescent protein (GFP)

Orthologs and Paralogs

Homologous sequences as described above can comprise orthologous or paralogous sequences. Several different methods are known by those of skill in the art for identifying and defining these functionally homologous sequences. General methods for identifying orthologs and paralogs, including phylogenetic methods, sequence similarity and hybridization methods, are described herein; an ortholog or paralog, including equivalogs, may be identified by one or more of the methods described below.

As described by Eisen (1998), evolutionary information may be used to predict gene function. It is common for groups of genes that are homologous in sequence to have diverse, although usually related, functions. However, in many cases, the identification of homologs is not sufficient to make specific predictions because not all homologs have the same function. Thus, an initial analysis of functional relatedness based on sequence similarity alone may not

provide one with a means to determine where similarity ends and functional relatedness begins. Fortunately, it is well known in the art that protein function can be classified using phylogenetic analysis; functional predictions can be greatly improved by focusing on how the genes became similar in sequence, i.e., by evolutionary processes, rather than on the sequence similarity itself (Eisen, 1998). In fact, many specific examples exist in which gene function has been shown to correlate well with gene phylogeny (Eisen, 1998). Thus, "[t]he first step in making functional predictions is the generation of a phylogenetic tree representing the evolutionary history of the gene of interest and its homologs. Such trees are distinct from clusters and other means of characterizing sequence similarity because they are inferred by techniques that help convert patterns of similarity into evolutionary relationships .... After the gene tree is inferred, biologically determined functions of the various homologs are overlaid onto the tree. Finally, the structure of the tree and the relative phylogenetic positions of genes of different functions are used to trace the history of functional changes, which is then used to predict functions of [as yet] uncharacterized genes" (Eisen, 1998).

Within a single plant species, gene duplication may cause two copies of a particular gene, giving rise to two or more genes with similar sequence and often similar function known as paralogs. A paralog is therefore a similar gene formed by duplication within the same species. Paralogs typically cluster together or in the same clade (a group of similar genes) when a gene family phylogeny is analyzed using programs such as CLUSTAL (Thompson et al., 1994; Higgins et al., 1996). Groups of similar genes can also be identified with pair-wise BLAST analysis (Feng and Doolittle, 1987). For example, a clade of very similar MADS domain transcription factors from *Arabidopsis* all share a related function in flowering time (Ratcliffe et al., 2001, 2003), and a group of very similar AP2 domain transcription factors from *Arabidopsis* are involved in tolerance of plants to freezing (Gilmour et al., 1998). Analysis of groups of similar genes with similar function that fall within one clade can yield sub-sequences that are particular to the clade. These sub-sequences, known as consensus sequences, can not only be used to define the sequences within each clade, but define the functions of these genes; genes within a clade may contain paralogous sequences, or orthologous sequences that share the same function (see also, for example, Mount, 2001).

Transcription factor gene sequences are conserved across diverse eukaryotic species lines (Goodrich et al., 1993; Lin et al., 1991; Sadowski et al., 1988). Plants are no exception to this observation; diverse plant species possess transcription factors that have similar sequences and functions. Speciation, the production of new species from a parental species, gives rise to two or

more genes with similar sequence and similar function. These genes, termed orthologs, often have an identical function within their host plants and are often interchangeable between species without losing function. Because plants have common ancestors, many genes in any plant species will have a corresponding orthologous gene in another plant species. Once a phylogenic tree for a gene family of one species has been constructed using a program such as CLUSTAL (Thompson et al., 1994; Higgins et al., 1996) potential orthologous sequences can be placed into the phylogenetic tree and their relationship to genes from the species of interest can be determined. Orthologous sequences can also be identified by a reciprocal BLAST strategy. Once an orthologous sequence has been identified, the function of the ortholog can be deduced from the identified function of the reference sequence..

By using a phylogenetic analysis, one skilled in the art would recognize that the ability to predict similar functions conferred by closely-related polypeptides is predictable. This predictability has been confirmed by our own many studies in which we have found that a wide variety of polypeptides have orthologous or closely-related homologous sequences that function as does the first, closely-related reference sequence. For example, distinct transcription factors, including:

(i) AP2 family *Arabidopsis* G47 (found in US patent publication 20040019925A1), a phylogenetically-related sequence from soybean, and two phylogenetically-related homologs from rice all conferred greater tolerance to drought, hyperosmotic stress, or delayed flowering in transgenic plants as compared to control plants;

(ii) CCAAT family and HAP3 *Arabidopsis* G481 (found in PCT patent publication WO2004076638), and numerous phylogenetically-related sequences from dicots and monocots conferred greater tolerance to drought-related stress as compared to control plants;

(iii) Myb-related Arabidopsis G682 (found in PCT patent publication WO2004076638) and numerous phylogenetically-related sequences from dicots and monocots conferred greater tolerance to heat , drought-related stress, cold , and salt as compared to control plants;

(iv) WRKY family *Arabidopsis* G1274 (found in US patent application 10/666,642) and numerous closely-related sequences from dicots and monocots have been shown to confer increased water deprivation tolerance, and

(v) AT-hook family soy sequence G3456 (found in US patent publication 20040128712A1) and numerous phylogenetically-related sequences from dicots and monocots, increased biomass compared to control plants when these sequences were overexpressed in plants.

The polypeptide sequences belong to distinct clades of polypeptides that include members from diverse species. In each case, most or all of the clade member sequences derived from both dicots and monocots have been shown to confer increased yield or tolerance to one or more abiotic stresses when the sequences were overexpressed. These studies and others demonstrate that evolutionarily conserved genes from diverse species are likely to function similarly (i.e., by regulating similar target sequences and controlling the same traits), and that polynucleotides from one species may be transformed into closely-related or distantly-related plant species to confer or improve traits.

As shown in Tables 1 - 4, polypeptides that are phylogenetically related homologs of the polypeptides of the invention may have conserved domains that share at least 34%, at least 37%, at least 39%, at least 44%, at least 47%, at least 52%, at least 55%, at least 61%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 70%, at least 71%, at least 72%, at least 73%, at least 75%, at least 76%, at least 78%, at least 80%, at least 81%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 89%, at least 90%, at least 91%, at least 92%, at least 96%, or 100% amino acid sequence identity to similar conserved domains of any of SEQ ID NO: 85-126, 199, 203, 211, or 214, and have similar functions in that the polypeptides of the invention may, when overexpressed, confer at least one regulatory activity selected from the group consisting of greater yield, more rapid growth, greater size, and increased biomass as compared to a control plant.

At the nucleotide level, the sequences of the invention will typically share at least about 30% or 35% nucleotide sequence identity, or 40% nucleotide sequence identity, preferably at least about 50%, or about 60%, or about 70% or about 80% sequence identity, or more preferably about 85%, or about 90%, or about 95% or about 97% or more sequence identity to one or more of the listed full-length sequences, or to a listed sequence but excluding or outside of the region(s) encoding a known consensus sequence or consensus DNA-binding site, or outside of the region(s) encoding one or all conserved domains. The degeneracy of the genetic code enables major variations in the nucleotide sequence of a polynucleotide while maintaining the amino acid sequence of the encoded protein.

Percent identity can be determined electronically, e.g., by using the MEGALIGN program (DNASTAR, Inc. Madison, Wis.). The MEGALIGN program can create alignments between two or more sequences according to different methods, for example, the clustal method (see, for example, Higgins and Sharp (1988). The clustal algorithm groups sequences into clusters by examining the distances between all pairs. The clusters are aligned pairwise and then

in groups. Other alignment algorithms or programs may be used, including FASTA, or BLAST, and which may be used to calculate percent similarity. These are available as a part of the GCG sequence analysis package (University of Wisconsin, Madison, WI), and can be used with or without default settings. In one embodiment, the percent identity of two sequences can be determined by the GCG program with a gap weight of 1, e.g., each amino acid gap is weighted as if it were a single amino acid or nucleotide mismatch between the two sequences (see USPN 6,262,333).

Software for performing BLAST analyses is publicly available, e.g., through the National Center for Biotechnology Information (see internet website at http://www.ncbi.nlm.nih.gov/). This 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, 1990; Altschul et al., 1993). These initial neighborhood word hits act 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 wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (*see* Henikoff & Henikoff, 1989). Unless otherwise indicated for comparisons of predicted polynucleotides, "sequence identity" refers to the % sequence identity generated from a tblastx using the NCBI version of the algorithm at the default settings using gapped alignments with the filter "off" (see, for example, internet website at http://www.ncbi.nlm.nih.gov/).

Other techniques for alignment are described by Doolittle (1996). Preferably, an alignment program that permits gaps in the sequence is utilized to align the sequences. The

Smith-Waterman is one type of algorithm that permits gaps in sequence alignments (see Shpaer (1997). Also, the GAP program using the Needleman and Wunsch alignment method can be utilized to align sequences. An alternative search strategy uses MPSRCH software, which runs on a MASPAR computer. MPSRCH uses a Smith-Waterman algorithm to score sequences on a massively parallel computer. This approach improves ability to pick up distantly related matches, and is especially tolerant of small gaps and nucleotide sequence errors. Nucleic acid-encoded amino acid sequences can be used to search both protein and DNA databases.

The percentage similarity between two polypeptide sequences, e.g., sequence A and sequence B, is calculated by dividing the length of sequence A, minus the number of gap residues in sequence A, minus the number of gap residues in sequence B, into the sum of the residue matches between sequence A and sequence B, times one hundred. Gaps of low or of no similarity between the two amino acid sequences are not included in determining percentage similarity.

Percent identity between polynucleotide sequences can also be counted or calculated by other methods known in the art, e.g., the Jotun Hein method (see, for example, Hein, 1990). Identity between sequences can also be determined by other methods known in the art, e.g., by varying hybridization conditions (see US Patent Application No. 20010010913).

Thus, the invention provides methods for identifying a sequence similar or paralogous or orthologous or homologous to one or more polynucleotides as noted herein, or one or more target polypeptides encoded by the polynucleotides, or otherwise noted herein and may include linking or associating a given plant phenotype or gene function with a sequence. In the methods, a sequence database is provided (locally or across an internet or intranet) and a query is made against the sequence database using the relevant sequences herein and associated plant phenotypes or gene functions.

In addition, one or more polynucleotide sequences or one or more polypeptides encoded by the polynucleotide sequences may be used to search against a BLOCKS (Bairoch et al., 1997), PFAM, and other databases which contain previously identified and annotated motifs, sequences and gene functions. Methods that search for primary sequence patterns with secondary structure gap penalties (Smith et al., 1992) as well as algorithms such as Basic Local Alignment Search Tool (BLAST; Altschul, 1990; Altschul et al., 1993), BLOCKS (Henikoff and Henikoff, 1991), Hidden Markov Models (HMM; Eddy, 1996; Sonnhammer et al., 1997), and the like, can be used to manipulate and analyze polynucleotide and polypeptide sequences

encoded by polynucleotides. These databases, algorithms and other methods are well known in the art and are described in Ausubel et al. (1997) and in Meyers (1995).

A further method for identifying or confirming that specific homologous sequences control the same function is by comparison of the transcript profile(s) obtained upon overexpression or knockout of two or more related polypeptides. Since transcript profiles are diagnostic for specific cellular states, one skilled in the art will appreciate that genes that have a highly similar transcript profile (e.g., with greater than 50% regulated transcripts in common, or with greater than 70% regulated transcripts in common, or with greater than 90% regulated transcripts in common) will have highly similar functions. Fowler and Thomashow (2002) have shown that three paralogous AP2 family genes (CBF1, CBF2 and CBF3) are induced upon cold treatment, and each of which can condition improved freezing tolerance, and all have highly similar transcript profiles. Once a polypeptide has been shown to provide a specific function, its transcript profile becomes a diagnostic tool to determine whether paralogs or orthologs have the same function.

Furthermore, methods using manual alignment of sequences similar or homologous to one or more polynucleotide sequences or one or more polypeptides encoded by the polynucleotide sequences may be used to identify regions of similarity and conserved (e.g., CCAAT binding) domains. Such manual methods are well-known by those of skill in the art and can include, for example, comparisons of tertiary structure between a polypeptide sequence encoded by a polynucleotide that comprises a known function and a polypeptide sequence encoded by a polynucleotide sequence that has a function not yet determined. Such examples of tertiary structure may comprise predicted alpha helices, beta-sheets, amphipathic helices, leucine zipper motifs, zinc finger motifs, proline-rich regions, cysteine repeat motifs, and the like.

Orthologs and paralogs of presently disclosed polypeptides may be cloned using compositions provided by the present invention according to methods well known in the art. cDNAs can be cloned using mRNA from a plant cell or tissue that expresses one of the present sequences. Appropriate mRNA sources may be identified by interrogating Northern blots with probes designed from the present sequences, after which a library is prepared from the mRNA obtained from a positive cell or tissue. Polypeptide-encoding cDNA is then isolated using, for example, PCR, using primers designed from a presently disclosed gene sequence, or by probing with a partial or complete cDNA or with one or more sets of degenerate probes based on the disclosed sequences. The cDNA library may be used to transform plant cells. Expression of the cDNAs of interest is detected using, for example, microarrays, Northern blots, quantitative PCR,

or any other technique for monitoring changes in expression. Genomic clones may be isolated using similar techniques to those.

Examples of *Arabidopsis* polypeptide sequences and functionally similar and phylogenetically-related sequences are listed in Tables 1 - 4 and the Sequence Listing. In addition to the sequences in Tables 1 - 4 and the Sequence Listing, the invention encompasses isolated nucleotide sequences that are phylogenetically and structurally similar to sequences listed in the Sequence Listing) and can function in a plant by increasing yield and/or and abiotic stress tolerance when ectopically expressed in a plant.

Since a significant number of these sequences are phylogenetically and sequentially related to each other and have been shown to increase yield from a plant and/or abiotic stress tolerance, one skilled in the art would predict that other similar, phylogenetically related sequences falling within the present clades of polypeptides would also perform similar functions when ectopically expressed.

Identifying Polynucleotides or Nucleic Acids by Hybridization

Polynucleotides homologous to the sequences illustrated in the Sequence Listing and tables can be identified, e.g., by hybridization to each other under stringent or under highly stringent conditions. Single stranded polynucleotides hybridize when they associate based on a variety of well characterized physical-chemical forces, such as hydrogen bonding, solvent exclusion, base stacking and the like. The stringency of a hybridization reflects the degree of sequence identity of the nucleic acids involved, such that the higher the stringency under which two polynucleotide strands hybridize, the more similar are the two strands. Stringency is influenced by a variety of factors, including temperature, salt concentration and composition, organic and non-organic additives, solvents, etc. present in both the hybridization and wash solutions and incubations (and number thereof), as described in more detail in the references cited below (e.g., Sambrook et al., 1989; Berger and Kimmel, 1987; and Anderson and Young, 1985).

Encompassed by the invention are polynucleotide sequences that are capable of hybridizing to the claimed polynucleotide sequences, including any of the polynucleotides within the Sequence Listing, and fragments thereof under various conditions of stringency (see, for example, Wahl and Berger, 1987; and Kimmel, 1987). In addition to the nucleotide sequences listed in the Sequence Listing, full length cDNA, orthologs, and paralogs of the present nucleotide sequences may be identified and isolated using well-known methods. The

cDNA libraries, orthologs, and paralogs of the present nucleotide sequences may be screened using hybridization methods to determine their utility as hybridization target or amplification probes.

With regard to hybridization, conditions that are highly stringent, and means for achieving them, are well known in the art (see, for example, Sambrook et al., 1989; Berger, 1987 , pages 467-469; and Anderson and Young, 1985).

Stability of DNA duplexes is affected by such factors as base composition, length, and degree of base pair mismatch. Hybridization conditions may be adjusted to allow DNAs of different sequence relatedness to hybridize. The melting temperature I is defined as the temperature when 50% of the duplex molecules have dissociated into their constituent single strands. The melting temperature of a perfectly matched duplex, where the hybridization buffer contains formamide as a denaturing agent, may be estimated by the following equations:

(I) DNA-DNA:

$T_m(^\circ C)=81.5+16.6(\log [Na+])+0.41(\%\ G+C)- 0.62(\%\ \text{formamide})-500/L$

(II) DNA-RNA:

$T_m(^\circ C)=79.8+18.5(\log [Na+])+0.58(\%\ G+C)+ 0.12(\%G+C)^2 - 0.5(\%\ \text{formamide}) - 820/L$

(III) RNA-RNA:

$T_m(^\circ C)=79.8+18.5(\log [Na+])+0.58(\%\ G+C)+ 0.12(\%G+C)^2 - 0.35(\%\ \text{formamide}) - 820/L$

where L is the length of the duplex formed, [Na+] is the molar concentration of the sodium ion in the hybridization or washing solution, and % G+C is the percentage of (guanine+cytosine) bases in the hybrid. For imperfectly matched hybrids, approximately 1° C is required to reduce the melting temperature for each 1% mismatch.

Hybridization experiments are generally conducted in a buffer of pH between 6.8 to 7.4, although the rate of hybridization is nearly independent of pH at ionic strengths likely to be used in the hybridization buffer (Anderson and Young, 1985). In addition, one or more of the following may be used to reduce non-specific hybridization: sonicated salmon sperm DNA or another non-complementary DNA, bovine serum albumin, sodium pyrophosphate, sodium

dodecyl sulfate (SDS), polyvinyl-pyrrolidone, ficoll and Denhardt's solution. Dextran sulfate and polyethylene glycol 6000 act to exclude DNA from solution, thus raising the effective probe DNA concentration and the hybridization signal within a given unit of time. In some instances, conditions of even greater stringency may be desirable or required to reduce non-specific and/or background hybridization. These conditions may be created with the use of higher temperature, lower ionic strength and higher concentration of a denaturing agent such as formamide.

Stringency conditions can be adjusted to screen for moderately similar fragments such as homologous sequences from distantly related organisms, or to highly similar fragments such as genes that duplicate functional enzymes from closely related organisms. The stringency can be adjusted either during the hybridization step or in the post-hybridization washes. Salt concentration, formamide concentration, hybridization temperature and probe lengths are variables that can be used to alter stringency (as described by the formula above). As a general guideline, high stringency is typically performed at T_m-5° C to T_m-20° C, moderate stringency at T_m-20° C to T_m-35° C and low stringency at T_m-35° C to T_m-50° C for duplex >150 base pairs. Hybridization may be performed at low to moderate stringency (25-50° C below T_m), followed by post-hybridization washes at increasing stringencies. Maximum rates of hybridization in solution are determined empirically to occur at T_m-25° C for DNA-DNA duplex and T_m-15° C for RNA-DNA duplex. Optionally, the degree of dissociation may be assessed after each wash step to determine the need for subsequent, higher stringency wash steps.

High stringency conditions may be used to select for nucleic acid sequences with high degrees of identity to the disclosed sequences. An example of stringent hybridization conditions obtained in a filter-based method such as a Southern or Northern blot for hybridization of complementary nucleic acids that have more than 100 complementary residues is about 5°C to 20°C lower than the thermal melting point I for the specific sequence at a defined ionic strength and pH. Conditions used for hybridization may include about 0.02 M to about 0.15 M sodium chloride, about 0.5% to about 5% casein, about 0.02% SDS or about 0.1% N-laurylsarcosine, about 0.001 M to about 0.03 M sodium citrate, at hybridization temperatures between about 50° C and about 70° C. More preferably, high stringency conditions are about 0.02 M sodium chloride, about 0.5% casein, about 0.02% SDS, about 0.001 M sodium citrate, at a temperature of about 50° C. Nucleic acid molecules that hybridize under stringent conditions will typically hybridize to a probe based on either the entire DNA molecule or selected portions, e.g., to a unique subsequence, of the DNA.

Stringent salt concentration will ordinarily be less than about 750 mM NaCl and 75 mM trisodium citrate. Increasingly stringent conditions may be obtained with less than about 500 mM NaCl and 50 mM trisodium citrate, to even greater stringency with less than about 250 mM NaCl and 25 mM trisodium citrate. Low stringency hybridization can be obtained in the absence of organic solvent, e.g., formamide, whereas high stringency hybridization may be obtained in the presence of at least about 35% formamide, and more preferably at least about 50% formamide. Stringent temperature conditions will ordinarily include temperatures of at least about 30° C, more preferably of at least about 37° C, and most preferably of at least about 42° C with formamide present. Varying additional parameters, such as hybridization time, the concentration of detergent, e.g., sodium dodecyl sulfate (SDS) and ionic strength, are well known to those skilled in the art. Various levels of stringency are accomplished by combining these various conditions as needed.

The washing steps that follow hybridization may also vary in stringency; the post-hybridization wash steps primarily determine hybridization specificity, with the most critical factors being temperature and the ionic strength of the final wash solution. Wash stringency can be increased by decreasing salt concentration or by increasing temperature. Stringent salt concentration for the wash steps will preferably be less than about 30 mM NaCl and 3 mM trisodium citrate, and most preferably less than about 15 mM NaCl and 1.5 mM trisodium citrate.

Thus, hybridization and wash conditions that may be used to bind and remove polynucleotides with less than the desired homology to the nucleic acid sequences or their complements that encode the present polypeptides include, for example:

6X SSC at 65° C;

50% formamide, 4X SSC at 42° C; or

0.5X SSC, 0.1% SDS at 65° C;

with, for example, two wash steps of 10 - 30 minutes each. Useful variations on these conditions will be readily apparent to those skilled in the art.

A person of skill in the art would not expect substantial variation among polynucleotide species encompassed within the scope of the present invention because the highly stringent conditions set forth in the above formulae yield structurally similar polynucleotides.

If desired, one may employ wash steps of even greater stringency, including about 0.2x SSC, 0.1% SDS at 65° C and washing twice, each wash step being about 30 minutes, or about 0.1 x SSC, 0.1% SDS at 65° C and washing twice for 30 minutes. The temperature for the wash

solutions will ordinarily be at least about 25° C, and for greater stringency at least about 42° C. Hybridization stringency may be increased further by using the same conditions as in the hybridization steps, with the wash temperature raised about 3° C to about 5° C, and stringency may be increased even further by using the same conditions except the wash temperature is raised about 6° C to about 9° C. For identification of less closely related homologs, wash steps may be performed at a lower temperature, e.g., 50° C.

An example of a low stringency wash step employs a solution and conditions of at least 25° C in 30 mM NaCl, 3 mM trisodium citrate, and 0.1% SDS over 30 minutes. Greater stringency may be obtained at 42° C in 15 mM NaCl, with 1.5 mM trisodium citrate, and 0.1% SDS over 30 minutes. Even higher stringency wash conditions are obtained at 65° C -68° C in a solution of 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. Wash procedures will generally employ at least two final wash steps. Additional variations on these conditions will be readily apparent to those skilled in the art (see, for example, US Patent Application No. 20010010913).

Stringency conditions can be selected such that an oligonucleotide that is perfectly complementary to the coding oligonucleotide hybridizes to the coding oligonucleotide with at least about a 5-10x higher signal to noise ratio than the ratio for hybridization of the perfectly complementary oligonucleotide to a nucleic acid encoding a polypeptide known as of the filing date of the application. It may be desirable to select conditions for a particular assay such that a higher signal to noise ratio, that is, about 15x or more, is obtained. Accordingly, a subject nucleic acid will hybridize to a unique coding oligonucleotide with at least a 2x or greater signal to noise ratio as compared to hybridization of the coding oligonucleotide to a nucleic acid encoding known polypeptide. The particular signal will depend on the label used in the relevant assay, e.g., a fluorescent label, a colorimetric label, a radioactive label, or the like. Labeled hybridization or PCR probes for detecting related polynucleotide sequences may be produced by oligolabeling, nick translation, end-labeling, or PCR amplification using a labeled nucleotide.

EXAMPLES

It is to be understood that this invention is not limited to the particular devices, machines, materials and methods described. Although particular embodiments are described, equivalent embodiments may be used to practice the invention.

The invention, now being generally described, will be more readily understood by reference to the following examples, which are included merely for purposes of illustration of

certain aspects and embodiments of the present invention and are not intended to limit the invention. It will be recognized by one of skill in the art that a polypeptide that is associated with a particular first trait may also be associated with at least one other, unrelated and inherent second trait that was not predicted by the first trait.

Example I. Project Types and Vector and Cloning Information

A number of constructs were used to modulate the activity of sequences of the invention. An individual project was defined as the analysis of transgenic plant lines for a particular construct (for example, this might include G929, G3926, G3911 or G3543 lines that constitutively overexpressed a sequence of the invention). In the present study, each gene was directly fused to a promoter that drove its expression in transgenic plants. Such a promoter could be the native promoter of that gene, or the cauliflower mosaic 35S promoter. Alternatively, a promoter that drives tissue specific or conditional expression could be used in similar studies.

In the present study, expression of a given polynucleotide from a particular promoter was achieved by either a direct-promoter fusion construct in which that sequence was cloned directly behind the promoter of interest, or a two-component system, described below. A direct fusion approach has the advantage of allowing for simple genetic analysis if a given promoter-polynucleotide line is to be crossed into different genetic backgrounds at a later date. The two-component method potentially allows for stronger expression to be obtained via an amplification of transcription.

For the two-component system, two separate constructs were used: Promoter::LexA-GAL4TA and opLexA::TF. The first of these (Promoter::LexA-GAL4TA) comprised a desired promoter (for example, the floral meristem-specific AP1 promoter, the epidermis and vascular tissue-specific LTP1 promoter, the shoot apical meristem-specific STM promoter, or the embryo-, endosperm-, and fruit-specific cruciferin promoter (SEQ ID NOs: 191, 193, 208 or 205, respectively) cloned in front of a LexA DNA binding domain fused to a GAL4 activation domain. The construct vector backbone (pMEN48, SEQ ID NO: 195) also carried a kanamycin resistance marker along with an opLexA::GFP reporter. Transgenic lines were obtained containing this first component, and a line was selected that showed reproducible expression of the reporter gene in the desired pattern through a number of generations. A homozygous population was established for that line, and the population was supertransformed with the second construct (opLexA::TF) carrying the transcription factor of interest cloned behind a LexA operator site, for example, G1819 (SEQ ID NO: 192), G2344 (SEQ ID NO: 194) or G929

(SEQ ID NO: 206). The backbone of these second construct vectors (pMEN53, SEQ ID NO: 196) also contained a sulfonamide resistance marker. After supertransformation, the LexA-GAL4 transcript was translated, and the resulting fusion protein activated the second component construct causing transcription of the transcription factor of interest.

For analysis of HAP2- or HAP5-overexpressing plants, transgenic lines were created with an expression vector, for example, P399 (SEQ ID NO: 127) or P26600 (SEQ ID NO: 135) containing HAP2 DNA clones, or P26591 (SEQ ID NO: 143) or P26598 (SEQ ID NO: 159) which contained HAP5 cDNA clones. These constructs constituted 35S::G929, 35S::G3926, 35S::G3911 or 35S::G3543 direct promoter-fusions, respectively in these examples, each carrying a kanamycin resistance marker. The constructs were introduced into *Arabidopsis* plants as indicated in following Examples.

A list of constructs (PIDs), indicating the promoter fragment that was used to drive the transgene, along with the cloning vector backbone, is provided in Table 5. Compilations of the sequences of promoter fragments and the expressed transgene sequences within the PIDs are provided in the Sequence Listing.

Table 5. Expression constructs, sequences of promoter fragments and the expressed transgene sequences

Gene Identifier	Construct (PID)	SEQ ID NO: of PID	Promoter	Project type	Vector
G929	P399	127	35S	Direct promoter-fusion	pMEN20
G2344	P1627	128	35S	Direct promoter-fusion	pMEN65
G931	P1608	129	35S	Direct promoter-fusion	pMEN65
G3920	P26608	130	35S	Direct promoter-fusion	pMEN65
G928	P143	131	35S	Direct promoter-fusion	pMEN20
G1782	P966	132	35S	Direct promoter-fusion	pMEN65
G1363	P26121	133	35S	Protein-YFP-C-fusion	P25800
G3924	P26602	134	35S	Direct promoter-fusion	pMEN65

G3926	P26600	135	35S	Direct promoter-fusion	pMEN65
G3925	P26597	136	35S	Direct promoter-fusion	pMEN65
G4264	P26593	137	35S	Direct promoter-fusion	pMEN65
G2632	P15494	138	35S	Direct promoter-fusion	pMEN65
G1334	P714	139	35S	Direct promoter-fusion	pMEN20
G926	P26217	140	35S	Direct promoter-fusion	pMEN65
G926	P26217	141	35S	Protein-YFP-C-fusion	P25800
G927	P142	142	35S	Direct promoter-fusion	pMEN20
G3911	P26591	143	35S	Direct promoter-fusion	pMEN65
G3546	P26603	144	35S	Direct promoter-fusion	pMEN65
G3909	P26596	145	35S	Direct promoter-fusion	pMEN20
G3552	P26595	146	35S	Direct promoter-fusion	pMEN65
G483	P48	147	35S	Direct promoter-fusion	pMEN20
G3547	P26758	148	35S	Direct promoter-fusion	pMEN65
G714	P111	149	35S	Direct promoter-fusion	pMEN20
G3542	P26604	150	35S	Direct promoter-fusion	pMEN65
G489	P26060	151	35S	Protein-YFP-C-fusion	P25800
G3544	P26599	152	35S	Direct promoter-fusion	pMEN65
G3550	P26606	153	35S	Direct promoter-fusion	pMEN65
G3548	P26610	154	35S	Direct promoter-fusion	pMEN65

G715	P15502	155	35S	Direct promoter-fusion	pMEN65
G3886	P26607	156	35S	Direct promoter-fusion	pMEN65
G3889	P26590	157	35S	Direct promoter-fusion	pMEN65
G1646	P964	158	35S	Direct promoter-fusion	pMEN65
G3543	P26598	159	35S	Direct promoter-fusion	pMEN65
G1820	P1284	160	35S	Direct promoter-fusion	pMEN65
G1836	P973	161	35S	Direct promoter-fusion	pMEN65
G1819	P1285	162	35S	Direct promoter-fusion	pMEN65
G1818	P1677	163	35S	Direct promoter-fusion	pMEN65
G490	P912	164	35S	Direct promoter-fusion	pMEN65
G3074	P2712	165	35S	Direct promoter-fusion	pMEN1963
G1249	P1184	166	35S	Direct promoter-fusion	pMEN65
G3075	P2797	167	35S	Direct promoter-fusion	pMEN1963
G3883	P26821	200	35S	Direct promoter-fusion	pMEN65
G3894	P26611	204	35S	Direct promoter-fusion	pMEN65
	P5326	191	AP1	AP1::LexA-GAL4TA driver construct in two-component system	pMEN48
G1819	P4039	192		Transcription factor component of two-component system (opLexA::G1819)	pMEN53
	P5287	193	LTP1	LTP1::LexA-GAL4TA driver	pMEN48

				construct in two-component system	
G2344	P6063	194		Transcription factor component of two-component system (opLexA::G2344)	pMEN53
	P5324	205	Cruciferin	CRU::LexA-GAL4TA driver construct in two-component system	pMEN48
G926	P5562	207		Transcription factor component of two-component system (opLexA::G926)	pMEN53
	P5318	208	STM	STM::LexA-GAL4TA driver construct in two-component system	pMEN48
G929	P9107	206		Transcription factor component of two-component system (opLexA::G929)	pMEN53
	P25800	168	35S	YFP fusion vector	
	pMEN1963	169	35S	35S expression vector	
	pMEN20	170	35S	35S expression vector	
	pMEN48	195	35S	Two component driver vector	
	pMEN53	196		LexA operator and polylinker sequence two component target vector	
	pMEN65	171	35S	35S expression vector	

Example II. Transformation of *Agrobacterium* with the Expression Vector

After the plasmid vector containing the gene was constructed, the vector was used to transform *Agrobacterium tumefaciens* cells expressing the gene products. The stock of *Agrobacterium tumefaciens* cells for transformation was made as described by Nagel et al. (1990) *FEMS MicroBiol. Letts.* 67: 325-328. *Agrobacterium* strain ABI was grown in 250 ml LB medium (Sigma) overnight at 28°C with shaking until an absorbance (A_{600}) of 0.5 – 1.0 was

reached. Cells were harvested by centrifugation at 4,000 x g for 15 min at 4° C. Cells were then resuspended in 250 µl chilled buffer (1 mM HEPES, pH adjusted to 7.0 with KOH). Cells were centrifuged again as described above and resuspended in 125 µl chilled buffer. Cells were then centrifuged and resuspended two more times in the same HEPES buffer as described above at a volume of 100 µl and 750 µl, respectively. Resuspended cells were then distributed into 40 µl aliquots, quickly frozen in liquid nitrogen, and stored at -80° C.

Agrobacterium cells were transformed with plasmids prepared as described above following the protocol described by Nagel et al. For each DNA construct to be transformed, 50 – 100 ng DNA (generally resuspended in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0) was mixed with 40 µl of *Agrobacterium* cells. The DNA/cell mixture was then transferred to a chilled cuvette with a 2 mm electrode gap and subject to a 2.5 kV charge dissipated at 25 µF and 200 µF using a Gene Pulser II apparatus (Bio-Rad). After electroporation, cells were immediately resuspended in 1.0 ml LB and allowed to recover without antibiotic selection for 2 – 4 hours at 28° C in a shaking incubator. After recovery, cells were plated onto selective medium of LB broth containing 100 µg/ml spectinomycin (Sigma) and incubated for 24-48 hours at 28° C. Single colonies were then picked and inoculated in fresh medium. The presence of the plasmid construct was verified by PCR amplification and sequence analysis.

Example III. Transformation of *Arabidopsis* Plants

Transformation of *Arabidopsis* was performed by an *Agrobacterium*-mediated protocol based on the method of Bechtold and Pelletier (1998). Most of the experiments were performed with the *Arabidopsis thaliana* ecotype Columbia (col-0). Some of the results, as noted, were obtained with transformed tomato plants (*Lycopersicon esculentum*).

Plant preparation. Seeds were sown on mesh covered pots. The seedlings were thinned so that 6-10 evenly spaced plants remained on each pot 10 days after planting. The primary bolts were cut off a week before transformation to break apical dominance and encourage auxiliary shoots to form. Transformation was typically performed at 4-5 weeks after sowing.

Bacterial culture preparation. *Agrobacterium* stocks were inoculated from single colony plates or from glycerol stocks and grown with the appropriate antibiotics and grown until saturation. On the morning of transformation, the saturated cultures were centrifuged and bacterial pellets were re-suspended in Infiltration Media (0.5X MS, 1X B5 Vitamins, 5% sucrose, 1 mg/ml benzylaminopurine riboside, 200 µl/L Silwet L77) until an A600 reading of 0.8 was reached.

Transformation and seed harvest. The *Agrobacterium* solution was poured into dipping containers. All flower buds and rosette leaves of the plants were immersed in this solution for 30 seconds. The plants were laid on their side and wrapped with plastic wrap to keep the humidity high. The plants were kept this way overnight at 4° C and then the pots were turned upright, unwrapped, and moved to growth racks.

The plants were maintained on growth racks under 24-hour light until seeds were ready to be harvested. Seeds were harvested when 80% of the siliques of the transformed plants were ripe (approximately 5 weeks after the initial transformation). This seed was deemed T0 seed, since it was obtained from the T0 generation, and was later plated on selection plates (either kanamycin or sulfonamide). Resistant plants that were identified on such selection plates comprised the T1 generation.

Example IV. Morphology

Morphological analysis was performed to determine whether changes in polypeptide levels affect plant growth and development. This was primarily carried out on the T1 generation, when at least 10-20 independent lines were examined. However, in cases where a phenotype required confirmation or detailed characterization, plants from subsequent generations were also analyzed.

Primary transformants were selected on MS medium with 0.3% sucrose and 50 mg/l kanamycin. T2 and later generation plants were selected in the same manner, except that kanamycin was used at 35 mg/l. In cases where lines carry a sulfonamide marker (as in all lines generated by super-transformation), seeds were selected on MS medium with 0.3% sucrose and 1.5 mg/l sulfonamide. KO lines were usually germinated on plates without a selection. Seeds were cold-treated (stratified) on plates for three days in the dark (in order to increase germination efficiency) prior to transfer to growth cabinets. Initially, plates were incubated at 22°C under a light intensity of approximately 100 microEinsteins for 7 days. At this stage, transformants were green, possessed two true leaves, and were easily distinguished from bleached kanamycin or sulfonamide-susceptible seedlings. Resistant seedlings were then transferred onto soil (Sunshine potting mix). Following transfer to soil, trays of seedlings were covered with plastic lids for 2-3 days to maintain humidity while they became established. Plants were grown on soil under fluorescent light at an intensity of 70-95 microEinsteins and a temperature of 18-23°C. Light conditions consisted of a 24-hour photoperiod unless otherwise stated. In instances where alterations in flowering time were apparent, flowering time was re-

examined under both 12-hour and 24-hour light to assess whether the phenotype was photoperiod dependent. Under our 24-hour light growth conditions, the typical generation time (seed to seed) was approximately 14 weeks.

Because many aspects of *Arabidopsis* development are dependent on localized environmental conditions, in all cases plants were evaluated in comparison to controls in the same flat. For a given construct, ten transformed lines were typically examined in subsequent plate based physiology assays. Controls for transgenic lines were wild-type plants or transgenic plants harboring an empty transformation vector selected on kanamycin or sulfonamide. Careful examination was made at the following stages: young seedling (1 week), rosette (2-3 weeks), flowering (4-7 weeks), and late seed set (8-12 weeks). Seed was also inspected. Young seedling size and morphology was assessed on selection plates. At all other stages, plants were macroscopically evaluated while growing on soil. All significant differences (including alterations in growth rate, size, biomass, etc., were recorded as noted in Example V.

Example V. Assessment of Growth Rate and Size

In subsequent Examples, unless otherwise indicted, morphological traits are disclosed in comparison to control plants. That is, a transformed plant that is described as large and/or has a faster growth rate was large and had a faster growth rate with respect to a control plant, the latter including wild-type plants, parental lines and lines transformed with a vector that does not contain the transcription factor sequence of interest (e.g., an "empty" vector). When a plant is said to have a better performance than controls, it generally was larger, had greater yield, and/or showed fewer stress symptoms than control plants.

Germination assays. All germination assays were performed in tissue culture. Growing the plants under controlled temperature and humidity on sterile medium produced uniform plant material that has not been exposed to additional stresses (such as water stress) which could cause variability in the results obtained.

Prior to plating, seed for all experiments were surface sterilized in the following manner: (1) 5 minute incubation with mixing in 70 % ethanol, (2) 20 minute incubation with mixing in 30% bleach, 0.01% triton-X 100, (3) 5X rinses with sterile water, (4) Seeds were re-suspended in 0.1% sterile agarose and stratified at 4° C for 3-4 days.

All germination assays followed modifications of the same basic protocol. Sterile seeds were sown on the conditional media that has a basal composition of 80% MS + Vitamins. Plates

were incubated at 22° C under 24-hour light (120-130 μE m^{-2} s^{-1}) in a growth chamber. Evaluation of germination and seedling vigor was performed five days after planting.

Growth assays. Assays were usually conducted on *Arabidopsis thaliana* ecotype Columbia (col-0) non-selected segregating T2 populations (in order to avoid the extra stress of selection). Control plants for assays on lines containing direct promoter-fusion constructs were wild-type Col-0 plants and/or Col-0 plants transformed with an empty transformation vector (pMEN65, SEQ ID NO: 171).

Example VI. Morphological observations with HAP2 and HAP5 overexpressors in *Arabidopsis* and tomato

Overexpression of HAP2 and HAP5 transcription factors in *Arabidopsis* or tomato plants produced the experimental observations related to size or growth rate that are listed in Tables 6 and 7. Experiments indicating larger seedlings or plants than controls also demonstrated a faster growth rate as the observed larger sizes were achieved in the same time period of growth for both controls and experimental plants. This may be particularly important for seedlings of overexpressors that were larger than controls as these plants may be more tolerant to environmental stresses encountered early in their growth.

Table 6. Yield-related experimental results obtained with HAP2 overexpressors

GID	SEQ ID NO: of polypeptide	SEQ ID NO: of conserved domain	% Identity of conserved domain in first column to conserved domain of G929	% Identity of conserved domain in first column to conserved domain of G3926	Experimental Observations
At/G929	2	85	100%	73%	Two 35S::G929 lines produced seedlings that were larger than controls, and three lines had larger rosettes at the flowering stage A transgenic tomato plant overexpressing G929 under the regulatory control of the cruciferin promoter was considerably larger than control plants (Figure 6).
At/G2344	4	86	86%	75%	Three of ten 35S::G2344 lines examined produced seedlings that were larger

					than controls, and one line had larger rosettes at the flowering stage. The average fruit weights of LTP1::G2344 tomato plants were within the top 1% of all tomato lines tested (plants comprised the two component expression system of SEQ ID NOs: 193 and 194), and STM::G2344 tomato plants were within the top 8% of all tomato lines tested (plants comprised the two component expression system of SEQ ID NOs: 208 and 194); empty vector controls were in the 56th percentile.
At/G931	6	87	76%	80%	Four 35S::G931 lines produced plants that were larger than controls at the rosette stage, and two of these lines maintained larger rosettes at the flowering stage.
Gm/G3920	8	88	76%	73%	Four of ten 35S::G3920 lines produced seedlings that were larger than controls; one line produced plants with larger rosettes than controls at the flowering stage.
At/G928	10	88	75%	78%	Two 35S::G928 lines produced seedlings that were larger than controls, and one of these lines was also larger at the rosette and flowering stages.
At/G1782	12	90	75%	85%	Four 35S::G1782 lines produced plants that were larger than controls at the rosette stage, and two of these lines maintained larger rosettes at the flowering stage.
At/G1363	14	91	73%	84%	One 35S::G1363 line produced plants that were

					larger at the flowering stage; seven G1363-YFP fusion lines had broader leaves at the rosette stage, and four G1363-YFP fusion lines were large at the flowering stage.
Os/G3924	16	92	73%	75%	Two of ten 35S::G3924 lines examined produced seedlings that were larger than controls.
Os/G3926	18	93	71%	100%	Three of ten 35S::G3926 lines tested produced seedlings that were larger than controls.
Os/G3925	20	94	71%	85%	Two of ten 35S::G3925 lines examine produced seedlings that were larger than controls.
Os/G4264	26	97	71%	91%	Two of ten 35S::G4264 lines produced seedlings that were larger than controls; four lines produced plants with larger rosettes than controls at the flowering stage.
At/G2632	28	98	73%	72%	None examined thus far have been found that were larger or had a faster growth rate than controls, and some 35S::G2632 seedlings were smaller than controls;
At/G1334	30	99	70%	65%	Four of ten 35S::G1334 lines tested produced seedlings that were larger than controls.
At/G926	32	100	66%	67%	Seedlings overexpressing G926-YFP fusion proteins were similar in size and growth rate to controls. However, the average fruit weights of 35S::G926 tomato plants were within the top 4% of all tomato lines tested (plants comprised the one component expression system of SEQ ID NO: 140), and the average fruit weights

					of cruciferin::G926 tomato plants were within the top 10% of all tomato lines tested (plants comprised the two component expression system of SEQ ID NOs: 205 and 207); empty vector controls were in the 56th percentile.
At/G927	34	101	64%	57%	35S::G927 seedlings were similar in size and growth rate to controls.

Table 7. Yield-related experimental results obtained with HAP5 overexpressors

GID	SEQ ID NO: of polypeptide	SEQ ID NO: of CCAAT-binding domain	% Identity of conserved domain in first column to conserved domain of G3911	% Identity of conserved domain in first column to conserved domain of G3543	Experimental Observations
Zm/G3911	36	102	100%	83%	Nine 35S::G3911 lines produced seedlings that were larger than controls, and two lines had larger rosettes at their late flowering stage.
Os/G3546	38	103	100%	83%	All ten 35S::G3546 lines tested produced seedlings that were larger than controls.
Zm/G3909	40	104	96%	80%	Seven 35S::G3909 lines produced seedlings that were larger than controls, and seven lines had larger rosettes at their early flowering stage.
Le/G3894	202	203	89%	81%%	Seedlings from a single line of 35S::G3894 plants of ten lines tested were larger and more vigorous than controls seven days after planting.
Zm/G3552	42	105	86%	84%	Eight 35S::G3552 lines produced seedlings that were larger than controls, and one line had slightly broader leaves than controls at its early flowering stage.

At/G483	44	106	86%	83%	35S::G483 overexpressors were wild-type in size and growth rate in experiments performed to date
Gm/G3547	46	107	87%	83%	Three 35S::G3547 lines produced seedlings that were larger than controls.
At/G714	48	108	87%	81%	Three 35S::G714 lines had larger rosettes at their late flowering stage.
Os/G3542	50	109	86%	84%	Five of ten 35S::G3542 lines tested produced seedlings that were larger than controls.
At/G489	52	110	87%	81%	Thirteen G489-YFP fusion lines had larger rosettes than controls at their late flowering stage.
Os/G3544	54	111	84%	84%	Two of ten 35S::G3544 lines tested produced seedlings that were larger than controls.
Gr/G3883	198	199	84%	90%	35S::G3883 overexpressors were wild-type in size and growth rate in experiments performed to date.
Gm/G3550	56	112	86%	84%	Five of ten 35S::G3550 lines tested produced seedlings that were larger than controls.
Gm/G3548	58	113	86%	84%	Three of ten 35S::G3548 lines tested produced seedlings that were larger than controls.
At/G715	60	114	84%	90%	One of ten 35S::G715 lines tested produced seedlings that were larger than controls, and another line was larger at the early flowering stage
Gm/G3886	62	115	84%	90%	One of ten 35S::G3886 lines tested produced seedlings that were slightly larger than controls.
Zm/G3889	64	116	84%	96%	Five of ten 35S::G3889 lines tested produced seedlings that were larger than controls.
At/G1646	66	117	84%	90%	One of ten 35S::G1646 lines tested produced seedlings that were slightly larger than controls; three lines had larger rosettes at early and late flowering stages.

Os/G3543	68	118	83%	100%	Five of ten 35S::G3543 lines tested produced seedlings that were larger than controls.
At/G1820	70	119	73%	71%	35S::G1820 overexpressors were wild-type in size and growth rate in experiments performed to date.
At/G1836	72	120	63%	63%	One 35S::G1836 line produced larger seedlings in germination and growth assays, and five lines had very full, large rosettes with broad leaves at the late flowering stage.
At/G1819	74	121	52%	61%	One 35S::G1819 *Arabidopsis* line was slightly larger than controls at the early flowering stage. The average fruit weights of AP1::G1819 overexpressing tomato plants were within the top 5% of all tomato lines tested (plants comprised the two component expression system of SEQ ID NOs: 191 and 192; empty vector controls were in the 56th percentile).
At/G1818	76	122	55%	55%	At late flowering to late stage, many 35S::G1818 lines produced large, full rosettes.
At/G490	78	123	44%	47%	Two 35S::G490 lines produced larger fuller rosettes than controls at the late flowering stage.
At/G3074	80	124	37%	39%	Two 35S::G3074 lines produced larger seedlings in germination and growth assays and five lines had slightly larger rosettes than controls.
At/G1249	82	125	34%	34%	Two 35S::G1249 lines of ten lines tested of overexpressors produced larger seedlings in germination assays.
At/G3075	84	126	22%	23%	35S::G3075 overexpressors were wild-type in size and growth rate in experiments performed to date.

Utilities of HAP2 and HAP5 transcription factors in plants

Based on the data obtained in the above-disclosed Examples, the increased size, height and/or biomass of plants that overexpress HAP2 or HAP5 transcription factors indicate that these sequences when overexpressed may be used to improve yield of commercially valuable plants or to help these plants become established more successfully or quickly.

Example VII. Transformation of dicots to produce increased yield

Crop species that overexpress polypeptides of the invention may produce plants with increased growth rate, size, biomass and/or yield in both stressed and non-stressed conditions. Thus, polynucleotide sequences listed in the Sequence Listing recombined into, for example, one of the expression vectors of the invention, or another suitable expression vector, may be transformed into a plant for the purpose of modifying plant traits for the purpose of improving yield and/or quality. The expression vector may contain a constitutive, tissue-specific or inducible promoter operably linked to the polynucleotide. The cloning vector may be introduced into a variety of plants by means well known in the art such as, for example, direct DNA transfer or *Agrobacterium tumefaciens*-mediated transformation. It is now routine to produce transgenic plants using most dicot plants (see Weissbach and Weissbach, 1989; Gelvin et al., 1990; Herrera-Estrella et al., 1983; Bevan, 1984; and Klee, 1985). Methods for analysis of traits are routine in the art and examples are disclosed above.

Numerous protocols for the transformation of tomato and soy plants have been previously described, and are well known in the art. Gruber et al. (1993) and Glick and Thompson (1993) describe several expression vectors and culture methods that may be used for cell or tissue transformation and subsequent regeneration. For soybean transformation, methods are described by Miki et al. (1993) and U.S. Pat. No. 5,563,055, (Townsend and Thomas), issued Oct. 8, 1996.

There are a substantial number of alternatives to *Agrobacterium*-mediated transformation protocols, other methods for the purpose of transferring exogenous genes into soybeans or tomatoes. One such method is microprojectile-mediated transformation, in which DNA on the surface of microprojectile particles is driven into plant tissues with a biolistic device (see, for example, Sanford et al., 1987; Christou et al., 1992; Sanford, 1993; Klein et al., 1987; U.S. Pat. No. 5,015,580 (Christou et al), issued May 14, 1991; and U.S. Pat. No. 5,322,783 (Tomes et al., issued Jun. 21, 1994)).

Alternatively, sonication methods (see, for example, Zhang et al., 1991); direct uptake of DNA into protoplasts using $CaCl_2$ precipitation, polyvinyl alcohol or poly-L-ornithine (see, for example, Hain et al., 1985; Draper et al., 1982); liposome or spheroplast fusion (see, for example, Deshayes et al., 1985); Christou et al., 1987); and electroporation of protoplasts and whole cells and tissues (see, for example, Donn et al., 1990; D'Halluin et al., 1992; and Spencer et al., 1994) have been used to introduce foreign DNA and expression vectors into plants.

After a plant or plant cell is transformed (and the latter regenerated into a plant), the transformed plant may be crossed with itself or a plant from the same line, a non-transformed or wild-type plant, or another transformed plant from a different transgenic line of plants. Crossing provides the advantages of producing new and often stable transgenic varieties. Genes and the traits they confer that have been introduced into a tomato or soybean line may be moved into distinct lines of plants using traditional backcrossing techniques well known in the art. Transformation of tomato plants may be conducted using the protocols of Koornneef et al (1986), and in U.S. Patent 6,613,962, the latter method described in brief here. Eight day old cotyledon explants are precultured for 24 hours in Petri dishes containing a feeder layer of Petunia hybrida suspension cells plated on MS medium with 2% (w/v) sucrose and 0.8% agar supplemented with 10 μM □-naphthalene acetic acid and 4.4 μM 6-benzylaminopurine. The explants are then infected with a diluted overnight culture of *Agrobacterium tumefaciens* containing an expression vector comprising a polynucleotide of the invention for 5-10 minutes, blotted dry on sterile filter paper and cocultured for 48 hours on the original feeder layer plates. Culture conditions are as described above. Overnight cultures of *Agrobacterium tumefaciens* are diluted in liquid MS medium with 2% (w/v/) sucrose, pH 5.7) to an OD_{600} of 0.8.

Following cocultivation, the cotyledon explants are transferred to Petri dishes with selective medium comprising MS salts with 4.56 μM zeatin, 67.3 μM vancomycin, 418.9 μM cefotaxime and 171.6 μM kanamycin sulfate, and cultured under the culture conditions described above. The explants are subcultured every three weeks onto fresh medium. Emerging shoots are dissected from the underlying callus and transferred to glass jars with selective medium without zeatin to form roots. The formation of roots in a kanamycin sulfate-containing medium is a positive indication of a successful transformation.

Transformation of soybean plants may be conducted using the methods found in, for example, U.S. Patent 5,563,055 (Townsend et al., issued October 8, 1996), described in brief here. In this method soybean seed is surface sterilized by exposure to chlorine gas evolved in a glass bell jar. Seeds are germinated by plating on 1/10 strength agar solidified medium without

plant growth regulators and culturing at 28° C. with a 16 hour day length. After three or four days, seed may be prepared for cocultivation. The seedcoat is removed and the elongating radicle removed 3-4 mm below the cotyledons.

Overnight cultures of *Agrobacterium tumefaciens* harboring the expression vector comprising a polynucleotide of the invention are grown to log phase, pooled, and concentrated by centrifugation. Inoculations are conducted in batches such that each plate of seed was treated with a newly resuspended pellet of *Agrobacterium*. The pellets are resuspended in 20 ml inoculation medium. The inoculum is poured into a Petri dish containing prepared seed and the cotyledonary nodes are macerated with a surgical blade. After 30 minutes the explants are transferred to plates of the same medium that has been solidified. Explants are embedded with the adaxial side up and level with the surface of the medium and cultured at 22° C. for three days under white fluorescent light. These plants may then be regenerated according to methods well established in the art, such as by moving the explants after three days to a liquid counter-selection medium (see U.S. Patent 5,563,055).

The explants may then be picked, embedded and cultured in solidified selection medium. After one month on selective media transformed tissue becomes visible as green sectors of regenerating tissue against a background of bleached, less healthy tissue. Explants with green sectors are transferred to an elongation medium. Culture is continued on this medium with transfers to fresh plates every two weeks. When shoots are 0.5 cm in length they may be excised at the base and placed in a rooting medium.

Example VIII: Transformation of monocots to produce increased yield

Members of the family Gramineae, including turfgrass or other grasses such as *Miscanthus*, *Panicum virgatum* or other *Panicum* species, or cereal plants such as barley, corn, rice, rye, sorghum, or wheat, may be transformed with the present polynucleotide sequences, including monocot or dicot-derived sequences such as those presented in the present Tables 1 - 7, cloned into a vector containing a kanamycin-resistance marker, and expressed constitutively under, for example, the CaMV 35S, STM, AP1, LPT1, cruciferin, or COR15 promoters, or with other tissue-specific or inducible promoters. The expression vectors may be one found in the Sequence Listing, or any other suitable expression vector may be similarly used. For example, pMEN020 may be modified to replace the NptII coding region with the Bar gene of *Streptomyces hygroscopicus* that confers resistance to phosphinothricin. The KpnI and BglII sites of the Bar gene are removed by site-directed mutagenesis with silent codon changes.

The cloning vector may be introduced into a variety of cereal plants by means well known in the art including direct DNA transfer or *Agrobacterium tumefaciens*-mediated transformation. The latter approach may be accomplished by a variety of means, including, for example, that of U.S. Patent No. 5,591,616, in which monocotyledon callus is transformed by contacting dedifferentiating tissue with the *Agrobacterium* containing the cloning vector.

The sample tissues are immersed in a suspension of $3x10^{-9}$ cells of *Agrobacterium* containing the cloning vector for 3-10 minutes. The callus material is cultured on solid medium at 25° C in the dark for several days. The calli grown on this medium are transferred to Regeneration medium. Transfers are continued every 2-3 weeks (2 or 3 times) until shoots develop. Shoots are then transferred to Shoot-Elongation medium every 2-3 weeks. Healthy looking shoots are transferred to rooting medium and after roots have developed, the plants are placed into moist potting soil.

The transformed plants are then analyzed for the presence of the NPTII gene/ kanamycin resistance by ELISA, using the ELISA NPTII kit from 5Prime-3Prime Inc. (Boulder, CO).

It is also routine to use other methods to produce transgenic plants of most cereal crops (Vasil, 1994) such as corn, wheat, rice, sorghum (Cassas et al., 1993), and barley (Wan and Lemeaux, 1994). DNA transfer methods such as the microprojectile method can be used for corn (Fromm et al., 1990; Gordon-Kamm et al., 1990; Ishida, 1990), wheat (Vasil et al., 1992; Vasil et al., 1993; Weeks et al., 1993), and rice (Christou, 1991; Hiei et al., 1994; Aldemita and Hodges, 1996; and Hiei et al., 1997). For most cereal plants, embryogenic cells derived from immature scutellum tissues are the preferred cellular targets for transformation (Hiei et al., 1997; Vasil, 1994). For transforming corn embryogenic cells derived from immature scutellar tissue using microprojectile bombardment, the A188XB73 genotype is the preferred genotype (Fromm et al., 1990; Gordon-Kamm et al., 1990). After microprojectile bombardment, the tissues are selected on phosphinothricin to identify the transgenic embryogenic cells (Gordon-Kamm et al., 1990). Transgenic plants are regenerated by standard corn regeneration techniques (Fromm et al., 1990; Gordon-Kamm et al., 1990).

Example IX: Expression and analysis of increased yield in non-*Arabidopsis* species

It is expected that structurally similar orthologs of the HAP2 or HAP5 polypeptide sequences, including those found in the Sequence Listing, can confer increased yield relative to control plants.

Northern blot analysis, RT-PCR or microarray analysis of the regenerated, transformed plants may be used to show expression of a polypeptide or the invention and related genes that are capable of inducing increased growth rate and/or larger size of the plants, including larger seed, plant products or plant parts, such as leaves, roots or stems.

After a dicot plant, monocot plant or plant cell has been transformed (and the latter regenerated into a plant) and shown to have greater size, improved planting density, that is, able to tolerate greater planting density with a coincident increase in yield in the presence or absence of stress conditions, the transformed monocot plant may be crossed with itself or a plant from the same line, a non-transformed or wild-type monocot plant, or another transformed monocot plant from a different transgenic line of plants.

The function of specific polypeptides of the invention, including closely-related orthologs, have been analyzed and may be further characterized and incorporated into crop plants. The ectopic overexpression of these sequences may be regulated using constitutive, inducible, or tissue specific regulatory elements. Genes that have been examined and have been shown to modify plant traits (including increasing growth rate and/or yield) encode polypeptides found in the Sequence Listing. In addition to these sequences, it is expected that newly discovered polynucleotide and polypeptide sequences closely related to polynucleotide and polypeptide sequences found in the Sequence Listing can also confer alteration of traits in a similar manner to the sequences found in the Sequence Listing, when transformed into any of a considerable variety of plants of different species, and including dicots and monocots. The polynucleotide and polypeptide sequences derived from monocots (e.g., the rice sequences) may be used to transform both monocot and dicot plants, and those derived from dicots (e.g., the *Arabidopsis* and soy genes) may be used to transform either group, although it is expected that some of these sequences will function best if the gene is transformed into a plant from the same group as that from which the sequence is derived.

It is expected that the same methods may be applied to identify other useful and valuable sequences of the present polypeptide clades, and the sequences may be derived from a broad range of plant species.

References cited:

Aldemita and Hodges (1996) *Planta* 199: 612-617

Altschul (1990) *J. Mol. Biol.* 215: 403-410

Altschul (1993) *J. Mol. Evol.* 36: 290-300

Anderson and Young (1985) "Quantitative Filter Hybridisation", In: Hames and Higgins, ed., Nucleic Acid Hybridisation, A Practical Approach. Oxford, IRL Press, 73-111

Ausubel et al. (1997) Short Protocols in Molecular Biology, John Wiley & Sons, New York, NY, unit 7.7

Bairoch et al. (1997) *Nucleic Acids Res.* 25: 217-221

Bechtold and Pelletier (1998) *Methods Mol. Biol.* 82: 259-266

Berger and Kimmel (1987), "Guide to Molecular Cloning Techniques", in Methods in Enzymology, vol. 152, Academic Press, Inc., San Diego, CA

Bevan (1984) *Nucleic Acids Res.* 12: 8711-8721

Bhattacharjee et al. (2001) *Proc. Natl. Acad. Sci. USA* 98: 13790-13795

Borevitz et al. (2000) *Plant Cell* 12: 2383-2393

Boss and Thomas (2002) *Nature*, 416: 847-850

Bruce et al. (2000) *Plant Cell* 12: 65-79

Bucher (1990) *J. Mol. Biol. 212*: 563-578

Bucher and Trifonov (1988) *J. Biomol. Struct. Dyn.* 5: 1231-1236

Cassas et al. (1993) *Proc. Natl. Acad. Sci. USA* 90: 11212-11216

Cheikh et al. (2003) U.S. Patent Application No. 20030101479

Christou et al. (1987) *Proc. Natl. Acad. Sci. USA* 84: 3962-3966

Christou (1991) *Bio/Technol.* 9:957-962

Christou et al. (1992) *Plant. J.* 2: 275-281

Coupland (1995) *Nature* 377: 482-483

Daly et al. (2001) *Plant Physiol.* 127: 1328-1333

Dang et al. (1996) *J. Bacteriol.* 178: 1842-1849

Deshayes et al. (1985) *EMBO J.*, 4: 2731-2737

D'Halluin et al. (1992) *Plant Cell* 4: 1495-1505

Donn et al.(1990) in Abstracts of VIIth International Congress on Plant Cell and Tissue Culture IAPTC, A2-38: 53

Doolittle, ed. (1996) *Methods in Enzymology*, vol. 266: "Computer Methods for Macromolecular Sequence Analysis" Academic Press, Inc., San Diego, Calif., USA

Draper et al. (1982) *Plant Cell Physiol.* 23: 451-458

Eddy (1996) *Curr. Opin. Str. Biol.* 6: 361-365

Edwards et al. (1998) *Plant Physiol.* 117: 1015-1022

Eisen (1998) *Genome Res.* 8: 163-167

Feng and Doolittle (1987) *J. Mol. Evol.* 25: 351-360

Forsburg and Guarente (1988) *Genes Dev.* 3: 1166-1178

Forsburg and Guarente (1989) Genes Dev. 3: 1166-1178

Fowler and Thomashow (2002) *Plant Cell* 14: 1675-1690

Fromm et al. (1990) *Bio/Technol.* 8: 833-839

Fu et al. (2001) *Plant Cell* 13: 1791-1802

Gancedo (1998) *Microbiol. Mol. Biol. Rev.* 62: 334-361

Gelinas et al. (1985) *Prog. Clin. Biol. Res.* 191: 125-139

Gelvin et al. (1990) Plant Molecular Biology Manual, Kluwer Academic Publishers

Gilmour et al. (1998) *Plant J.* 16: 433-442

Glick and Thompson (1993) Methods in Plant Molecular Biology and Biotechnology. eds., CRC Press, Inc., Boca Raton

Gruber et al. (in Glick and Thompson (1993) Methods in Plant Molecular Biology and Biotechnology. eds., CRC Press, Inc., Boca Raton

Goodrich et al. (1993) Cell 75: 519-530

Gordon-Kamm et al. (1990) *Plant Cell* 2: 603-618

Hain et al. (1985) *Mol. Gen. Genet.* 199: 161-168

Haymes et al. (1985) Nucleic Acid Hybridization: A Practical Approach, IRL Press, Washington, D.C.

He et al. (2000) *Transgenic Res.* 9: 223-227

Hein (1990) *Methods Enzymol.* 183: 626-645

Henikoff and Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915

Henikoff and Henikoff (1991) Nucleic Acids Res. 19: 6565-6572

Herrera-Estrella et al. (1983) *Nature* 303: 209

Hiei et al. (1994) *Plant J.* 6:271-282

Hiei et al. (1997) *Plant Mol. Biol.* 35:205-218

Higgins and Sharp (1988) Gene 73: 237-244

Higgins et al. (1996) *Methods Enzymol.* 266: 383-402

Ishida (1990) *Nature Biotechnol.* 14:745-750

Jaglo et al. (2001) *Plant Physiol.* 127: 910-917

Kashima et al. (1985) *Nature* 313: 402-404

Kim et al. (2001) *Plant J.* 25: 247-259

Kimmel (1987) *Methods Enzymol.* 152: 507-511

Klee (1985) *Bio/Technology* 3: 637-642

Klein et al. (1987) *Nature* 327: 70-73

Koornneef et al (1986) in Tomato Biotechnology: Alan R. Liss, Inc., 169-178

Ku et al. (2000) *Proc. Natl. Acad. Sci. USA* 97: 9121-9126

Kyozuka and Shimamoto (2002) *Plant Cell Physiol.* 43: 130-135

Li et al. (1992) *Nucleic Acids Res.* 20: 1087-1091

Lin et al. (1991) *Nature* 353: 569-571

Mandel (1992a) *Nature* 360: 273-277

Mandel et al. (1992b) *Cell* 71-133-143

Masiero et al. (2002) *J. Biol. Chem.* 277, 26429-26435

Mazon et al. (1982) *Eur. J. Biochem.* 127: 605-608

Meyers (1995) Molecular Biology and Biotechnology, Wiley VCH, New York, NY, p 856-853

Miki et al. (1993) in Methods in Plant Molecular Biology and Biotechnology, p. 67-88, Glick and Thompson, eds., CRC Press, Inc., Boca Raton

Mount (2001), in Bioinformatics: Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, p. 543

Müller et al. (2001) *Plant J.* 28: 169-179

Nandi et al. (2000) *Curr. Biol.* 10: 215-218

Peng et al. (1997) *Genes Development* 11: 3194-3205)

Peng et al. (1999) *Nature* 400: 256-261

Ratcliffe et al. (2001) *Plant Physiol.* 126: 122-132

Ratcliffe et al. (2003) *Plant Cell* 15: 1159-1169

Riechmann et al. (2000a) *Science* 290: 2105-2110

Riechmann and Ratcliffe (2000b) *Curr. Opin. Plant Biol.* 3, 423-434

Rieger et al. (1976) Glossary of Genetics and Cytogenetics: Classical and Molecular, 4th ed., Springer Verlag, Berlin

Robson et al. (2001) *Plant J.* 28: 619-631

Romier et al. (2003) *J. Biol. Chem.* 278: 1336-1345

Sadowski et al. (1988) *Nature* 335: 563-564

Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Sanford et al. (1987) *Part. Sci. Technol.* 5:27-37

Sanford (1993) *Methods Enzymol.* 217: 483-509

Shpaer (1997) Methods Mol. Biol. 70: 173-187

Smith et al. (1992) *Protein Engineering* 5: 35-51

Sonnhammer et al. (1997) *Proteins* 28: 405-420

Spencer et al. (1994) *Plant Mol. Biol.* 24: 51-61

Suzuki et al. (2001) *Plant J.* 28: 409-418

Tasanen et al. (1992) J. Biol. Chem. 267: 11513-11519

Thompson et al. (1994) *Nucleic Acids Res.* 22: 4673-4680

Tudge (2000) in The Variety of Life, Oxford University Press, New York, NY pp. 547-606

Vasil et al. (1992) *Bio/Technol.* 10:667-674

Vasil et al. (1993) *Bio/Technol.* 11:1553-1558

Vasil (1994) *Plant Mol. Biol.* 25: 925-937

Wahl and Berger (1987) Methods Enzymol. 152: 399-407

Wan and Lemeaux (1994) *Plant Physiol.* 104: 37-48

Weeks et al. (1993) *Plant Physiol.* 102:1077-1084

Weigel and Nilsson (1995) *Nature* 377: 482-500

Weissbach and Weissbach (1989) Methods for Plant Molecular Biology, Academic Press

Xu et al. (2001) *Proc. Natl. Acad. Sci. USA* 98: 15089-15094

Zhang et al. (1991) *Bio/Technology* 9: 996-997

All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

The present invention is not limited by the specific embodiments described herein. The invention now being fully described, it will be apparent to one of ordinary skill in the art that many changes and modifications can be made thereto without departing from the spirit or scope of the Claims. Modifications that become apparent from the foregoing description and accompanying figures fall within the scope of the following Claims.

CLAIMS

1. An expression vector comprising a recombinant nucleic acid sequence encoding a polypeptide sharing a percentage of amino acid identity with any of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 198, 202, 210 or 213; wherein

when the polypeptide is overexpressed in a plant, the polypeptide regulates transcription and confers at least one regulatory activity resulting in an altered trait in the plant as compared to a control plant;

wherein the percentage of amino acid identity is selected from the group consisting of at least 35%, at least 36%, at least 37%, at least 38%, at least 39%, at least 40%, at least 41%, at least 42%, at least 45%, at least 46%, at least 47%, at least 49%, at least 50%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 61%, at least 62%, at least 63%, at least 72%, at least 78%, at least 79%, and at least 86%.

2. An expression vector comprising a recombinant nucleic acid sequence encoding a polypeptide sharing a percentage of amino acid identity with any of SEQ ID NO: 85 - 126, 199, 203, 211, or 214; wherein

when the polypeptide is overexpressed in a plant, the polypeptide regulates transcription and confers at least one regulatory activity resulting in an altered trait in the plant as compared to a control plant;

wherein the percentage of amino acid identity is selected from the group consisting of at least 34%, at least 37%, at least 39%, at least 44%, at least 47%, at least 52%, at least 55%, at least 61%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 70%, at least 71%, at least 72%, at least 73%, at least 75%, at least 76%, at least 78%, at least 80%, at least 81%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 89%, at least 90%, at least 91%, at least 92%, at least 96%, and 100%.

3. The expression vector of Claim 1 or Claim 2, wherein the expression vector encodes a polypeptide comprising any of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 198, 202, 210, or 213.

4. The expression vector of Claim 1 or Claim 2, wherein the expression vector further comprises a constitutive, inducible, or tissue-specific promoter operably lined to the recombinant nucleic acid sequence.

5. The expression vector of Claim 4, wherein the tissue-specific promoter regulates transcription in a tissue selected from the group consisting of floral meristem, epidermis, vascular, shoot apical meristem, embryo, endosperm, and fruit.

6. The expression vector of Claim 4, wherein the altered trait is selected from the group consisting of greater yield, greater size, greater biomass, and faster growth rate, as compared to the control plant.

7. A recombinant host cell comprising an expression vector of any of Claims 1-6.

8. A transgenic plant comprising an expression vector of any of Claims 1-6, wherein when a polypeptide encoded by the expression vector is overexpressed in the transgenic plant, the polypeptide confers at least one regulatory activity resulting in an altered trait selected from the group consisting of greater size, greater biomass, and faster growth rate, as compared to the control plant.

9. The transgenic plant of Claim 8, wherein the transgenic plant is a monocot.

10. The transgenic plant of Claim 9, wherein the transgenic plant is a member of the family Gramineae.

11. The transgenic plant of Claim 8, wherein the transgenic plant is a dicot.

12. The transgenic plant of Claim 11, wherein the transgenic plant is a tomato plant.

13. A transgenic seed produced from the transgenic plant of any of Claims 8-12.

14. A method for increasing yield of a plant as compared to yield of a control plant, the method comprising:

(a) providing an expression vector of any of Claims 1-6; wherein

when a polypeptide encoded by the expression vector is overexpressed in a plant, the polypeptide confers at least one regulatory activity resulting in an altered trait selected from the group consisting of greater size, greater biomass, and faster growth rate, as compared to the control plant; and

(b) transforming a target plant with the expression vector to produce a transgenic plant; wherein the transgenic plant produces greater yield than the wild-type plant.

15. The method of Claim 14, wherein the methods further comprises the step of:

(c) selecting a transgenic plant that ectopically expresses the polypeptide.

16. The method of Claim 14, wherein the expression vector further comprises a constitutive, inducible, or tissue-specific promoter operably lined to the recombinant nucleic acid sequence.

17. The method of Claim 14, wherein the tissue-specific promoter regulates transcription in a tissue selected from the group consisting of floral meristem, epidermis, vascular, shoot apical meristem, embryo, endosperm, and fruit.

18. The method of Claim 14, wherein the expression vector encodes a polypeptide comprising any of SEQ ID NO: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 198, 202, 210, or 213.

19. The method of Claim 14, wherein the method steps further comprise selfing or crossing the transgenic plant with itself or another plant, respectively, to produce transgenic seed.

20. A method for producing a transgenic plant having greater biomass, growth rate or yield as compared to a control plant, the method comprising:

(a) providing an expression vector of any of Claims 1-6; wherein

when a polypeptide encoded by the expression vector is overexpressed in a plant, the polypeptide confers at least one regulatory activity resulting in an altered trait selected from

the group consisting of greater size, greater biomass, and faster growth rate, as compared to the control plant;

(b) transforming a target plant with the expression vector to produce a transgenic plant; and

(c) optionally, selecting the transgenic plant that has greater biomass, growth rate or yield than the wild-type plant.

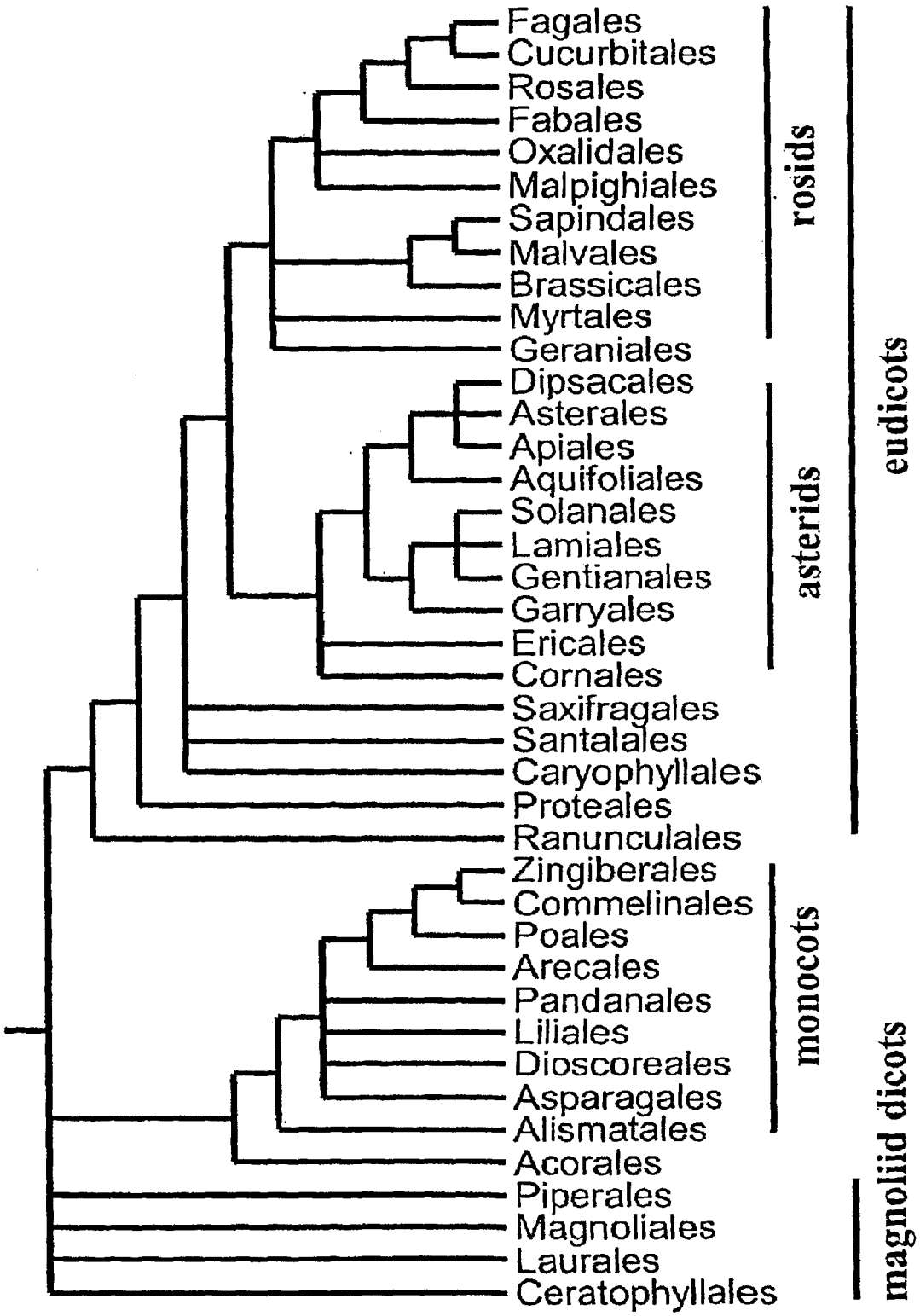

Fagales
Cucurbitales
Rosales
Fabales
Oxalidales
Malpighiales
Sapindales
Malvales
Brassicales
Myrtales
Geraniales
Dipsacales
Asterales
Apiales
Aquifoliales
Solanales
Lamiales
Gentianales
Garryales
Ericales
Cornales
Saxifragales
Santalales
Caryophyllales
Proteales
Ranunculales
Zingiberales
Commelinales
Poales
Arecales
Pandanales
Liliales
Dioscoreales
Asparagales
Alismatales
Acorales
Piperales
Magnoliales
Laurales
Ceratophyllales
rosids
asterids
eudicots
monocots
magnoliid dicots

FIG. 1

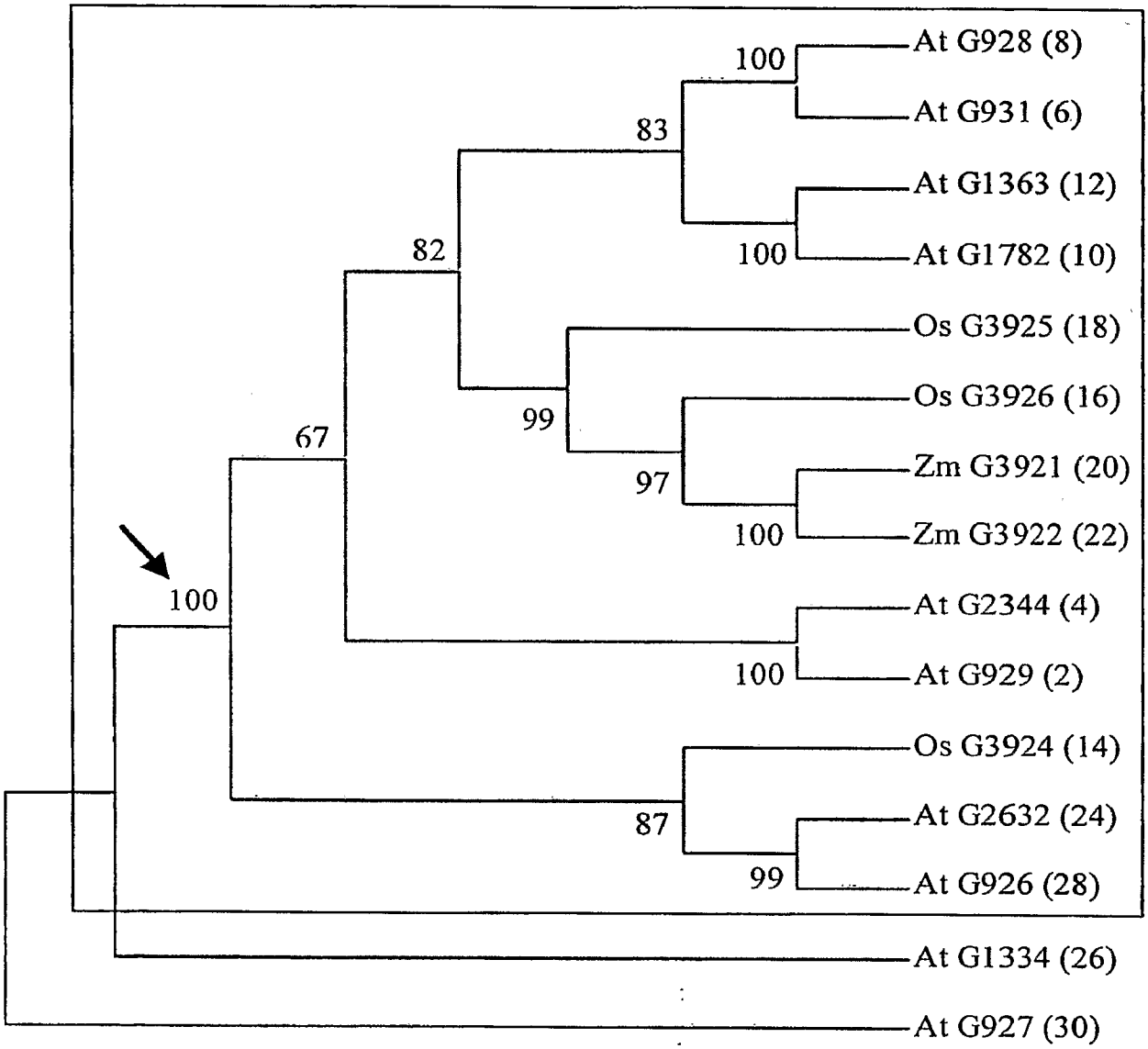

At G928 (8)
At G931 (6)
At G1363 (12)
At G1782 (10)
Os G3925 (18)
Os G3926 (16)
Zm G3921 (20)
Zm G3922 (22)
At G2344 (4)
At G929 (2)
Os G3924 (14)
At G2632 (24)
At G926 (28)
At G1334 (26)
At G927 (30)
100
83
82
100
99
97
100
67
100
100
87
99

FIG. 2

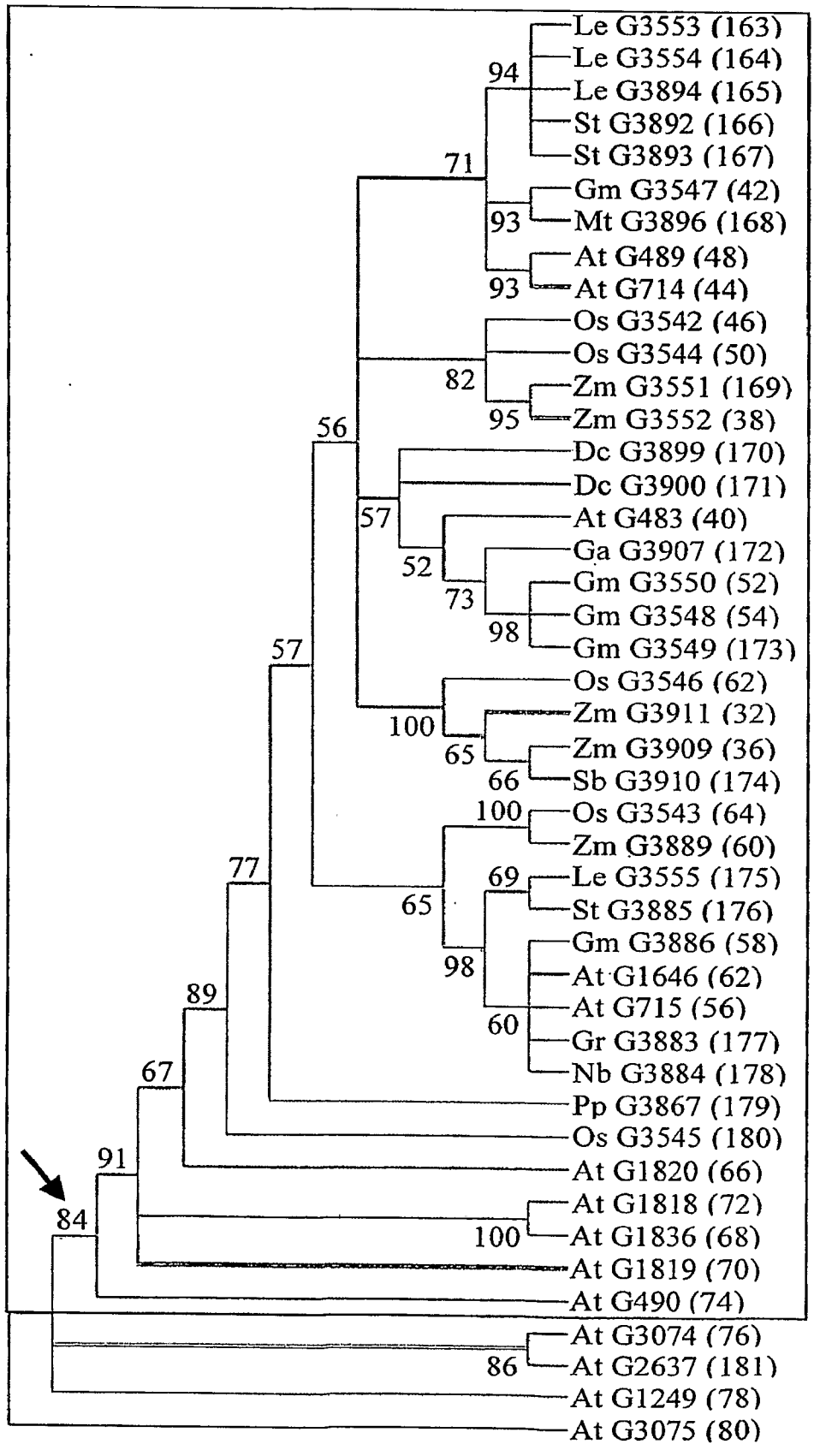

Le G3553 (163)
Le G3554 (164)
Le G3894 (165)
St G3892 (166)
St G3893 (167)
Gm G3547 (42)
Mt G3896 (168)
At G489 (48)
At G714 (44)
Os G3542 (46)
Os G3544 (50)
Zm G3551 (169)
Zm G3552 (38)
Dc G3899 (170)
Dc G3900 (171)
At G483 (40)
Ga G3907 (172)
Gm G3550 (52)
Gm G3548 (54)
Gm G3549 (173)
Os G3546 (62)
Zm G3911 (32)
Zm G3909 (36)
Sb G3910 (174)
Os G3543 (64)
Zm G3889 (60)
Le G3555 (175)
St G3885 (176)
Gm G3886 (58)
At G1646 (62)
At G715 (56)
Gr G3883 (177)
Nb G3884 (178)
Pp G3867 (179)
Os G3545 (180)
At G1820 (66)
At G1818 (72)
At G1836 (68)
At G1819 (70)
At G490 (74)
At G3074 (76)
At G2637 (181)
At G1249 (78)
At G3075 (80)
94
71
93
93
82
95
56
57
52
73
98
57
100
65
66
100
69
65
98
89
60
77
67
91
84
100
86

FIG. 3

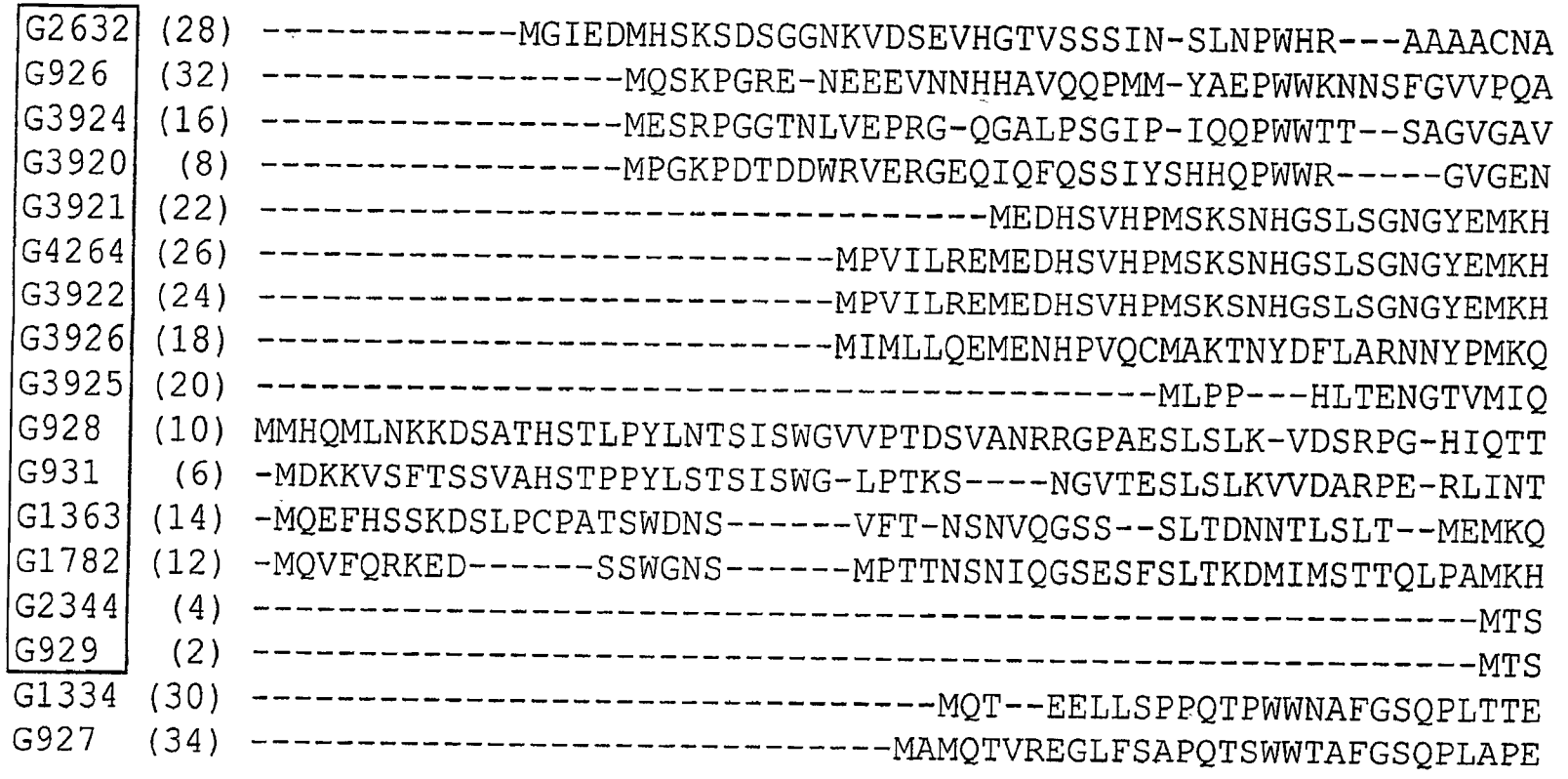


FIG. 4A

WO 2008/021021 PCT/US2007/017321

```
G2632 (28) NSSVEAGDKSSKSIALALESNGSKSPS-----NRDN-TVNKESQVTTSPQSAGDYSDKNQ
G926  (32) RPSGIPSNSSSLDCPNGSESNDVHSASEDGALNGENDGTWKDSQAATSSRSDNHGMEGND
G3924 (16) SPAVVAPGSGAGISLSGRDGGGD----------DAAEESSDDSRRSGETKDGSTDQEKHH
G3920  (8) -----ASKSSSDDQLNGSIVNG-----------ITRSETNDKSGGGVAKEY---QNIKH
G3921 (22) PGHEVCDRDSS-SESDR-SHQEASAASESSPD-EHTSTQ-----SD-NDEDHGKDNQDTL
G4264 (26) SGHKVCDRDSS-SESDR-SHQEASAASESSPN-EHTSTQ-----SD-NDEDHGKDNQDTM
G3922 (24) SGHKVCDRDSS-SESDR-SHQEASAASESSPN-EHTSTQ-----SD-NDEDHGKDNQDTM
G3926 (18) LVQRNSDGDSSPTKSGE-SHQEASAVSDSSLNGQHTSPQSVFVPSDINNNDSCGERDHGT
G3925 (20) FGHKMPDYESSATQSTSGSPREVSGMSEGSLNEQNDQSG--------NLDGYTKSDEGKM
G928  (10) KQISFQDQDSSSTQSTGQSYTEVASSGDDNPSRQISFSAKS----GSEITQRKGFASNPK
G931   (6) KNISFQDQDSSSTLSSAQSSNDVTSSGDDNPSRQISFLAHSDVCKGFEETQRKRFAIKS-
G1363 (14) TGFQMQHYDSSSTQSTGGES-YSEVASLSEPTNRYGHNIVVTHLSGYKENPENPIG--SH
G1782 (12) SGLQLQNQDSTSSQSTEEESGGGEVASFGEYK-RYGCSIVNNNLSGYIENLGKPIENYTK
G2344  (4) SIHELSDNIGSHEKQEQRDSHFQPPIPSARNYESIVTSLVYSDPGTTNS----------
G929   (2) SVHELSDNNESHAKKERPDSQTRPQVPSGRSSESIDTNSVYSEP---------------
G1334 (30) SLSGEASDSFTGVKAVTTEAEQGVVDKQT-STTLFTFSPGGEKSSRDVPKPH-VAFAMQS
G927  (34) SLAGD-SDSFAGVKVGSVGETRQRVDKQSNSATHLAFSLGDVKSPRLVPKPHGATFSMQS
```

FIG. 4B

```
G2632  (28)  ESLHHG--ITQPPPHP-QLVGHTVGWASSNPYQDPYYAGVMGAYGHHPLGFVP-------
G926   (32)  PALSIRNMHDQPLVQPPELVGHYIACVP-NPYQDPYYGGLMGAYGHQQLGFRP-------
G3924  (16)  ATSQMTALASDYLTPFSQLELNQPIASAAYQYPDSYYMGMVGPYGPQAMSAQTH------
G3920   (8)  AMLSTPFTMEKHLAPNPQMELVGHSVVLTSPYSDAQYGQILTTYGQQVMINP--------
G3921  (22)  KPVLSLGKEGSATGAPKLHYSPSFACI--PYTADAYYGAVGVLTGYPPHAIV--------
G4264  (26)  KPVLSLGKEGSAFLAPKLHYSPSFACI--PYTSDAYYSAVGVLTGYPPHAIV--------
G3922  (24)  KPVLSLGKEGSAFLAPKLHYSPSFACI--PYTADAYYSGVGVSTGYAPHAIVCSLLIFQF
G3926  (18)  KSVLSLGNTEAAFPPSKFDYNQPFACVSYPYGTDPYYG--GVSTGYTSHAFV--------
G3925  (20)  MSALSLGKSETVYAHSEPDRSQPFGIS--YPYADSFYG--GAVATYGTHAIM--------
G928   (10)  QGSMTGFPNIHFAPA-----QANFSFH----YADP-HYGGLLAATYLPQAPTC-------
G931    (6)  GSSTAGIADIHSSPS-----KANFSFH----YADP-HFGGLMPAAYLPQATIW-------
G1363  (14)  SISK---VSQDSVVLPIEA--ASWPLH----GNVTPHFNGFLSFPYASQHTVQ-------
G1782  (12)  SITTSSMVSQDSVFPAPTSGQISWSLQ----CAETSHFNGFLAPEYASTPTALP------
G2344   (4)  --------------------MAPGQYP----YPDPYYRSIFAPP---PQPYTGV------
G929    (2)  --------------------MAHGLYP----YPDPYYRSVFAQQAYLPHPYPGV------
G1334  (30)  ACFEFG--------------FAQPMMYTKHPHVEQ-YYGVVSAYGSQRSSGRVMI-----
G927   (34)  PCLELG--------------FSQPPIYTKYPYGEQQYYGVVSAYGSQ---SRVML-----
```

FIG. 4C

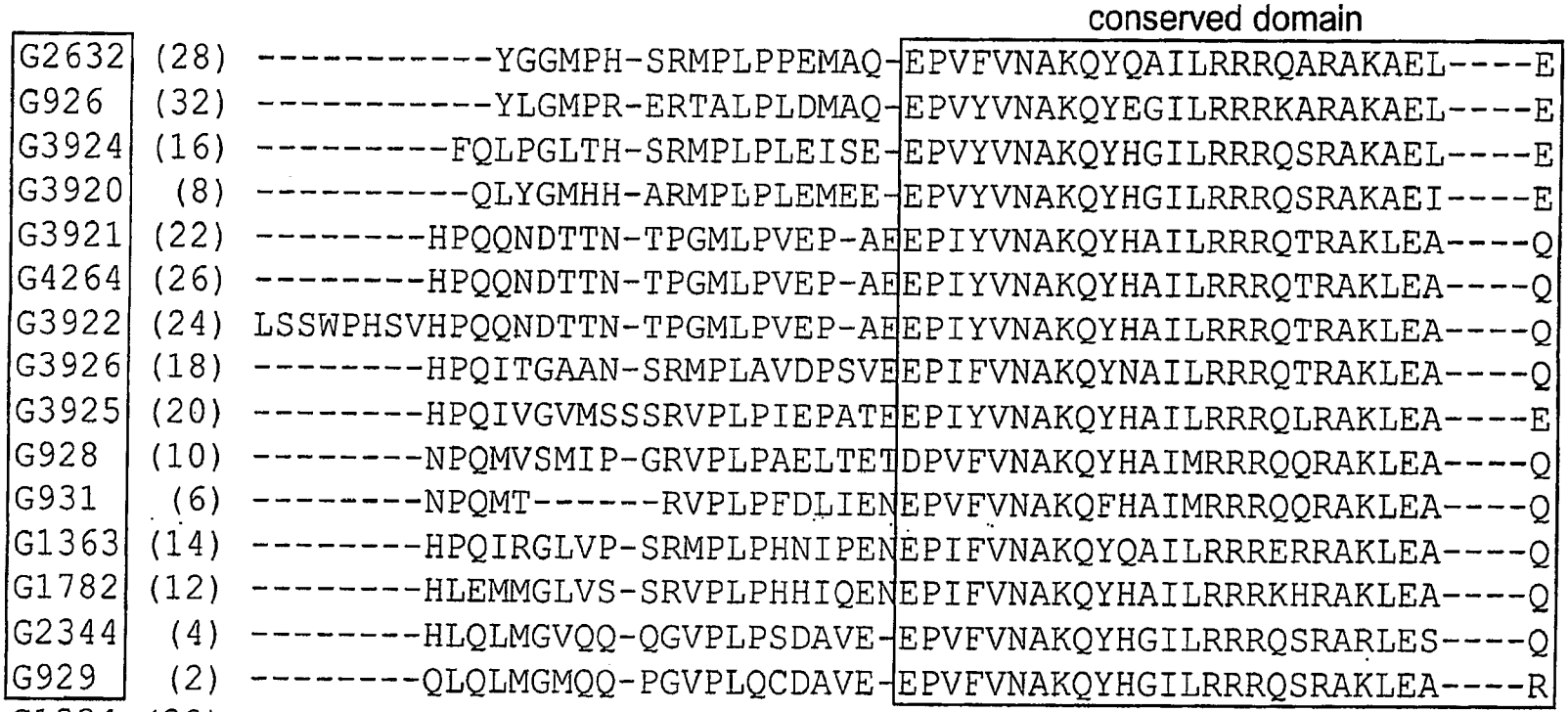

conserved domain
G2632 (28) ----------YGGMPH-SRMPLPPEMAQ-EPVFVNAKQYQAILRRRQARAKAEL----E
G926 (32) ----------YLGMPR-ERTALPLDMAQ-EPVYVNAKQYEGILRRRKARAKAEL----E
G3924 (16) ---------FQLPGLTH-SRMPLPLEISE-EPVYVNAKQYHGILRRRQSRAKAEL----E
G3920 (8) ----------QLYGMHH-ARMPLPLEMEE-EPVYVNAKQYHGILRRRQSRAKAEI----E
G3921 (22) --------HPQQNDTTN-TPGMLPVEP-AEEPIYVNAKQYHAILRRRQTRAKLEA----Q
G4264 (26) --------HPQQNDTTN-TPGMLPVEP-AEEPIYVNAKQYHAILRRRQTRAKLEA----Q
G3922 (24) LSSWPHSVHPQQNDTTN-TPGMLPVEP-AEEPIYVNAKQYHAILRRRQTRAKLEA----Q
G3926 (18) --------HPQITGAAN-SRMPLAVDPSVEEPIFVNAKQYNAILRRRQTRAKLEA----Q
G3925 (20) --------HPQIVGVMSSSRVPLPIEPATEEPIYVNAKQYHAILRRRQLRAKLEA----E
G928 (10) --------NPQMVSMIP-GRVPLPAELTETDPVFVNAKQYHAIMRRRQQRAKLEA----Q
G931 (6) --------NPQMT------RVPLPFDLIENEPVFVNAKQFHAIMRRRQQRAKLEA----Q
G1363 (14) --------HPQIRGLVP-SRMPLPHNIPENEPIFVNAKQYQAILRRRERRAKLEA----Q
G1782 (12) --------HLEMMGLVS-SRVPLPHHIQENEPIFVNAKQYHAILRRRKHRAKLEA----Q
G2344 (4) --------HLQLMGVQQ-QGVPLPSDAVE-EPVFVNAKQYHGILRRRQSRARLES----Q
G929 (2) --------QLQLMGMQQ-PGVPLQCDAVE-EPVFVNAKQYHGILRRRQSRAKLEA----R
G1334 (30) ----------PLKMETEEDGTIYVNSKQYHGIIRRRQSRAKA------E
G927 (34) ----------PLNMET-EDSTIYVNSKQYHGIIRRRQSRAKAAAVLDQK

FIG. 4D

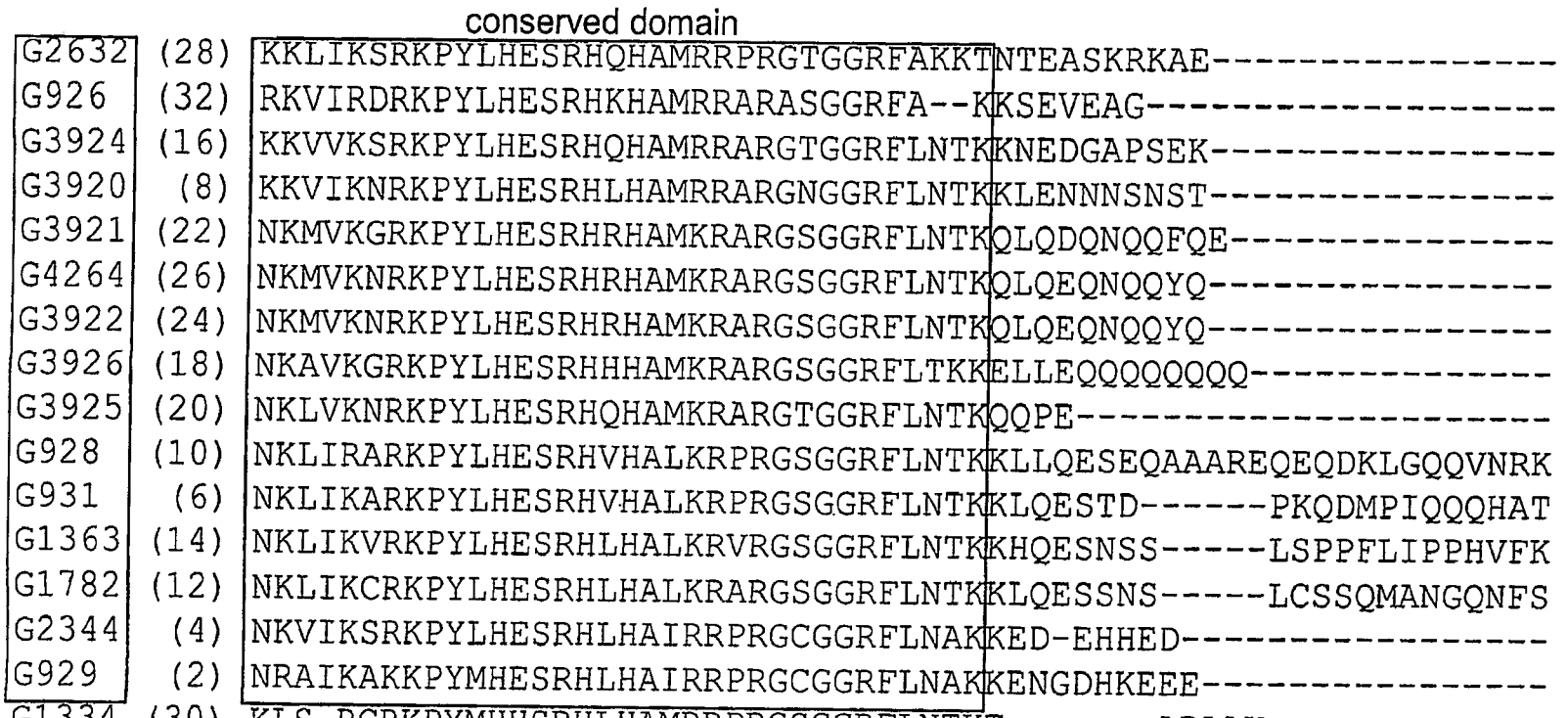


FIG. 4E

WO 2008/021021 PCT/US2007/017321

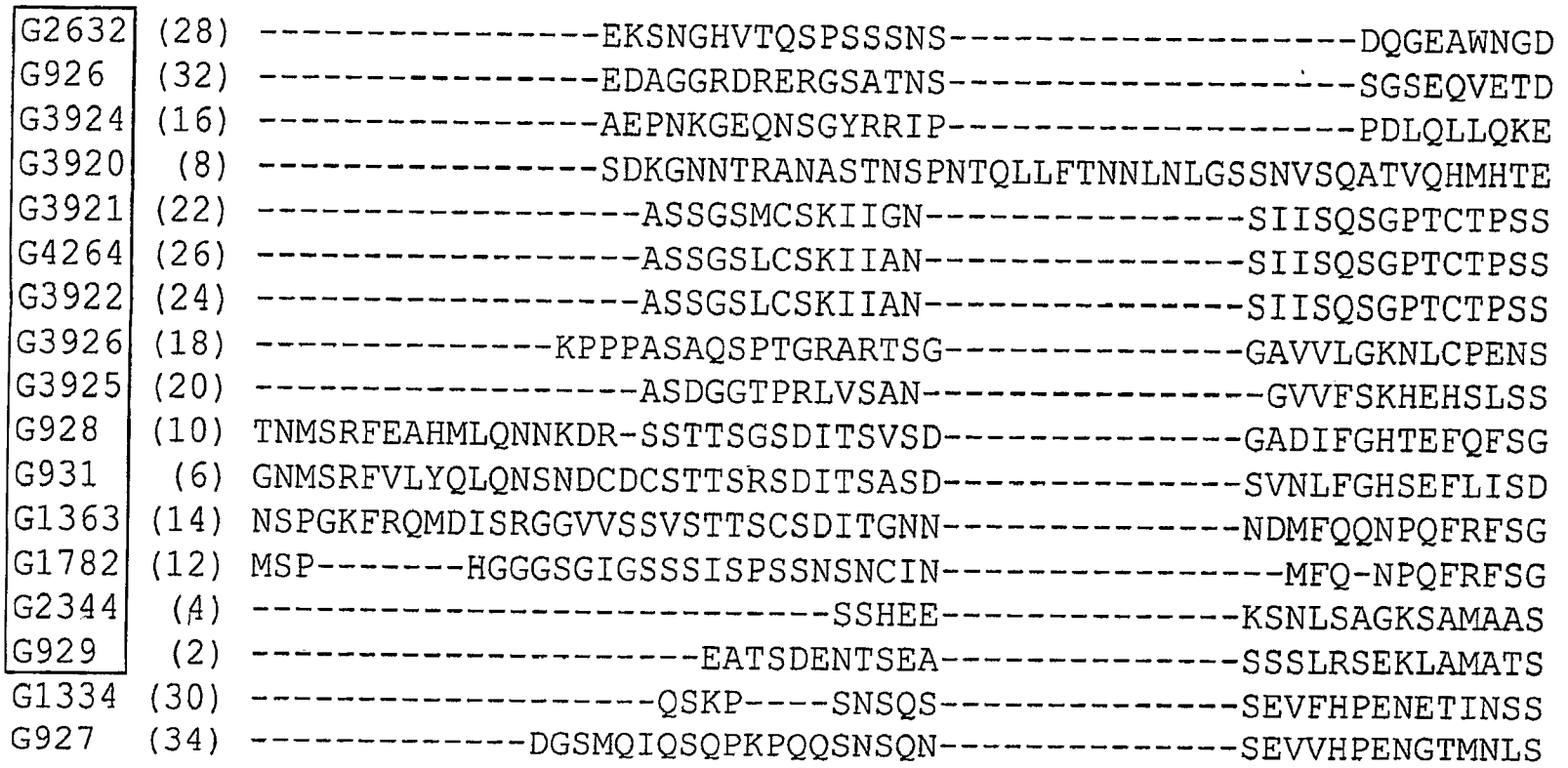


FIG. 4F

WO 2008/021021 PCT/US2007/017321

```
G2632  (28)  YRTPQGDEMQSSAYKRREEGECSGQQWNSLSSNHPSQARLAIK---------
G926   (32)  ----------------------SNETLN--SSGAP----------------
G3924  (16)  T-----------------------------------------------
G3920   (8)  QSFTIGYHNGNGLTALYRSQANGKKEGNCFGKERDPNGDFK----------
G3921  (22)  GTAGASTAS-QDRSCLPSVGFRPTTNFSDQGRGGLKLAVIGMQQRVSTIR--
G4264  (26)  GTAGASTAG-QDRSCLPSVGFRPTTNFSDQGRGGLKLAVIGMQQRVSTIR--
G3922  (24)  DTAGLQQPA-RTAAACPRWASAPQP---------TSVSKVEAARSWS-----
G3926  (18)  TSCSPSTPTGSEISSISFGGGMLAHQEHISFASADRHPTMNQNHRVPVMR--
G3925  (20)  ----------SDLHHRAKEGA-------------------------------
G928   (10)  FPTPINRAMLVHGQSNDMHGGGDMHHFSVHI--------------------
G931    (6)  CPSQTNPTMYVHGQSNDMHGGRNTHHFSVHI--------------------
G1363  (14)  YPSNHHVSVLM-------------------------------------
G1782  (12)  YPSTHHASALMSGT-----------------------------------
G2344   (4)  --SGTS-----------------------------------------
G929    (2)  GPNGRS-----------------------------------------
G1334  (30)  REANESNLSDSAVTSMDYFLSSSAYSPGGMVMPIKWN--AAAMDIGCCKLNI
G927   (34)  ---NGLNVSGSEVTSMNYFLSSPVHSLGGMVMPSKWIAAAAAMDNGCCNFKT
```

FIG. 4G

WO 2008/021021 PCT/US2007/017321

```
G3547  (46)   MDHQGHS---QNPSMGVVGSGAQLAYGSNPYQPGQITG--PPGSVVTSVGTIQSTGQPAG
G3894  (174)  MDQHGNG---QPPGIGVVTSSA--PIYGAPYQANQMAGPSPPAVSAGAIQSPQAAGLAAS
G3892  (175)  MDHHGNG---QPP-------------------------------VSAGAIQSPQAAGLSAS
G489   (52)   MDQQDHG---QSGAMN-------------------------YGTNPYQTNPMSTTAATVA
G714   (48)   MDQQ--G---QSSAMN-------------------------YGSNPYQTNAMTTTPT---
G3542  (50)   MEPSSQPQPVMGVATGGSQAYPP--PAAAYPPQAMVPGAPAVVPPGSQPSAPFPTNPAQL
G3544  (54)   MEPSSQPQPAIGVVAGGSQVYP------AYRPAATVPTAPAVIPAGSQPAPSFPANPDQL
G3551  (178)  MEPSPQP---MGVAAGGSQVYP----ASAYPPAATVAPA-SVVSAGLQSGQPFPANPGHM
G3552  (42)   MEPSPQP---MGVAAGGSQVYP----ASAYPPAATVAPA-SVVSAGLQSGQPFPANPGHM
G3548  (58)   ----------------MGVATG-ASQMAYSSHYPTAPMVASGTPAVAVPSPTQAPAAFS
G3549  (182)  MDKSEQT-QQQQQQQHVMGVAAG-ASQMAYSSHYPTASMVASGTPAVTAPSPTQAPAAFS
G3550  (56)   MDKSEQTQQQQQQQQHVMGVAAG-ASQMAYSSHYPTASMVASGTPAVTAPSPTQAPAAFS
G483   (44)   MEQSEEG--QQQQQQGVMDYVP----PHAYQ----------SGP---------------V
G3899  (179)  MDESEEP---QQQQEAVIDSAS--QMTYGVPHYHAVGLGVATGTPVVPVSAPTQHPTG-T
G1646  (66)   MDNNNNNNNQQPP----------------------------PTSVYPPGSAVTTVIPPPP
G715   (60)   MDTNN----QQPP----------------------------PSAAG---------IPPPP
G3883  (186)  MDSNQQTQS-----------------------------TPYPP-------QPPTSA
G3886  (62)   METNNQQQQQQGA-------------------------QAQSGP-------YPVAGA
G3884  (187)  MENN-QQS-----------------------------AANAA----------A
G3543  (68)   MDNQQLPY-AGQP----------------------AAAGAG----------APV
G3889  (64)   MDNQPLPYSTGQP----------------------PAPG-G----------APV
G3909  (40)   MEP-KSTTPPPPP-----------------------VMGAPIAYPPPPGAAYPAGPY
G3546  (38)   MEP-KSTTPPPPPPPP-------------------VLGAPVPYPPAGAYPPPVGPY
G3911  (36)   MDPNKSSTPPPPP-----------------------VMGAPVAYPPP-AYPPGVAAG
G1818  (76)   MENNNNNH-----------------------------------------------
G1836  (72)   MENNNGNN-----------------------------------------------
G1819  (74)   MEENNGNNNHYLP-----------------------------------------QPSSS
G1820  (70)   MAENNNNNGDNMN-------------------------------------------NDNHQ
G490   (78)   MRRPKSSHVRMEP-----------------------------------------VAPRSHN
G1249  (82)   -----------------------------------------------------------
G3074  (80)   MRKKLDTRFPAARIKKIMQADEDVG--------KIALAVPVLVSKSLELFLQDLCDRTYE
G3075  (84)   MVSSKKPKEKKARSDVVVNKASG---------RSKRSSGSRTKKTSNKVNIVKKKPEIYE
```

FIG. 5A

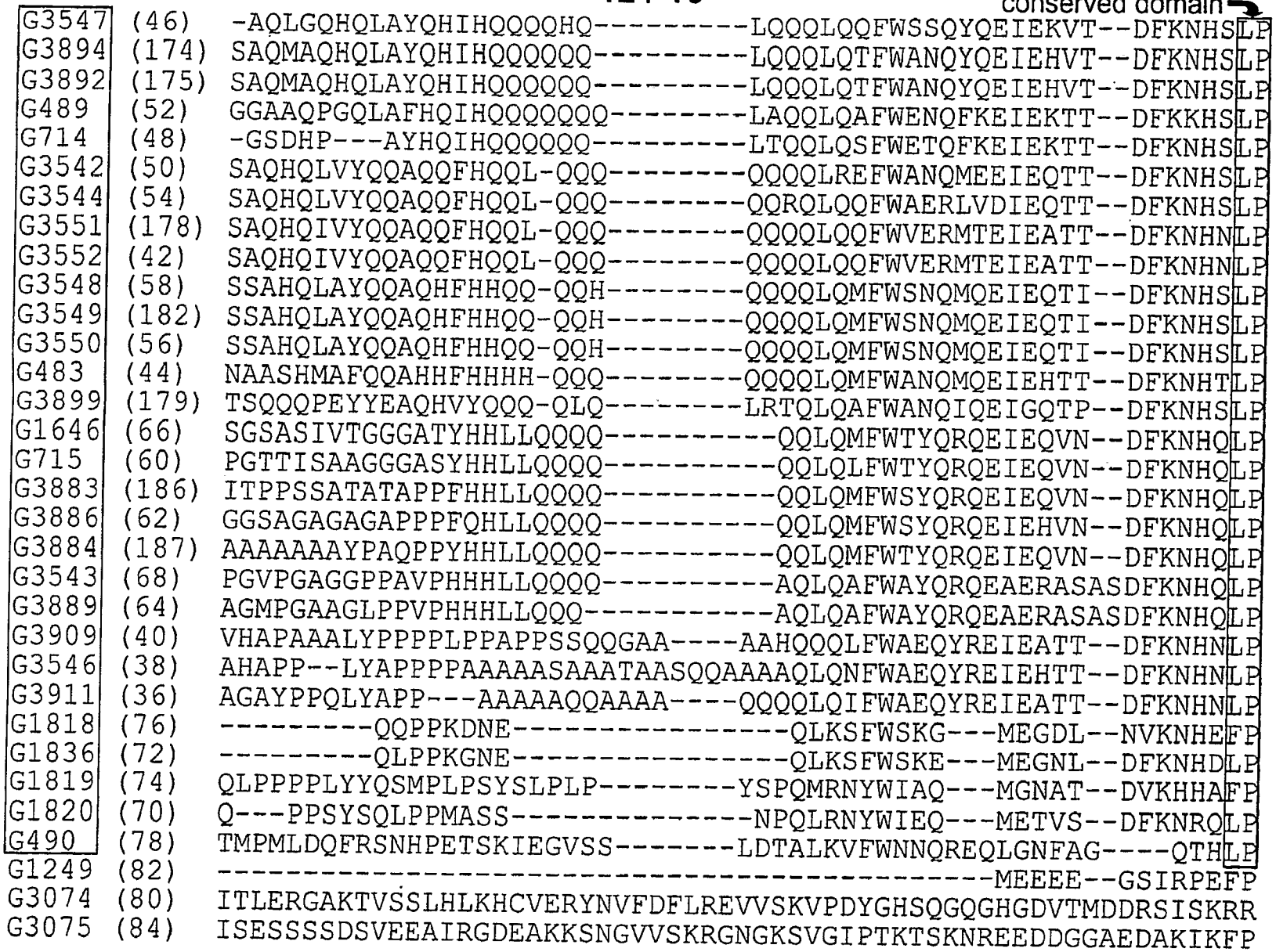

conserved domain
G3547 (46) -AQLGQHQLAYQHIHQQQQHQ---------LQQQLQQFWSSQYQEIEKVT--DFKNHSLP
G3894 (174) SAQMAQHQLAYQHIHQQQQQQ---------LQQQLQTFWANQYQEIEHVT--DFKNHSLP
G3892 (175) SAQMAQHQLAYQHIHQQQQQQ---------LQQQLQTFWANQYQEIEHVT--DFKNHSLP
G489 (52) GGAAQPGQLAFHQIHQQQQQQQ--------LAQQLQAFWENQFKEIEKTT--DFKKHSLP
G714 (48) -GSDHP---AYHQIHQQQQQQ---------LTQQLQSFWETQFKEIEKTT--DFKNHSLP
G3542 (50) SAQHQLVYQQAQQFHQQL-QQQ--------QQQQLREFWANQMEEIEQTT--DFKNHSLP
G3544 (54) SAQHQLVYQQAQQFHQQL-QQQ--------QQRQLQQFWAERLVDIEQTT--DFKNHSLP
G3551 (178) SAQHQIVYQQAQQFHQQL-QQQ--------QQQQLQQFWVERMTEIEATT--DFKNHNLP
G3552 (42) SAQHQIVYQQAQQFHQQL-QQQ--------QQQQLQQFWVERMTEIEATT--DFKNHNLP
G3548 (58) SSAHQLAYQQAQHFHHQQ-QQH--------QQQQLQMFWSNQMQEIEQTI--DFKNHSLP
G3549 (182) SSAHQLAYQQAQHFHHQQ-QQH--------QQQQLQMFWSNQMQEIEQTI--DFKNHSLP
G3550 (56) SSAHQLAYQQAQHFHHQQ-QQH--------QQQQLQMFWSNQMQEIEQTI--DFKNHSLP
G483 (44) NAASHMAFQQAHHFHHHH-QQQ--------QQQQLQMFWANQMQEIEHTT--DFKNHTLP
G3899 (179) TSQQQPEYYEAQHVYQQQ-QLQ--------LRTQLQAFWANQIQEIGQTP--DFKNHSLP
G1646 (66) SGSASIVTGGGATYHHLLQQQQ----------QQLQMFWTYQRQEIEQVN--DFKNHQLP
G715 (60) PGTTISAAGGGASYHHLLQQQQ----------QQLQLFWTYQRQEIEQVN--DFKNHQLP
G3883 (186) ITPPSSATATAPPFHHLLQQQQ----------QQLQMFWSYQRQEIEQVN--DFKNHQLP
G3886 (62) GGSAGAGAGAPPPFQHLLQQQQ----------QQLQMFWSYQRQEIEHVN--DFKNHQLP
G3884 (187) AAAAAAAYPAQPPYHHLLQQQQ----------QQLQMFWTYQRQEIEQVN--DFKNHQLP
G3543 (68) PGVPGAGGPPAVPHHHLLQQQQ----------AQLQAFWAYQRQEAERASASDFKNHQLP
G3889 (64) AGMPGAAGLPPVPHHHLLQQQ----------AQLQAFWAYQRQEAERASASDFKNHQLP
G3909 (40) VHAPAAALYPPPPLPPAPPSSQQGAA----AAHQQQLFWAEQYREIEATT--DFKNHNLP
G3546 (38) AHAPP--LYAPPPPAAAAASAAATAASQQAAAAQLQNFWAEQYREIEHTT--DFKNHNLP
G3911 (36) AGAYPPQLYAPP---AAAAAQQAAAA----QQQQLQIFWAEQYREIEATT--DFKNHNLP
G1818 (76) ---------QQPPKDNE----------QLKSFWSKG---MEGDL--NVKNHEFP
G1836 (72) ---------QLPPKGNE----------QLKSFWSKE---MEGNL--DFKNHDLP
G1819 (74) QLPPPPLYYQSMPLPSYSLPLP--------YSPQMRNYWIAQ---MGNAT--DVKHHAFP
G1820 (70) Q---PPSYSQLPPMASS----------NPQLRNYWIEQ---METVS--DFKNRQLP
G490 (78) TMPMLDQFRSNHPETSKIEGVSS-------LDTALKVFWNNQREQLGNFAG----QTHLP
G1249 (82) ----------MEEEE--GSIRPEFP
G3074 (80) ITLERGAKTVSSLHLKHCVERYNVFDFLREVVSKVPDYGHSQGQGHGDVTMDDRSISKRR
G3075 (84) ISESSSSDSVEEAIRGDEAKKSNGVVSKRGNGKSVGIPTKTSKNREEDDGGAEDAKIKFP

FIG. 5B

conserved domain

G3547	(46)	LARIKKIMKADEDVRMISAEAP------VIFARACEMFILELTLRSWNHTEENKRRTLQK
G3894	(174)	LARIKKIMKADEDVRMISAEAP------VVFARACEMFILELTLRAWNHTEENKRRTLQK
G3892	(175)	LARIKKIMKADEDVRMISAEAP------VVFARACEMFILELTLRAWNHTEENKRRTLQK
G489	(52)	LARIKKIMKADEDVRMISAEAP------VVFARACEMFILELTLRSWNHTEENKRRTLQK
G714	(48)	LARIKKIMKADEDVRMISAEAP------VVFARACEMFILELTLRSWNHTEENKRRTLQK
G3542	(50)	LARIKKIMKADEDVRMISAEAP------VVFAKACEVFILELTLRSWMHTEENKRRTLQK
G3544	(54)	LARIKKIMKADEDVRMISAEAP------VIFAKACEIFILELTLRSWMHTEENKRRTLQK
G3551	(178)	LARIKKIMKADEDVRMISAEAP------VVFAKACEIFILELTLRSWMHTEVNKRRTLQK
G3552	(42)	LARIKKIMKADEDVRMISAEAP------VVFAKACEIFILELTLRSWMHTEENKRRTLQK
G3548	(58)	LARIKKIMKADEDVRMISAEAP------VIFAKACEMFILELTLRSWIHTEENKRRTLQK
G3549	(182)	LARIKKIMKADEDVRMISAEAP------VIFAKACEMFILELTLRSWIHTEENKRRTLQK
G3550	(56)	LARIKKIMKADEDVRMISAEAP------VIFAKACEMFILELTLRSWIHTEENKRRTLQK
G483	(44)	LARIKKIMKADEDVRMISAEAP------VIFAKACEMFILELTLRAWIHTEENKRRTLQK
G3899	(179)	LARIKKIMKADEDVRMISSEAP------VIFAKACEMFILELTMRSWLLTEENKRRTLQK
G1646	(66)	LARIKKIMKADEDVRMISAEAP------ILFAKACELFILELTIRSWLHAEENKRRTLQK
G715	(60)	LARIKKIMKADEDVRMISAEAP------ILFAKACELFILELTIRSWLHAEENKRRTLQK
G3883	(186)	LARIKKIMKADEDVRMISAEAP------ILFAKACELFILELTIRSWLHAEENKRRTLQK
G3886	(62)	LARIKKIMKADEDVRMISAEAP------ILFAKACELFILELTIRSWLHAEENKRRTLQK
G3884	(187)	LARIKKIMKADEDVRMISAEAP------ILFAKACELFILELTIRSWLHAEENKRRTLQK
G3543	(68)	LAGIKKIMKADEDVRMISAEAP------VLFAKACELFILELTIRSWLHAEENKRRTLQR
G3889	(64)	LARIKKIMKADEDVRMISAEAP------VLFAKACELFILELTIRSWLHAEENKRRTLQR
G3909	(40)	LARIKKIMKADEDVRMIAAEAP------VVFSRACEMFILELTHRGWAHAEENKRRTLQK
G3546	(38)	LARIKKIMKADEDVRMIAAEAP------VVFARACEMFILELTHRGWAHAEENKRRTLQK
G3911	(36)	LARIKKIMKADEDVRMIAAEAP------VVFARACEMFILELTHRGWAHAEENKRRTLQK
G1818	(76)	ISRIKRIMKFDPDVSMIAAEAP------NLLSKACEMFVMDLTMRSWLHAQESNRLTIRK
G1836	(72)	ITRIKKIMKYDPDVTMIASEAP------ILLSKACEMFIMDLTMRSWLHAQESKRVTLQK
G1819	(74)	LTRIKKIMKSNPEVNMVTAEAP------VLISKACEMLILDLTMRSWLHTVEGGRQTLKR
G1820	(70)	LARIKKIMKADPDVHMVSAEAP------IIFAKACEMFIVDLTMRSWLKAEENKRHTLQK
G490	(78)	LSRVRKILKSDPEVKKISCDVP------ALFSKACEYFILEVTLRAWMHTQSCTRETIRR
G1249	(82)	IGRVKKIMKLDKDINKINSEAL------HVITYSTELFLHFLAEKSAVVTAEKKRKTVNL
G3074	(80)	KPISDEVNDSDEEYKKSKTQEIGS----AKTSGRGGRGRGRGRGRGGRAAKAAEREGLNR
G3075	(84)	MNRIRRIMRSDNSAPQIMQDAV------FLVNKATEMFIERFSEEAYDSSVKDKKKFIHY

FIG. 5C

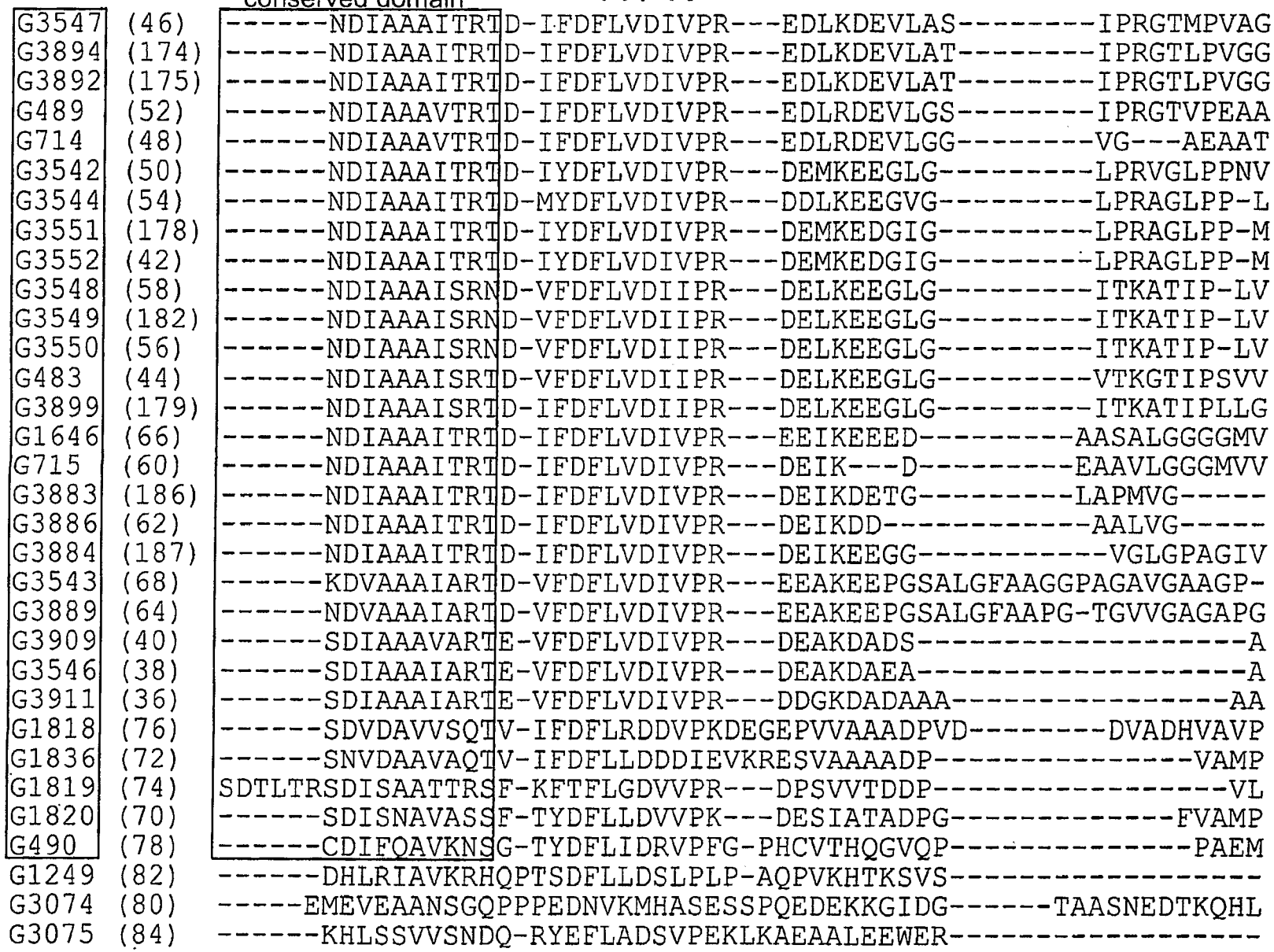

conserved domain
G3547 (46) ------NDIAAAITRTD-IFDFLVDIVPR---EDLKDEVLAS--------IPRGTMPVAG
G3894 (174) ------NDIAAAITRTD-IFDFLVDIVPR---EDLKDEVLAT--------IPRGTLPVGG
G3892 (175) ------NDIAAAITRTD-IFDFLVDIVPR---EDLKDEVLAT--------IPRGTLPVGG
G489 (52) ------NDIAAAVTRTD-IFDFLVDIVPR---EDLRDEVLGS--------IPRGTVPEAA
G714 (48) ------NDIAAAVTRTD-IFDFLVDIVPR---EDLRDEVLGG--------VG---AEAAT
G3542 (50) ------NDIAAAITRTD-IYDFLVDIVPR---DEMKEEGLG---------LPRVGLPPNV
G3544 (54) ------NDIAAAITRTD-MYDFLVDIVPR---DDLKEEGVG---------LPRAGLPP-L
G3551 (178) ------NDIAAAITRTD-IYDFLVDIVPR---DEMKEDGIG---------LPRAGLPP-M
G3552 (42) ------NDIAAAITRTD-IYDFLVDIVPR---DEMKEDGIG---------LPRAGLPP-M
G3548 (58) ------NDIAAAISRND-VFDFLVDIIPR---DELKEEGLG---------ITKATIP-LV
G3549 (182) ------NDIAAAISRND-VFDFLVDIIPR---DELKEEGLG---------ITKATIP-LV
G3550 (56) ------NDIAAAISRND-VFDFLVDIIPR---DELKEEGLG---------ITKATIP-LV
G483 (44) ------NDIAAAISRTD-VFDFLVDIIPR---DELKEEGLG---------VTKGTIPSVV
G3899 (179) ------NDIAAAISRTD-IFDFLVDIIPR---DELKEEGLG---------ITKATIPLLG
G1646 (66) ------NDIAAAITRTD-IFDFLVDIVPR---EEIKEEED---------AASALGGGGMV
G715 (60) ------NDIAAAITRTD-IFDFLVDIVPR---DEIK---D---------EAAVLGGGMVV
G3883 (186) ------NDIAAAITRTD-IFDFLVDIVPR---DEIKDETG---------LAPMVG-----
G3886 (62) ------NDIAAAITRTD-IFDFLVDIVPR---DEIKDD----------AALVG-----
G3884 (187) ------NDIAAAITRTD-IFDFLVDIVPR---DEIKEEGG----------VGLGPAGIV
G3543 (68) ------KDVAAAIARTD-VFDFLVDIVPR---EEAKEEPGSALGFAAGGPAGAVGAAGP-
G3889 (64) ------NDVAAAIARTD-VFDFLVDIVPR---EEAKEEPGSALGFAAPG-TGVVGAGAPG
G3909 (40) ------SDIAAAVARTE-VFDFLVDIVPR---DEAKDADS----------------A
G3546 (38) ------SDIAAAIARTE-VFDFLVDIVPR---DEAKDAEA----------------A
G3911 (36) ------SDIAAAIARTE-VFDFLVDIVPR---DDGKDADAAA----------------AA
G1818 (76) ------SDVDAVVSQTV-IFDFLRDDVPKDEGEPVVAAADPVD--------DVADHVAVP
G1836 (72) ------SNVDAAVAQTV-IFDFLLDDDIEVKRESVAAAADP----------VAMP
G1819 (74) SDTLTRSDISAATTRSF-KFTFLGDVVPR---DPSVVTDDP----------VL
G1820 (70) ------SDISNAVASSF-TYDFLLDVVPK---DESIATADPG----------FVAMP
G490 (78) ------CDIFQAVKNSG-TYDFLIDRVPFG-PHCVTHQGVQP----------PAEM
G1249 (82) ------DHLRIAVKRHQPTSDFLLDSLPLP-AQPVKHTKSVS----------
G3074 (80) -----EMEVEAANSGQPPPEDNVKMHASESSPQEDEKKGIDG------TAASNEDTKQHL
G3075 (84) ------KHLSSVVSNDQ-RYEFLADSVPEKLKAEAALEEWER----------

FIG. 5D

```
G3547  (46)   --PADALPYCY-MPPQH----------PSQVGAAGVIMGKPVMDPNMYAQQSHPYMAP
G3894  (174)  --PTEGLPFYYGMPPQS----------AQPIGAPGMYMGKPVDQA-LYAQQPRPYMAQ
G3892  (175)  --PTEGLPFYYGMPPQS----------AQPIGAPGMYMGKPVDQA-LYAQQPRPFMAQ
G489   (52)   --AAG---YPYGYLPAG----------TAPIGNPGMVMGNPGG-----AYPPNPYMGQ
G714   (48)   --AAG---YPYGYLPPG----------TAPIGNPGMVMGNPG------AYPPKAYMGQ
G3542  (50)   GGAADTYPY-YYVPAQQ----------GPGSGMMYGGQQGHPVT--YVWQQPQEQQ
G3544  (54)   GVPADSYPYGYYVPQQQ----------VPGAGIAYGGQQGHPG---YLWQDPQEQQ
G3551  (178)  GAPADAYPY-YYMPQQQ----------VPGSGMVYGAQQGHPVT--YLWQEPQQQQ
G3552  (42)   GAPADAYPY-YYMPQQQ----------VPGSGMVYGAQQGHPVT--YLWQEPQQQQ
G3548  (58)   NSPAD-MPY-YYVPPQHPVVGPPGMIMGKPVGAEQATLYSTQQPRPPMAFMPWPHTQPQQ
G3549  (182)  NSPAD-MPY-YYVPPQHPVVGPPGMIMGKPVGAEQATLYSTQQPRPPMAFMPWPHTQPQQ
G3550  (56)   GSPAD-MPY-YYVPPQHPVVGPPGMIMGKPIGAEQATLYSTQQPRPPVAFMPWPHTQPLQ
G483   (44)   GSP----PY-YYLQQQ-------GMMQ----------------------------HWPQE
G3899  (179)  -SPADSAPY-YYVPQQH----------AVEQAGFYPDQQAHPQLPYMSWQ--QPHE
G1646  (66)   APAASGVPYYYPPMGQPAV----------PGGMMIGR--PAMDPS--GVYAQPP-SQAWQ
G715   (60)   APTASGVPYYYPPMGQPAG----------PGGMMIGR--PAMDPN--GVYVQPP-SQAWQ
G3883  (186)  -ATASGVPYFYPPMGQPAAGG--------PGGMMIGR--PAVDPTG-GIYGQPP-SQAWQ
G3886  (62)   -ATASGVPYYYPPIGQPAG----------MMIGR--PAVDPAT-GVYVQPP-SQAWQ
G3884  (187)  GSTASGVPYYYPPMGQPAP----------PGVMMGR--PAMPGVDPSMYVQPPPSQAWQ
G3543  (68)   ---AAGLPYYYPPMGQPAP----------MMPAWHVPAWDPAW-QQGAAPDVDQGAA
G3889  (64)   GAPAAGMPYYYPPMGQPAP----------MMPAWHVPAWDPAW-QQGAA-DVDQ--S
G3909  (40)   AMGAAGIP--HPAAGLPAA----------DPMG-YYYVQPPQ-----
G3546  (38)   AAVAAGIP--HPAAGLPAT----------DPMA-YYYVQPQ------
G3911  (36)   AAAAAGIP--RPAAGVPAT----------DPLA-YYYVPQQ------
G1818  (76)   DLNNE-----ELPPGTVIG----------TPVCYGLGIHAPHPQM---PGAWT-----EE
G1836  (72)   PIDDG-----ELPPGMVIG----------TPVCCSLGIHQPQPQMQAWPGAWTSVSGEEE
G1819  (74)   HPDGE-----VLPPGTVIG----------YPVFDCNGVYASPPQM----QEWPAVPGDGE
G1820  (70)   HPDGGGVPQYYYPPGVVMG----------TPMVGS-GMYAPS-------QAWPAAAGDGE
G490   (78)   ILPDMNVPIDMDQIEEE----------NMMEERSVGFDLNCDLQ
G1249  (82)   -------DKKIPAPPIG----------TRRIDDFFSKGKAKTDSA
G3074  (80)   QSPKEGIDFDLNAESLDLN--------ETKLAPATGTTTTTTAATDSEEYSGWPMMDISK
G3075  (84)   -----------GMTDAG----------
```

FIG. 5E

```
G3547  (46)   Q----------MWPQPPDQRQSSPEH--------------------------------
G3894  (174)  P----------IWPQQQQPPSDS----------------------------------
G3892  (175)  P----------IWPQQQQPPSDS----------------------------------
G489   (52)   P----------MWQQQAPDQPDQEN--------------------------------
G714   (48)   P----------MWQQPGPEQQDPDN--------------------------------
G3542  (50)   E----------EAPEEQHSLPESS--------------------------------
G3544  (54)   E----------EPPAEQQSD-----------------------------------
G3551  (178)  E----------QAPEEQQSA-----------------------------------
G3552  (42)   E----------QAPEEQQSA-----------------------------------
G3548  (58)   Q----------QPPQHQQTDS----------------------------------
G3549  (182)  Q----------QPPQHQQTDS----------------------------------
G3550  (56)   Q----------QPPQHQQTDS----------------------------------
G483   (44)   Q----------HPDES--------------------------------------
G3899  (179)  H----------KDQE--ENGD----------------------------------
G1646  (66)   SVWQNSAGGGDDVSYGSGGSSGHGNLDSQG-------------------------
G715   (60)   SVWQTSTGTGDDVSYGSGGSSGQGNLDGQG-------------------------
G3883  (186)  SVWQTAG--TDDGSYGSGVTGGQGNLDGQG-------------------------
G3886  (62)   SVWQSA---AEDASYGTGPAGAQRSLDGQS-------------------------
G3884  (187)  SVWQTA----EDNSYASGGSSGQGNLDGQS-------------------------
G3543  (68)   GSFSEEGQQGFAGHGGAAASFPPAPPSSESIWFHACIAEVLASYYSCSNQMLNVSVINCG
G3889  (64)   GSFSEEGQ-GFGAGHGGAASFPPAPPTSE-------------------------
G3909  (40)   ----------------------------------------------------
G3546  (38)   ----------------------------------------------------
G3911  (36)   ----------------------------------------------------
G1818  (76)   DAT----------GANGGN-GGN----------------------------------
G1836  (72)   EAR----------GKKGGD-DGN----------------------------------
G1819  (74)   EAA----------GEIGGSSGGN----------------------------------
G1820  (70)   DDA----------EDNGGNGGGN----------------------------------
G490   (78)   ----------------------------------------------------
G1249  (82)   ----------------------------------------------------
G3074  (80)   MDPAQLASLGKRIDEDEEDYDEEG----------------------------------
G3075  (84)   ----------------------------------------------------
```

FIG. 5F

G3547	(46)	----------
G3894	(174)	----------
G3892	(175)	----------
G489	(52)	----------
G714	(48)	----------
G3542	(50)	----------
G3544	(54)	----------
G3551	(178)	----------
G3552	(42)	----------
G3548	(58)	----------
G3549	(182)	----------
G3550	(56)	----------
G483	(44)	----------
G3899	(179)	----------
G1646	(66)	----------
G715	(60)	----------
G3883	(186)	----------
G3886	(62)	----------
G3884	(187)	----------
G3543	(68)	TYLNRCTWTLKLIKGCTIY
G3889	(64)	----------
G3909	(40)	----------
G3546	(38)	----------
G3911	(36)	----------
G1818	(76)	----------
G1836	(72)	----------
G1819	(74)	----------
G1820	(70)	----------
G490	(78)	----------
G1249	(82)	----------
G3074	(80)	----------
G3075	(84)	----------

FIG. 5G

FIG. 6