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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 May 2013 (16.05.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/071112 A1

(51) **International Patent Classification**:
C07H 21/04 (2006.01) ***C12P 21/04*** (2006.01)
C12N 1/00 (2006.01)

(21) **International Application Number**:
PCT/US2012/064457

(22) **International Filing Date**:
9 November 2012 (09.11.2012)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
61/557,971 10 November 2011 (10.11.2011) US

(71) **Applicant**: **MASCOMA CORPORATION** [US/US]; 67 Etna Road, Suite 300, Lebanon, New Hampshire 03766 (US).

(72) **Inventors**: **ARGYROS, D. Aaron**; 74 Runnals Road, White River Junction, Vermont 05001 (US). **CAIAZZA, Nicky**; P.O. Box 1484, Rancho Santa Fe, California 92067 (US). **BARRET, Trisha F.**; 2885 South Road, Bradford, Vermont 05033 (US). **WARNER, Anne K.**; 42 Wolf Road, Unit 832, Lebanon, New Hampshire 03766 (US).

(74) **Agents**: **JACKMAN, Peter A.** et al.; Sterne, Kessler, Goldstein & Fox PLLC, 1100 New York Avenue N.W., Washington, District of Columbia 20005-3934 (US).

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

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

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

WO 2013/071112 A1

(54) **Title**: A GENETICALLY MODIFIED STRAIN OF S. CEREVISIAE ENGINEERED TO FERMENT XYLOSE AND ARABINOSE

(57) **Abstract**: The present invention provides a microorganism capable of fermenting arabinose to a desired product such as ethanol. In some embodiments, the organism is also capable of fermenting xylose. In some embodiments, the organism is capable of fermenting arabinose and xylose, and expresses one or more cellulases.

A GENETICALLY MODIFIED STRAIN OF S. CEREVISIAE ENGINEERED TO FERMENT XYLOSE AND ARABINOSE

BACKGROUND OF THE INVENTION

[0001] Energy conversion, utilization and access underlie many of the great challenges of our time, including those associated with sustainability, environmental quality, security, and poverty. New applications of emerging technologies are required to respond to these challenges. Biotechnology, one of the most powerful of the emerging technologies, can give rise to important new energy conversion processes. Plant biomass and derivatives thereof are a resource for the biological conversion of energy to forms useful to humanity.

[0002] Among forms of plant biomass, lignocellulosic biomass ("biomass") is particularly well-suited for energy applications because of its large-scale availability, low cost, and environmentally benign production. In particular, many energy production and utilization cycles based on cellulosic biomass have near-zero greenhouse gas emissions on a life-cycle basis. The primary obstacle impeding the more widespread production of energy from biomass feedstocks is the general absence of low-cost technology for overcoming the recalcitrance of biomass materials to conversion into useful products. Lignocellulosic biomass contains carbohydrate fractions (*e.g.*, cellulose and hemicellulose) including pentose sugars (*e.g.*, xylose and arabinose) that can be converted into ethanol or other products such as lactic acid and acetic acid. In order to convert the lignocellulose fractions, the cellulose, hemicellulose, and pentoses must ultimately be converted into monosaccharides; it is this conversion step that has historically been problematic.

[0003] Biomass processing schemes involving enzymatic or microbial hydrolysis commonly involve four biologically mediated transformations: (1) the production of saccharolytic enzymes (cellulases and hemicellulases); (2) the hydrolysis of carbohydrate components present in pretreated biomass to sugars; (3) the fermentation of hexose sugars (*e.g.*, glucose, mannose, and galactose); and (4) the fermentation of pentose sugars (*e.g.*, xylose and arabinose). These four transformations occur in a single step in a process configuration called consolidated bioprocessing (CBP), which is distinguished from other

less highly integrated configurations in that it does not involve a dedicated process step for cellulase and/or hemicellulase production.

[0004] CBP offers the potential for lower cost and higher efficiency than processes featuring dedicated cellulase production. The benefits result in part from avoided capital costs, substrate and other raw materials, and utilities associated with cellulase production. In addition, several factors support the realization of higher rates of hydrolysis, and hence reduced reactor volume and capital investment using CBP, including enzyme-microbe synergy and the use of thermophilic organisms and/or complexed cellulase systems. Moreover, cellulose-adherent cellulolytic microorganisms are likely to compete successfully for products of cellulose hydrolysis with non-adhered microbes, *e.g.*, contaminants. Successful competition of desirable microbes increases the stability of industrial processes based on microbial cellulose utilization. Progress in developing CBP-enabling microorganisms is being made through two strategies: engineering naturally occurring cellulolytic microorganisms to improve product-related properties, such as yield and titer; and engineering non-cellulolytic organisms that exhibit high product yields and titers to express a heterologous cellulase and hemicellulase system enabling cellulose and hemicellulose utilization.

[0005] One way to meet the demand for ethanol production is to convert sugars found in biomass, *i.e.*, materials such as agricultural wastes, corn hulls, corncobs, cellulosic materials, and the like to produce ethanol. Efficient biomass conversion in large-scale industrial applications requires a microorganism that is able to tolerate high concentrations of sugar and ethanol, and which is able to ferment more than one sugar simultaneously.

[0006] Pentoses appear in great abundance in nature. As much as 40% of a lignocellulosic material can be comprised of pentoses (Ladisch *et al.*, "Process considerations in the enzymatic hydrolysis of biomass." *Enz. Microb. Technol.*, 5: 82-100. (1983)). By fermentation, pentoses can be converted to ethanol which can be used as a liquid fuel or a chemical feedstock. Although many microorganisms have the ability to ferment simple hexose sugars, the pentose sugars, xylose and arabinose, are among the most difficult sugars in biomass to metabolize. Some microorganisms can ferment pentoses to ethanol and other co-products, and microorganisms with improved ethanol production from pentose sugars have been genetically engineered. However, many of

these studies have been conducted in bacteria that are sensitive to low pH and high concentrations of ethanol. Therefore, their use in fermentations is associated with undesired co-product formation, and the level of ethanol they are capable of producing remains low.

[0007] Bakers' yeast (*Saccharomyces cerevisiae*) is the preferred microorganism for the production of ethanol (Hahn-Hägerdal, B., *et al.*, *Adv. Biochem. Eng. Biotechnol. 73*, 53–84 (2001)). Attributes in favor of this microbe are (i) high productivity at close to theoretical yields (0.51 g ethanol produced/g glucose used), (ii) high osmo- and ethanol tolerance, (iii) natural robustness in industrial processes, also (iv) being generally regarded as safe (GRAS) due to its long association with wine and bread making, and beer brewing. Furthermore, *S. cerevisiae* exhibits tolerance to inhibitors commonly found in hydrolysates resulting from biomass pretreatment. However, *S. cerevisiae* does not naturally break down components of cellulose, nor does it efficiently use pentose sugars.

[0008] Progress has been made in engineering *S. cerevisiae* to express heterologous enzymes that enable it to break down cellulose. (*See e.g.* U.S. Appl. No. 13/130,549 and PCT/US2011/039192, incorporated herein by reference in their entirety). However, utilization of arabinose for industrial ethanologenic fermentation has not been demonstrated in yeast. In addition, there is a need for an ethanologenic organism capable of efficiently utilizing arabinose and xylose that is also capable of breaking down cellulose. The highest products yields are obtained when all the cellulose and hemicellulose are broken down into monomer sugars and fermented into the desired product.

[0009] Therefore, there is a need in the art for an ethanologenic organism capable of fermenting pentose sugars in quantities sufficient for commercial applicability. There is also a need to combine efficient pentose utilization with cellulose digestion in order to maximize the efficiency of cellulosic feedstock use and to generate the highest yield of product.

BRIEF SUMMARY OF THE INVENTION

[0010] The present invention provides a microorganism capable of fermenting arabinose to a desired product such as ethanol. In some embodiments, the organism is also capable

of fermenting xylose. In some embodiments, the organism is capable of fermenting arabinose and xylose, and expresses one or more cellulases.

[0011] In some embodiments, the invention provides a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (R5PE).

[0012] In some embodiments, the invention provides a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (R5PE), wherein one or more of the AI, RK and R5PE is derived from an AI, RK and R5PE of *B. thetaiotamicron.*

[0013] In some embodiments, the recombinant eukaryotic host contains an AraT derived from an AraT of an organism selected from the group consisting of *Ambrosiozyma monospora* (LAT2), *Candida arabinofermentans*, *Ambrosiozyma monospora* (LAT1), *Kluveromyces marxianus* (LAT1), *Pichia guillermondii* (LAT1), *Pichia guillermondii* (LAT2), *Pichia stipites*, *Ambrosiozyma monospora* (LAT2), *Debaryomyces hensenii*, *Apergillus flavus*, *Aspergillus terreus*, *Neosartorya fischeri*, *Aspergillus niger*, *Penicillium marneffei*, *Coccidioides posadasii*, *Gibberella zeae*, *Magnaporthe oryzae*, *Schizophyllum commune*, *Pichia stipites*, *Saccaharomyces* HXT2, *Aspergillus clavatus* (ACLA_032060), *Sclerotinia sclerotiorum* (SS1G_01302), *Arthroderma benhamiae* (ARB_03323), *Trichophyton equinum* (TEQG_03356), *Trichophyton tonsurans* (G_04876), *Coccidioides immitis* (CIMG_09387), *Coccidioides posadasii* (CPSG_03942), *Coccidioides posadasii* (CPC735_017640), *Botryotinia fuckeliana* (BC1G_08389), *Pyrenophora tritici-repentis* (PTRG_10527), *Ustilago maydis* (UM03895.1), *Clavispora lusitaniae* (CLUG_02297), *Pichia guillermondii* (LAT1), *Pichia guillermondii* (LAT2), *Debaryomyces hansenii* (DEHA2E01166g), *Pichia stipites*, *candida albicans*, *Debaryomyces hansenii* (DEHA2B16082g), *Kluveromyces marxianus* (LAT1), *Kluyveromyces lactis* (KLLA-ORF10059), *Lachancea thermotolerans* (KLTH0H13728g), *Vanderwaltozyma polyspora* (Kpol_281p3), *Zygosaccharomyces rouxii* (ZYRO0E03916g), *Pichia pastoris* (0.1833), *Candida arabinofermentans* (0.1378), *Ambrosiozyma monospora* (LAT1), *Aspergillus clavatus* (ACLA_044740), *Neosartorya*

fischeri (NFIA_094320), *Aspergillus flavus* (AFLA_116400), *Aspergillus terreus* (ATEG_08609), *Aspergillus niger* (ANI_1_1064034), *Telaromyces stipitatus* (TSTA_124770), *Penicillium chrysogenum* (Pc20g01790), *Penicillium chrysogenum* (Pc20g01790)#2, *Gibberella zeae* (FG10921.1), *Nectria hematococco*, and *Glomerella graminicola* (GLRG_10740).

[0014] In some embodiments, the recombinant eukaryotic host cell comprises an AraT that encodes an amino acid sequence at least 80% identical to any one of the amino acid sequences of SEQ ID NOs: 9-20. In some embodiments, the recombinant eukaryotic host comprises a heterologous AI, RK and R5PE wherein one or more of the AI, RK and R5PE is derived from an AI, RK and R5PE of *B. thetaiotamicron.*

[0015] In other embodiments, the invention comprises a recombinant eukaryotic host cell expressing an arabinose isomerase (AI), a ribulokinase (RK) and a ribulose-5-phosphate epimerase (R5PE) wherein the AI comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 6; the RK comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 7 ; and, the R5PE comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 8 . In further embodiments, the recombinant host cell of the invention further comprises a heterologous polynucleotide encoding a xylose isomerase (XI).

[0016] In some embodiments, expression of one or more heterologous polynucleotides confers an ability to ferment arabinose to the recombinant host cell. In some embodiments, the recombinant eukaryotic host cell further comprises a heterologous polynucleotide encoding a xylose isomerase (XI). In further embodiments, the XI is derived from an XI of *B. thetaiotamicron.* In some embodiments, the XI comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 24 or SEQ ID NO: 26.

[0017] In some embodiments, the host cell is a yeast cell selected from the group consisting of *Saccharomyces cerevisiae, Schizzosaccharomyces pombe, Candida albicans, Pichia pastoris, Pichia stipitis, Yarrowia lipolytica, Hansenula polymorpha, Phaffia rhodozyma, Candida utilis, Arxula adeninivorans, Debaryomyces hansenii, Debaryomyces polymorphus, Schizosaccharomyces pombe and Schwanniomyces occidentalis.*

[0018] In some embodiments, the yeast cell comprises a heterologous sequence encoding a xylulokinase, ribulose 5-phosphate isomerase, ribulose 5-phophate epimerase, transketolase and transaldolase, and the yeast cell does not express an aldose reductase that is capable of catalyzing the conversion of xylose to xylitol.

[0019] In some embodiments, the invention relates to a method of producing a fermentation product comprising:

a) combining a recombinant eukaryotic host cell of the invention with a substrate;

b) allowing the host cell to ferment the substrate; and,

c) recovering one or more products of the fermentation,

wherein the substrate is selected from the group consisting of cellulosic substrate, biomass feedstock, and combinations thereof.

[0020] In some embodiments, the invention relates to a composition comprising a carbon source and the recombinant eukaryotic host cell.

[0021] In some embodiments, the invention relates to a media supernatant generated by incubating the host cell with a medium containing a carbon source.

BRIEF DESCRIPTION OF THE DRAWINGS/FIGURES

[0022] Figure 1 depicts a phylogenetic representation of the similarity of protein sequences with arabinose transporter activity.

[0023] Figure 2 depicts a phylogenetic representation of the similarity of protein sequences with possible arabinose transporter activity.

[0024] Figure 3 depicts the results from an arabinose uptake assay. *S. cerevisiae* strains were transformed with a plasmid expressing an arabinose transporter derived from the species indicated. The negative control strain was transformed with an empty vector plasmid. Arabinose concentration was measured before inoculation and 48 hours after inoculation.

[0025] Figure 4 depicts an assembly of the DNA construct used to compile the components of the arabinose utilization pathway and target the assembly to the rDNA sites for recombination into the yeast genome, (*i.e.* an arabinose transporter (AraT), an arabinose isomerase (AI), a ribulokinase (RK) and a ribulose 5-phosphate epimerase (R5PE), collectively the arabinose-utilization construct).

[0026] Figure 5 depicts an electrophoretic gel image depicting the individual amplicons that were cotransformed into yeast to assemble into the arabinose-utilization construct depicted in Figure 4, and integrate into the yeast genome.

[0027] Figure 6 depicts the results of a yeast transformation either with the DNA components of the arabinose-utilization construct added to the transformation or with no DNA added to the transformation. Transformants were selected on media containing arabinose as the only sugar. Colonies grew only upon successful transformation and integration of the arabinose-utilization construct into the genome of a progenitor cell.

[0028] Figure 7 depicts the levels of arabinose and ethanol over time in a fermentation experiment using an *S. cerevisiae* strain containing the arabinose-utilization construct (2874+ara) versus the control strain (2874). The control strain uses no arabinose and produces no ethanol whereas the strain containing the arabinose-utilization construct uses arabinose and is able to ferment arabinose to ethanol.

[0029] Figure 8 depicts the levels of arabinose, xylose and ethanol over time in a fermentation experiment using an *S. cerevisiae* strain containing the arabinose-utilization construct (2874+ara) versus a control strain (2874). The control strain uses no arabinose, but can produce ethanol from the xylose present in the media. The strain containing the arabinose-utilization construct (2874+ara) is able to ferment arabinose to ethanol which accounts for the increased ethanol produced as compared to the control strain.

[0030] Figure 9 depicts the levels of arabinose, glucose and ethanol over time in a fermentation experiment using an *S. cerevisiae* strain containing the arabinose-utilization construct (2874+ara) versus a control strain (2874). The control strain uses no arabinose, but can produce ethanol from the glucose present in the media. The strain containing the arabinose-utilization construct (2874+ara) is able to ferment arabinose to ethanol which accounts for the increased ethanol produced as compared to the control strain.

[0031] Figure 10 depicts levels of arabinose, xylose, glucose and ethanol over time in a fermentation experiment using an *S. cerevisiae* strain containing the arabinose-utilization construct (2874+ara) versus a control strain (2874). The control strain uses no arabinose, but can produce ethanol from the glucose and xylose present in the media. The strain containing the arabinose-utilization construct (2874+ara) is able to ferment arabinose to ethanol which accounts for the increased ethanol produced as compared to the control strain.

[0032] Figures 11-13 depict results from an arabinose utilization assay in an *S. cerevisiae* strain expressing the arabinose utilization genes of the invention and also expressing a xylose isomerase from *Pyromyces spp.* The yeast strains were pregrown on YPX, washed, and used to inoculate 150 ml sealed bottles with 25 ml medium, which were flushed with N_2 and incubated at 35°C at 250 RPM. Cultures were sampled for HPLC at 0, 24, and 48 hours. No growth was observed on YPA. 3.7 g l^{-1} arabinose was consumed for clone 1 in YPAXD. Over half of this (2.1 g l^{-1}) was consumed between 24 and 48 h, when xylose and glucose already had been depleted. Legend: 0 (parental strain), 1 (parental strain + ara genes, clone 1), 2 (parental strain + ara genes, clone 2).

DETAILED DESCRIPTION OF THE INVENTION

[0033] This specification discloses one or more embodiments that incorporate the features of this invention. The disclosed embodiment(s) merely exemplify the invention. The scope of the invention is not limited to the disclosed embodiment(s). The invention is defined by the claims appended hereto.

[0034] In the following description, for purposes of explanation, specific numbers, materials and configurations are set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one having ordinary skill in the art that the invention may be practiced without these specific details. In some instances, well-known features may be omitted or simplified so as not to obscure the present invention.

[0035] The embodiment(s) described, and references in the specification to "one embodiment", "an embodiment", "an example embodiment", etc., indicate that the embodiment(s) described can include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is understood that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0036] The description of "a" or "an" item herein may refer to a single item or multiple items. It is understood that wherever embodiments are described herein with the

language "comprising," otherwise analogous embodiments described in terms of "consisting of" and/or "consisting essentially of" are also provided.

Definitions

[0037] The term "heterologous" when used in reference to a polynucleotide, a gene, a polypeptide, or an enzyme refers to a polynucleotide, gene, polypeptide, or an enzyme not normally found in the host organism. "Heterologous" also includes a native coding region, or portion thereof, that is reintroduced into the source organism in a form that is different from the corresponding native gene, *e.g.*, not in its natural location in the organism's genome or with different regulatory sequences. The heterologous polynucleotide or gene may be introduced into the host organism by, *e.g.*, gene transfer. A heterologous gene may include a native coding region that is a portion of a chimeric gene including non-native regulatory regions that is reintroduced into the native host. Foreign genes can comprise native genes inserted into a non-native organism, or chimeric genes. A heterologous polynucleotide, gene, polypeptide, or an enzyme may be derived from any source, *e.g.*, eukaryotes, prokaryotes, viruses, or synthetic polynucleotide fragments.

[0038] The term "promoter" as used herein is a region of DNA that facilitates the transcription of a particular gene. Promoters are typically located near the genes they regulate, on the same strand and upstream (towards the 5' region of the sense strand). The terms "promoter" or "surrogate promoter" is intended to include a polynucleotide that can transcriptionally control a gene-of-interest that it does not transcriptionally control in nature. In certain embodiments, the transcriptional control of a surrogate promoter results in an increase in expression of the gene-of-interest. In certain embodiments, a surrogate promoter is placed 5' to the gene-of-interest. A surrogate promoter may be used to replace the natural promoter, or may be used in addition to the natural promoter. A surrogate promoter may be endogenous with regard to the host cell in which it is used, or it may be a heterologous polynucleotide sequence introduced into the host cell, *e.g.*, exogenous with regard to the host cell in which it is used.

[0039] As used herein, the term "terminator" or "transcription terminator" is a section of genetic sequence that marks the end of a gene or operon on genomic DNA for transcription. Promoters and terminators suitable for use in the present invention include, for example, those found in Table 1.

Table 1. Sequences of promoters and terminators used in exemplary embodiments.

Promoter	Sequence
ADH1 promoter	cgatttttttctaaaccgtggaatatttcggatatccttttgttgtttccgggtgtacaatat ggacttcctcttttctggcaaccaaacccatacatcgggattcctataataccttcgttg gtctccctaacatgtaggtggcggaggggagatatacaatagaacagataccagac aagacataatgggctaaacaagactacaccaattacactgcctcattgatggtggta cataacgaactaatactgtagccctagacttgatagccatcatcatatcgaagtttcac tacccttttttccatttgccatctattgaagtaataataggcgcatgcaacttcttttcttttttt tttcttttctctctcccccgttgttgtctcaccatatccgcaatgacaaaaaaatgatgga agacactaaaggaaaaaattaacgacaaagacagcaccaacagatgtcgttgttcc agagctgatgaggggtatctcgaagcacacgaaacttttttccttccttcattcacgca cactactctctaatgagcaacggtatacggccttccttccagttacttgaatttgaaata aaaaaaagtttgctgtcttgctatcaagtataaatagacctgcaattattaatcttttgttt cctcgtcattgttctcgttccctttcttccttgtttctttttctgcacaatatttcaagctatac caagcatacaatcaactatctcatataca
HXT7 promoter	ccagaaaggcaacgcaaaatttttttttccagggaataaacttttatgacccactacttc tcgtaggaacaatttcgggcccctgcgtgttcttctgaggttcatctttacatttgcttct gctggataattttcagaggcaacaaggaaaaattagatggcaaaaagtcgtctttca aggaaaaatccccaccatctttcgagatcccctgtaacttattggcaactgaaagaat gaaaaggaggaaaatacaaaatatactagaactgaaaaaaaaaaagtataaatagа gacgatatatgccaatacttcacaatgttcgaatctattcttcatttgcagctattgtaaa ataataaaacatcaagaacaaacaagctcaacttgtcttttctaagaacaaagaataa acacaaaaacaaaaagttttttaattttaatcaaaaa
TPI1 promoter	ctacttattcccttcgagattatatctaggaacccatcaggttggtggaagattacccg ttctaagacttttcagcttcctctattgatgttacacctggacacccctttttctggcatcc agtttttaatcttcagtggcatgtgagattctccgaaattaattaaagcaatcacacaatt ctctcggataccacctcggttgaaactgacaggtggtttgttacgcatgctaatgcaa aggagcctatataccttggctcggctgctgtaacagggaatataaagggcagcata atttaggagtttagtgaacttgcaacatttactattttcccttcttacgtaaatatttttctttt taattctaaatcaatcttttttcaatttttttgtttgtattctttttcttgcttaaatctataactacaa aaaacacatacataaactaaaa
ENO1 promoter	ctagtcttctaggcgggttatctactgatccgagcttccactaggatagcacccaaac acctgcatatttggacgacctttacttacaccaccaaaaaccactttcgcctctcccgc ccctgataacgtccactaattgagcgattacctgagcggtcctcttttgtttgcagcat gagacttgcatactgcaaatcgtaagtagcaacgtctcaaggtcaaaactgtatgga aaccttgtcacctcacttaattctagctagcctaccctgcaagtcaagaggtctccgtg attcctagccacctcaaggtatgcctctccccggaaactgtggccttttctggcacac atgatctccacgatttcaacatataaatagcttttgataatggcaatattaatcaaatttat tttacttctttcttgtaacatctctcttgtaatcccttattccttctagctatttttcataaaaa accaagcaactgcttatcaacacacaaacactaaatcaaa
PDC1 terminator	gcgatttaatctctaattattagttaaagttttataagcatttttatgtaacgaaaaataaat tggttcatattattactgcactgtcacttaccatggaaagaccagacaagaagttgcc gacagtctgttgaattggcctggttaggcttaagtctgggtccgcttctttacaaatttg gagaatttctcttaaacgatatgtatattcttttcgttggaaaagatgtcttccaaaaaaa aaaccgatgaattagtggaaccaaggaaaaaaaaagaggtatccttgattaaggaa

	cactgtttaaacagtgtggtttccaaaaccctgaaactgcattagtgtaatagaagact agacacctcgatacaaataatggttactcaattcaaaactgccagcgaattcgactct gcaattgctcaagacaagctagttgtcgtagatttctacgccacttggtgcggtccat gtaaaatgattgctccaatgattgaaa
PMA1 terminator	tcctgttgaagtagcatttaatcataatttttgtcacattttaatcaacttgatttttctggttt aatttttctaattttaatttaatttttttatcaatgggaactgatacactaaaaagaattag gagccaacaagaataagccgcttatttcctactagagtttgcttaaaatttcatctcga attgtcattctaatattttatccacacacacaccttaaaatttttagattaaatggcatcaa ctcttagcttcacacacacacacacaccgaagctggttgttttatttgatttgatataatt ggtttctctggatggtacttttttctttcttggttatttcctattttaaaatatgaaacgcaca caagtcataattattctaatagagcacaattcacaacacgcacatttcaacttttaatattt ttttagaaacactttatttagtctaattcttaatttttaatatatataatgcacacacactaat tt
FBA1 terminator	gttaattcaaattaattgatatagtttttttaatgagtattgaatctgtttagaaataatggaa tattatttttatttatttatttatattattggtcggctcttttcttctgaaggtcaatgacaaaat gatatgaaggaaataatgatttctaaaatttttacaacgtaagatatttttacaaaagcct agctcatcttttgtcatgcactattttactcacgcttgaaattaacggccagtccactgc ggagtcatttcaaagtcatcctaatcgatctatcgtttttgatagctcattttggagttcg cgattgtcttctgttattcacaactgttttaatttttatttcattctggaactcttcgagttctt gtaaagtctttcatagtagcttactttatcctccaacatatttaacttcatgtcaatttcgg ctcttaaattttccacatcatcaagttcaacatcatcttttaacttgaatttattctctagc
ENO1 terminator	tcgagagcttttgattaagccttctagtccaaaaaacacgtttttttgtcatttatttcatttt cttagaatagtttagtttattcattttatagtcacgaatgttttatgattctatatagggttgc aaacaagcatttttcattttatgttaaaacaatttcaggtttacctttattctgcttgtggtg acgcgtgtatccgcccgctcttttggtcacccatgtatttaattgcataaataattcttaa aagtggagctagtctatttctatttacatacctctcatttctcatttcctcctaatgtgtcaa tgatcatattcttaactggaccgatcttattcgtcagattcaaaccaaaagttcttaggg ctaccacaggaggaaaattagtgtgatataatttaaataatttatccgccattcctaata gaacgttgttcgacggatatctttctgcccaaaagggttctaagctcaatgaagagcc aatgtctaaacctc

[0040] As used herein, the term "operon" refers to a functioning unit of genomic material containing a cluster of genes under the control of a single regulatory signal or promoter. The genes are transcribed together into an mRNA strand and either translated together in the cytoplasm, or undergo *trans*-splicing to create monocistronic mRNAs that are translated separately, *i.e.* several strands of mRNA that each encode a single gene product. The result of this is that the genes contained in the operon are either expressed together or not at all. Originally operons were thought to exist solely in prokaryotes but since the discovery of the first operons in eukaryotes in the early 1990s, more evidence has arisen to suggest they are more common than previously assumed. Operons occur

primarily in prokaryotes but also in some eukaryotes, including *Drosophila melanogaster* and *C. elegans*.

[0041] The terms "gene(s)" or "polynucleotide" or "polynucleotide sequence(s)" are intended to include nucleic acid molecules, *e.g.*, polynucleotides which include an open reading frame encoding a polypeptide, and can further include non-coding regulatory sequences, and introns. In addition, the terms are intended to include one or more genes that map to a functional locus. In addition, the terms are intended to include a specific gene for a selected purpose. The gene may be endogenous to the host cell or may be recombinantly introduced into the host cell, *e.g.*, as a plasmid maintained episomally or a plasmid (or fragment thereof) that is stably integrated into the genome. In addition to the plasmid form, a gene may, for example, be in the form of linear DNA. As used herein, the terms may refer to, *inter alia*, genes encoding enzymes in the D-xylose pathway, such as xylose isomerase and xylulokinase and enzymes in the L-arabinose pathway, such as arabinose transporters (AraT), arabinose isomerase (AI), ribulokinase (RK), and ribulose 5-phosphate epimerase (R5PE). The term gene is also intended to cover all copies of a particular gene, *e.g.*, all of the DNA sequences in a cell encoding a particular gene product.

[0042] The term "expression" is intended to include the expression of a gene at least at the level of mRNA production.

[0043] The term "expression product" is intended to include the resultant product, *e.g.*, a polypeptide, of an expressed gene.

[0044] The term "cellulolytic activity" is intended to include the ability to hydrolyze glycosidic linkages in oligohexoses and polyhexoses. Cellulolytic activity may also include the ability to depolymerize or debranch cellulose and hemicellulose.

[0045] A "xylose metabolizing enzyme" can be any enzyme involved in xylose digestion, metabolism and/or hydrolysis, including a xylose isomerase, xylulokinase, xylose reductase, xylose dehydrogenase, xylitol dehydrogenase, xylonate dehydratase, a transketolase, and a transaldolase protein.

[0046] A "xylulokinase" (XK) as used herein, is meant for refer to an enzyme that catalyzes the chemical reaction: ATP + D-xylulose $\rightleftharpoons$ADP + D-xylulose 5-phosphate. Thus, the two substrates of this enzyme are ATP and D-xylulose, whereas its two products are ADP and D-xylulose 5-phosphate. This enzyme belongs to the family of

transferases, specifically those transferring phosphorus-containing groups (phosphotransferases) with an alcohol group as acceptor. The systematic name of this enzyme class is ATP:D-xylulose 5-phosphotransferase. Other names in common use include xylulokinase (phosphorylating), and D-xylulokinase. This enzyme participates in pentose and glucuronate interconversions. XK includes those enzymes that correspond to Enzyme Commission Number 2.7.1.17.

[0047] A "xylose isomerase" (XI) as used herein, is meant to refer to an enzyme that catalyzes the chemical reaction: D-xylose $\rightleftharpoons$D-xylulose. This enzyme belongs to the family of isomerases, specifically those intramolecular oxidoreductases interconverting aldoses and ketoses. The systematic name of this enzyme class is D-xylose aldose-ketose-isomerase. Other names in common use include D-xylose isomerase, D-xylose ketoisomerase, and D-xylose ketol-isomerase. This enzyme participates in pentose and glucuronate interconversions and fructose and mannose metabolism. The enzyme is used industrially to convert glucose to fructose in the manufacture of high-fructose corn syrup. It is sometimes referred to as "glucose isomerase". XI includes those enzymes that correspond to Enzyme Commission Number 5.3.1.5.

[0048] An "arabinose transporter" as used herein is meant to refer to an enzyme that is capable of efficiently transporting arabinose across a membrane. In general, arabinose transporters are transmembrane proteins that selectively transport pentoses, specifically arabinose, into the cell. Arabinose transporters according to the invention can be derived from a number of species. These include, *e.g.*, transporters derived from *Ambrosiozyma monospora*, *Candida arabinofermentans*, *Ambrosiozyma monospora*, *Kluveromyces marxianus*, *Pichia guillermondii* (LAT1), *Pichia guillermondii* (LAT2), *Pichia stipites*, *Ambrosiozyma monospora* (LAT2), *Debaryomyces hensenii*, *Apergillus flavus*, *Aspergillus terreus*, *Neosartorya fischeri*, *Aspergillus niger*, *Penicillium marneffei*, *Coccidioides posadasii*, *Gibberella zeae*, *Magnaporthe oryzae*, *Schizophyllum commune*, *Pichia stipites*, *Saccaharomyces* HXT2, *Aspergillus clavatus* (ACLA_032060), *Sclerotinia sclerotiorum* (SS1G_01302), *Arthroderma benhamiae* (ARB_03323), *Trichophyton equinum* (TEQG_03356), *Trichophyton tonsurans* (G_04876), *Coccidioides immitis* (CIMG_09387), *Coccidioides posadasii* (CPSG_03942), *Coccidioides posadasii* (CPC735_017640), *Botryotinia fuckeliana* (BC1G_08389), *Pyrenophora tritici-repentis* (PTRG_10527), *Ustilago maydis* (UM03895.1), *Clavispora lusitaniae* (CLUG_02297),

Pichia guillermondii (LAT1), *Pichia guillermondii* (LAT2), *Debaryomyces hansenii* (DEHA2E01166g), *Pichia stipites, candida albicans, Debaryomyces hansenii* (DEHA2B16082g), *Kluveromyces marxianus* (LAT1), *Kluyveromyces lactis* (KLLA-ORF10059), *Lachancea thermotolerans* (KLTH0H13728g), *Vanderwaltozyma polyspora* (Kpol_281p3), *Zygosaccharomyces rouxii* (ZYRO0E03916g), *Pichia pastoris* (0.1833), *Candida arabinofermentans* (0.1378), *Ambrosiozyma monospora* (LAT1), *Aspergillus clavatus* (ACLA_044740), *Neosartorya fischeri* (NFIA_094320), *Aspergillus flavus* (AFLA_116400), *Aspergillus terreus* (ATEG_08609), *Aspergillus niger* (ANI_1_1064034), *Telaromyces stipitatus* (TSTA_124770), *Penicillium chrysogenum* (Pc20g01790), *Penicillium chrysogenum* (Pc20g01790)#2, *Gibberella zeae* (FG10921.1), *Nectria hematococco*, and *Glomerella graminicola* (GLRG_10740) and *Arabidopsis thaliana* or any suitable source of the enzyme.

[0049] As used herein, "taken up" or "take up" is used to refer to the ability of a cell to transport a chemical moiety from the extracellular space into the intracellular space. For example, arabinose transporters of the invention allow the host cells of the invention to "take up" arabinose from the extracellular medium into the host cell.

[0050] An "arabinose isomerase (AI)" as used herein is meant to refer to an enzyme that is capable of catalyzing the chemical conversion of arabinose to ribulose (EC 5.3.1.3). Arabinose isomerase belongs to the oxidoreductase family of enzymes capable of interconverting aldoses and ketoses. Arabinose isomerases of the invention include those derived from various species including both prokaryotic and eukaryotic species. Arabinose isomerases may be derived from *B. subtilis, M. smegmatis, B. licheniformis, L. plantarum, Arthrobacter aurescens, Clavibacter michiganensis, Gramella forsetii, B. thetaiotamicron* or any other suitable source of the enzyme.

[0051] A "ribulokinase" (RK) as used herein is meant to refer to an enzyme that is capable of catalyzing the chemical reaction that phosphorylates ribulose to yield ribulose-5-phosphate (EC 2.7.1.16). Ribulokinases of the invention include those derived from various species including both prokaryotic and eukaryotic species. Ribulokinases may be derived from *E. coli, L. plantarum, A. aurescens, C. michiganensis, G. forsetii, B. thetaiotamicron* or any other suitable source of the enzyme.

[0052] A "ribulose 5-phosphate epimerase" (R5PE) as used herein is meant to refer to an enzyme that is capable of catalyzing the interconversion of ribulose-5-phosphate and

xylulose-5-phosphate (EC 5.1.3.4). Ribulose 5-phosphate epimerases of the invention include those derived from various species including both prokaryotic and eukaryotic species. Ribulose 5-phosphate epimerases of the present invention include those derived from various species including both prokaryotic and eukaryotic species. Ribulose 5-phosphate epimerases may be derived from *E. coli, L. plantarum, Arthrobacter aurescens, C. michiganensis, G. forsetii, B. thetaiotamicron* or any other suitable source of the enzyme.

[0053] Certain embodiments of the present invention provide for the "inactivation" or "deletion" of certain genes or particular polynucleotide sequences within microorganisms, which "inactivation" or "deletion" of genes or particular polynucleotide sequences may be understood to encompass "genetic modification(s)" or "transformation(s)" such that the resulting strains of said microorganisms may be understood to be "genetically modified" or "transformed." In certain embodiments, strains of microorganisms may be of bacterial, fungal, or yeast origin.

[0054] Certain embodiments of the present invention provide for the "insertion," (*e.g.*, the addition, integration, incorporation, or introduction) of certain genes or particular polynucleotide sequences within microorganisms, which insertion of genes or particular polynucleotide sequences may be understood to encompass "genetic modification(s)" or "transformation(s)" such that the resulting strains of microorganisms may be understood to be "genetically modified" or "transformed." In certain embodiments, strains of microorganisms may be of bacterial, fungal, or yeast origin.

[0055] The term "CBP organism" is intended to include certain microorganisms of the invention, *e.g.*, microorganisms that have properties suitable for CBP.

[0056] The terms "fermenting" and "fermentation" are intended to include the enzymatic process (*e.g.*, cellular or acellular, *e.g.*, a lysate or purified polypeptide mixture) by which ethanol, or another fermentation product is produced from a carbohydrate, in particular, as a result of fermentation.

[0057] In one aspect of the invention, the genes or particular polynucleotide sequences are inserted to confer upon the cell the activity encoded by the polynucleotide, such as the expression of an enzyme. In certain embodiments, genes encoding enzymes in the metabolic production of ethanol, *e.g.*, enzymes that metabolize pentose and/or hexose sugars, may be added to a mesophilic or thermophilic organism. In certain embodiments

of the invention, the enzyme may confer the ability to metabolize a pentose sugar and be involved, for example, in a xylose-utilization pathway and/or an arabinose-utilization pathway. In certain embodiments of the invention, genes encoding enzymes in the conversion of acetate to a non-charged solvent, including but not limited to, acetone, isopropanol, ethyl acetate, or ethanol, may be added to an organism.

[0058] In one aspect of the invention, genes or particular polynucleotide sequences are partially, substantially, or completely deleted, silenced, inactivated, or down-regulated in order to inactivate the activity for which they encode, such as the expression of an enzyme. Deletions provide maximum stability because there is no opportunity for a reverse mutation to restore function. Alternatively, genes can be partially, substantially, or completely deleted, silenced, inactivated, or down-regulated by insertion of nucleic acid sequences that disrupt the function and/or expression of the gene (*e.g.*, P1 transduction or other methods known in the art). The terms "eliminate," "elimination," and "knockout" are used interchangeably with the terms "deletion," "partial deletion," "substantial deletion," or "complete deletion." In certain embodiments, strains of microorganisms of interest may be engineered by site directed homologous recombination to knockout specific genes. In still other embodiments, RNAi or antisense DNA (asDNA) may be used to partially, substantially, or completely silence, inactivate, or down-regulate a particular gene of interest.

[0059] In certain embodiments, the genes targeted for deletion or inactivation as described herein may be endogenous to the native strain of the microorganism, and may thus be understood to be referred to as "native gene(s)" or "endogenous gene(s)." An organism is in "a native state" if it has not been genetically engineered or otherwise manipulated by the hand of man in a manner that intentionally alters the genetic and/or phenotypic constitution of the organism. For example, wild-type organisms may be considered to be in a native state. In other embodiments, the gene(s) targeted for deletion or inactivation may be non-native to the organism.

[0060] Similarly, the enzymes of the invention as described herein can be endogenous to the native strain of the microorganism, and can thus be understood to be referred to as "native" or "endogenous."

Biomass

[0061] Biomass can include any type of biomass known in the art or described herein. The terms "lignocellulosic material," "lignocellulosic substrate," and "cellulosic biomass" mean any type of biomass comprising cellulose, hemicellulose, lignin, or combinations thereof, such as but not limited to woody biomass, forage grasses, herbaceous energy crops, non-woody-plant biomass, agricultural wastes and/or agricultural residues, forestry residues and/or forestry wastes, paper-production sludge and/or waste paper sludge, waste-water-treatment sludge, municipal solid waste, corn fiber from wet and dry mill corn ethanol plants, and sugar-processing residues. The terms "hemicellulosics," "hemicellulosic portions," and "hemicellulosic fractions" mean the non-lignin, non-cellulose elements of lignocellulosic material, such as but not limited to hemicellulose (*i.e.*, comprising xyloglucan, xylan, glucuronoxylan, arabinoxylan, mannan, glucomannan, and galactoglucomannan, inter alia), pectins (*e.g.*, homogalacturonans, rhamnogalacturonan I and II, and xylogalacturonan), and proteoglycans (*e.g.*, arabinogalactan-protein, extensin, and proline-rich proteins).

[0062] In a non-limiting example, the lignocellulosic material can include, but is not limited to, woody biomass, such as recycled wood pulp fiber, sawdust, hardwood, softwood, and combinations thereof; grasses, such as switch grass, cord grass, rye grass, reed canary grass, miscanthus, or a combination thereof; sugar-processing residues, such as but not limited to sugar cane bagasse; agricultural wastes, such as but not limited to rice straw, rice hulls, barley straw, corn cobs, cereal straw, wheat straw, canola straw, oat straw, oat hulls, and corn fiber; stover, such as but not limited to soybean stover, corn stover; succulents, such as but not limited to, Agave; and forestry wastes, such as but not limited to, recycled wood pulp fiber, sawdust, hardwood (*e.g.*, poplar, oak, maple, birch, willow), softwood, or any combination thereof. Lignocellulosic material may comprise one species of fiber; alternatively, lignocellulosic material may comprise a mixture of fibers that originate from different lignocellulosic materials. Other lignocellulosic materials are agricultural wastes, such as cereal straws, including wheat straw, barley straw, canola straw and oat straw; corn fiber; stovers, such as corn stover and soybean stover; grasses, such as switch grass, reed canary grass, cord grass, and miscanthus; or combinations thereof.

[0063] Paper sludge is also a viable feedstock for lactate or acetate production. Paper sludge is solid residue arising from pulping and paper-making, and is typically removed from process wastewater in a primary clarifier.

Consolidated Bioprocessing

[0064] Consolidated bioprocessing (CBP) is a processing strategy for cellulosic biomass that involves consolidating into a single process step four biologically-mediated events: enzyme production, hydrolysis, hexose fermentation, and pentose fermentation. Implementing this strategy requires development of microorganisms that both utilize cellulose, hemicellulosics, and other biomass components, such as hexose and pentose sugars, while also producing a product of interest at sufficiently high yield and concentrations. The feasibility of CBP is supported by kinetic and bioenergetic analysis. *See e.g.* van Walsum and Lynd (1998) *Biotech. Bioeng. 58*:316.

Pentose metabolism

[0065] The term "pentose" includes the five-carbon monosaccharides xylose and arabinose that can be metabolized into useful products by the organisms of the present invention. There are two main pathways of xylose metabolism, each pathway is unique in the characteristic enzymes it utilizes. One pathway is called the "Xylose Reductase-Xylitol Dehydrogenase" or XR-XDH pathway. Xylose reductase (XR) and xylitol dehydrogenase (XDH) are the two main enzymes used in this method of xylose degradation. XR, encoded by the *GRE3* gene in *S. cerevisiae*, is responsible for the reduction of xylose to xylitol and is aided by cofactors NADH or NADPH. Xylitol is then oxidized to xylulose by XDH, which is expressed through the XYL2 gene, and accomplished exclusively with the cofactor NAD^+. Because of the varying cofactors needed in this pathway and the degree to which they are available for usage, an imbalance can result in an overproduction of xylitol byproduct and an inefficient production of desirable ethanol. In some embodiments of the invention the *GRE3* gene is deleted to remove the XR-XDH pathway for xylose metabolism from operating in cells of the invention.

[0066] The other pathway for xylose metabolism is called the "Xylose Isomerase" (XI) pathway. XI is responsible for direct conversion of xylose into xylulose, and does not proceed via a xylitol intermediate. Both pathways create xylulose, although the enzymes

utilized are different. After production of xylulose both the XR-XDH and XI pathways proceed through enzyme xylulokinase (XK), encoded on gene XKS1, to further modify xylulose into xylulose-5-P where it then enters the pentose phosphate pathway for further catabolism.

[0067] The xylose isomerase pathway is considered advantageous because it does not produce a cofactor imbalance; however the XR-XDH is endogenous to yeast, whereas the XI pathway is absent. In some embodiments, the cells of the invention comprise an exogenous XI. XI includes those enzymes that correspond to Enzyme Commission Number 5.3.1.5. Suitable xylose isomerases of the present invention include xylose isomerases derived from *Pyromyces sp.,* and *B. thetaiotamicron,* although any xylose isomerase that functions when expressed in host cells of the invention can be used.

[0068] Arabinose is not efficiently taken up by wild-type *S. cerevisiae*, but small amounts of arabinose present in the cell are naturally processed by an $NAD(P)^+$-dependant dehydrogenase to yield arabinono-1,4-lactone. However, arabinono-1,4-lactone is not a useful intermediate for the production of fermentation products and this pathway does not efficiently use arabinose. Exogenously expressed arabinose transporters of the present invention can confer upon host cells such as *S. cerevisiae* the ability to take up arabinose from the external media. In addition to the sequences listed in SEQ ID NOs: 9-22, Table 2 lists the amino acid sequences of additional arabinose transporters of the invention.

Table 2. Amino acid sequences of additional arabinose transporters of the invention and the species from which they are derived.

Zygosaccharomyces rouxii	mkidkkqigcalmgkrinyrvtiydkfpkiynifvigftscisglmfgfdvssmssmig tdgykeyfgtpgpteqggitacmpagsfvasliapyfsdnfgrrvslhlcaifwmigavl qcasqdlamlcvgrvvsglgigfgssvapvycseiappkirgaigglfqfsvtlgimilffi gygahfingagsfrltwgielvpgaclliavfflpesprwlalhdyweeaedivirvaakg nreneqvmiqleeireqveidkeaeafqlkdlfrpktrvktmvgmmaqmwqqmcg mnvmmyyivyiftmagfkggavlvsgsiqyvlnvvmtipalflmdkcgrrpvlligg llmcawlfavggllatysdpyphgfegdetvriaipqsnkpaangviacsylfvcsyapt wgvciwiycaeifnnterakgsglctavnwifnfalalfvpsafknltwktyimfgvfcv altintfllfpetkgktleeidqmweahipawkthswvptipsaskfdqemhktdlehve dtgdsdrispkddsekgsvtgleevaksnpnstslse
Vanderwaltozyma polyspora	mgsfkdtilmknikyegklyerfpkiyniyvigfvscisglmfgfdissmssmigtday kqyfgspdatkqggitssmaagsfvgsllsplfsdvfgrrvslhicstfwligatlqcasqdl amlvvgrlvsgigigfgsavapvycsevappkirgaiaglfqlsvtlgililyyvgygahfi tsassfrltwgiqlvpgfvllvatfflpesprwlankgfwekatynicrinntdpdniseev aiqleemntqvmddkeadsftyanlfrkktikktivgmsaqmwqqlsginvmmyyi vyifqmagysgnavlvsgsinyilnvamtipalfvidklgrrpilivggilmfvwlfava gllsvysvpvpggvggnetvnimipdnhkhaakgviaccylfvctfaptwgigiwivc

	seifnnserakgsslsaavnwifnfalglfvpsafqnitwktylmfgifsvaltihtflmfpe tkgktleeidqmweanipawrsaswkptlpshlhddfknlhtgesssnfveddgkae mekpvvdhiestdksl
Debaryomyces hansenii	mnsifnysgfvmkftipekyllenkvkgkklthcvvalsalaifffgydqgmmagvnt spdyvekmkygyfnengdvtvtnstrqggivaiyyfgtlvgcvfgglfsdrhgrikaial galiaifgaalqcaaqqmswmcgarfvngigtgilnavvpvyssetaehtsrgafiaieft lnifgvcvaywleyglsyidsgfsafqwrfpiafqiipllvllgivwffpesprwlvknge edhakrillnmrgvergnqefaeivgamrfeqesalsssywrmflgyfpdkdskksak aktlhiarrvqiviwmqifqewvgiagvtvyqpeifkqagfgtrksawlsgvnnifycls tlinfftvdrfgrrftlfwgaigqgismflaggfsklqqknpknssygaaaasfvfiytsifg atwlavpwlypteifplkvraqgnafgvvgwsigngwltllcpimfskigektlyifgac nfislalvylfcpetanrtledidylfandswlaskseadfkrikieqvdkqvgrekqiidid ssekenfstehfe
Aspergillus niger	myrisniyvlagfgtiggalfgfdvssmsawigtdqyleyfnhpdsdlqggitasmsag sfagalaagfisdrigrryslmlacciwvigaaiqcsaqnvahlvagrvisglsvgitssqv cvylaelaparirgrivgiqqwaiewgmlimylisygcgqglagaasfrvswgvqgipa lillaalpffpesprwlaskerweealdtlallhakgdrndpvvqveyeevqeaariaqea kdisffslfgpkiwkrtlcgvsaqvwqqllggnvamyyvvyifnmagmsgnttlyssa iqyviflvttgtilpfvdrigrrlllltgsvlcmachfaiaglmasrghhvdsvdgnanlkw sitgppgkgviacsyifvavygftwapvawiyasevfplkyrakgvglsaagnwifnfa layfvapaftniqwktyiifgvfctvmtfhvfffypetarrsledidlmfetdmkpwkth qihdrfgeeverhkhkdmadqekgvvsthdema
Penicillium chrysogenum	mytitniyvlaafgtiggalfgfdvssmsawigvdtytdyfdspdsnlqggitasmsags fagsiaagwladilgrryalmiaslvwivgavvqcsaqnvthlvagrvvsglavgvtssq tcvylselaparirgrivgiqqwaiewgilimyliaygcvvgvsgpaafricwgvqavp glilfialfffpesprwlasqerweealdtlaiihangdrhdpvvqvefeevqeavrvahe srdvsfmalfgprvwkrtmcgmsvqmwqqllggnvamyyvvyifemagmtgntt lwssaiqyviflvttgcmlpfidrvgrrnllligsvtcmvvhyiiaavmaskgkpvpdvn gnanltweikgsagmtviafsyiftgiygltwaptawiyaaevfplkfrakgvglsaatn wifnfalayfvapafhniqwktyiifgvfctvmtfhvffmypetvgrsleeidlvfetdvk pwrthkigdifgeeierrkelgaktetggatheevv
Pichia guilermondii	mgyedklvapalkfrnfidrtpntynvyviasiscisgamfgfdissmsvfvgqtpylnf fhspksdlqgfitaamslgsffgsllssfvsepfgrrasllicgflwcvgaaiqcssqnvaql iigriisgfgvgfgssvapvygsemaprkirgtiggffqfsvtlgifimfligygcskidav gsfripwgvqivpglflllgcffipesprwlakqgyweeaeiivaniqakgnredpdvli eiseikeqllldehakaftyadlfskkylprtitaisaqiwqqltgmnvmmyyivyifqm agyegdtnlipsliqyiintvvtipslylldrvgrrkmllfgaaammawqfgvagilatys epydlndtvkitipdkhksaakgviaccylfvasfastwgvgiwvycsevwgdsqsrq rgaavataanwifnfaigmftpssfknitwktyciyatfcgcmfihvffffpetkgkrleei aqiweekvpawktskwqphvpllsdhelaekmstkhdenmlqsqsseekptv
A. flavus	mcdqipkwnvvhrlekrklliginsvaalsilffgydqgmmagvnnskdyidlmgfg ytemkdgyltpvvtdsllqggivsvyylgtlfgallggwigdrigriktiaagalwailgaa lqcsaqnhnwmicsrfingigtgilnaivpvwatetaehtsrgqfiaieftlnifgvvlayw lefglsfidggrspfrwrfpiafqiiflvllfvvvwffpesprwlvkvgreqearyilgrlrgs sdedavraeaefrdiqnvaemeksmnhstsylamlfgyktgklhlgrrvqlviwlqim qewvgiagvtvyaptifsiagfdsmksqwisglnnvfymfatlvcvftldrigrrwtlyw gsiaqgiamflaggfsrlaidaradgnisransfgaaaasmvfiftsvfgatwltvvpwiy paeiyplavrakgnawgvvgwsigngwltllcpvmfeaigektlyvfaasnvitgmv

	walypesnqrtledmdllfaaetpwvwdaertfarlkaenpgyietanrknsavdpem gkptdaheehassas
C. lusitnaea	mgyeeklvapalklrrfldrtpntynvyfiasiscisgmmfgfdissmsvfvsdkpylny fdhpssvmqgfitaamslgsffgslsssfvsepfgrrasllicgflwcvgaaiqcsaqnra qliigriisgwgvgfgssvspvygselsprkirgfvggmfqfsvtfgilimfliaygmshv hgkasfrvswgvqivpglvlliglffipesprwlakqgywdeaefivakiqakgnredp evqielseikeqllleehaknftyadlfspkyrvrtvtavfaqiwqqltgmnvmmyyiv yifemagyegntnlipsliqyiinsavtvpslylldkvgrrtlllfgaagmmafqfavagll atysipheykgndtvritipkknkpaargviaccylfvvcfastwgvgiwvycsevwg dnrsrqrgaslstsanwifnfaiamftpssfknitwktyiiyavfcccmfvhvffcfpetrg krleeiaqiwdekvpawktrnwqphvpllsdaqleeklnvnhaenagedkavqshss sdgqv
C. albicans (SC5314)	mksplelalggtalkistfldklpkiynvyfiasistiagmmfgfdissmsafigtetymdf fnspgsdiqgfitssmalgsffgsiassfisepfgrrlsliicaffwmvgaaiqssvqnraql iigriisgvgvgfgssvatiygaelaprkirgfiggmfqffvtlgilimfylsfglghikgva sfriawglqivpglmlfigcffipesprwlakqnrweqaeyivsriqakgnredpdvliei seikdqllieeaaksvsyatlfrkkyllrtftaifaqiwqqltgmnvmmyyivyifqmag ysgnanlvassiqyvintgvtipalffvdrigrrpvlitgavlmmtfqfglagilgqysvp wtdsgndsvniripednksaskgaiaccylfvasfastwgptiwiycseiwgdnrvaqr gnslataanwilnfaigmytpagfksiswrtyiiygvmcftmaihvyfgfpetkgkrlee igqmweehvpawksrswqphvpiasdaelarkmdvehkegglmnedtnseakaes v

[0069] Conversely, some organisms contain an arabinose pathway that converts arabinose into xylulose, which can then enter the pentose phosphate pathway for further catabolism, leading to fermentation. This arabinose pathway comprises an arabinose isomerase (AI) (encoded by AraA), a ribulokinase (RK) (encoded by AraB) and an epimerase (R5PE) (encoded by AraD). The AraA isomerase converts arabinose to ribulose; the AraB ribulokinase phosphorylates the ribulose to yield ribulose-phosphate; and the AraD ribulose 5-phosphate epimerase converts ribulose-phosphate to xylulose-phosphate which enters the pentose phosphate pathway and can ultimately be converted to glucose, or glycolytic intermediates to yield fermentation products.

Host Cells

[0070] Host cells useful in the present invention include prokaryotic or eukaryotic cells; for example, microorganisms selected from bacterial and yeast cells. Among host cells suitable for the present invention are microorganisms, for example, of the genera *Aeromonas*, *Aspergillus*, *Clostridium, Bacillus*, *Escherichia*, *Kluyveromyces*, *Pichia*, *Rhodococcus*, *Saccharomyces* and *Streptomyces*.

[0071] In some embodiments, the host cells are microorganisms. In one embodiment the microorganism is a yeast. According to the present invention the yeast host cell can be, for example, from the genera *Saccharomyces*, *Kluyveromyces*, *Candida*, *Pichia*, *Schizosaccharomyces*, *Hansenula*, *Kloeckera*, *Schwanniomyces*, and *Yarrowia*. In one particular embodiment, the yeast is *Saccharomyces cerevisiae*. In another embodiment, the yeast is a thermotolerant *Saccharomyces cerevisiae*. The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein.

[0072] In some embodiments, the host cell is an oleaginous cell. The oleaginous host cell can be an oleaginous yeast cell. For example, the oleaginous yeast host cell can be from the genera *Blakeslea, Candida, Cryptococcus, Cunninghamella, Lipomyces, Mortierella, Mucor, Phycomyces, Pythium, Rhodosporidum, Rhodotorula, Trichosporon* or *Yarrowia*.

[0073] In some embodiments, the host cell is a thermotolerant host cell. Thermotolerant host cells can be particularly useful in simultaneous saccharification and fermentation processes by allowing externally produced cellulases and ethanol-producing host cells to perform optimally in similar temperature ranges.

[0074] In some particular embodiments, the host cell is a *Kluyveromyces* host cell. The host cells can contain antibiotic markers or can contain no antibiotic markers. In another embodiment, the host cells are bacteria selected from the genus *Clostridium, Acinetobacter*, *Thermoanaerobacterium,* and other bacteria having characteristics resembling those of *Clostridium* species.

[0075] Several microorganisms that are reported in the literature to be cellulolytic or have cellulolytic activity have been characterized by a variety of means, including their ability to grow on microcrystalline cellulose as well as a variety of other sugars. Additionally, such organisms may be characterized by other means, including but not limited to, their ability to depolymerize and debranch cellulose and hemicellulose.

[0076] Certain microorganisms, including, for example, *S. cerevisiae*, cannot metabolize pentose sugars, such as xylose or arabinose, but are able to metabolize hexose sugars. Both xylose and arabinose are abundant sugars in biomass with xylose accounting for approximately 16-20% in soft and hard woods and L-arabinose accounting for approximately 25% in corn fiber. Accordingly, one embodiment of the invention is a genetically-modified cellulolytic microorganism, with the ability to metabolize pentose sugars, such as xylose and arabinose, thereby to enhance its use as a biocatalyst for

fermentation in the biomass-to-acetic acid or lactic acid or ethanol industries. Therefore, in some embodiments, the host cell is a *S. cerevisiae* strain.

[0077] In some embodiments, the thermotolerant host cell can grow at temperatures above about 30° C, about 31° C, about 32° C, about 33° C, about 34° C, about 35° C, about 36° C, about 37° C, about 38° C, about 39° C, about 40° C, about 41° C or about 42° C. In some embodiments of the present invention the thermotolerant host cell can produce ethanol from xylose at temperatures above about 30° C, about 31° C, about 32° C, about 33° C, about 34° C, about 35° C, about 36° C, about 37° C, about 38° C, about 39° C, about 40° C, about 41° C, about 42° C, or about 43 °C, or about 44 °C, or about 45 °C, or about 50° C.

[0078] In some embodiments of the present invention, the thermotolerant host cell can grow at temperatures from about 30° C to 60° C, about 30° C to 55° C, about 30° C to 50° C, about 40° C to 60° C, about 40° C to 55° C or about 40° C to 50° C. In some embodiments of the present invention, the thermotolerant host cell can produce ethanol from xylose at temperatures from about 30° C to 60° C, about 30° C to 55° C, about 30° C to 50° C, about 40° C to 60° C, about 40° C to 55° C or about 40° C to 50° C.

[0079] In some embodiments, the host cell is a host cell that exhibits high ethanol tolerance. High ethanol tolerant strains are able to produce an ethanol titer higher than a non-high ethanol tolerant strain containing the same modifications according to the present invention, under identical growth conditions. Such industrially hearty strains are known in the art of industrial ethanol production. In some embodiments, the high ethanol tolerant strains can produce ethanol titers of up to 4%, up to 5%, up to 6%, up to 7%, up to 8%, up to 9%, up to 10%, up to 11%, up to 12%, up to 13%, up to 14%, up to 15%, up to 16%, and up to 17% ethanol in the fermentation media.

[0080] The present invention provides cellulolytic microorganisms expressing enzymes that allow the microorganisms to ferment xylose and/or arabinose. When genes encoding enzymes involved in the metabolic pathway of lactate or acetate, including, for example, xylose and/or arabinose, are introduced into a microorganism that lacks one or more of these genes, for example, *S. cerevisiae*, one may select transformed strains for growth on xylose or growth on arabinose. *S. cerevisiae* may lack one or more known genes or enzymes in the arabinose to ethanol pathway and/or the arabinose utilization pathway.

[0081] In one embodiment, host cells are genetically engineered (transduced or transformed or transfected) with the polynucleotides encoding arabinose-utilizing enzymes of this invention. In some embodiments, the polynucleotides encoding arabinose-utilizing enzymes can be introduced to the host cell on a vector, which may be, for example, a cloning vector or an expression vector comprising a sequence encoding a heterologous xylose metabolizing enzyme. The host cells can comprise polynucleotides of the invention as integrated copies or plasmid copies.

[0082] In some embodiments, the arabinose-utilizing enzymes may be introduced as a series of overlapping segments and cotransformed into the host cells so as to accomplish multiple homologous recombination events, which render a complete construct within the host cell. In further embodiments, the assembled construct contains homology to a target segment of the host cell genome and the entire, assembled construct integrates into the host cell genome.

[0083] In certain aspects, the present invention relates to host cells containing the polynucleotide constructs described below by way of specific examples. The host cells of the present invention can express one or more heterologous polypeptides expressing xylose metabolizing enzymes in addition to the arabinose-utilizing enzymes. In some embodiments, the host cell comprises a combination of polynucleotides that encode heterologous xylose metabolizing enzymes or fragments, variants or derivatives thereof. The host cell can, for example, comprise multiple copies of the same nucleic acid sequence, for example, to increase expression levels, or the host cell can comprise a combination of unique polynucleotides. In other embodiments, the host cell comprises a single polynucleotide that encodes a heterologous xylose metabolizing enzyme or a fragment, variant or derivative thereof. For example, in some embodiments the host cells of the invention have an up-regulated pentose phosphate pathway. In some embodiments, the pentose phosphate pathway genes that are up-regulated are transketolase, transaldolase, ribulose-5-phosphate-3-epimerase, and/or ribulose-5-phosphate isomerase. However, in alternative embodiments, the endogenous pentose phosphate pathway of the host cell is not genetically manipulated.

[0084] In some embodiments, the microorganisms of the invention contain enzymes involved in cellulose digestion, metabolism and/or hydrolysis. A "cellulolytic enzyme" can be any enzyme involved in cellulose digestion, metabolism, and/or hydrolysis. The

term "cellulase" refers to a class of enzymes produced chiefly by fungi, bacteria, and protozoans that catalyze cellulolysis (*i.e.* the hydrolysis) of cellulose. However, there are also cellulases produced by other types of organisms such as plants and animals. Several different kinds of cellulases are known, which differ structurally and mechanistically. There are general types of cellulases based on the type of reaction catalyzed: endocellulase breaks internal bonds to disrupt the crystalline structure of cellulose and expose individual cellulose polysaccharide chains; exocellulase cleaves 2-4 units from the ends of the exposed chains produced by endocellulase, resulting in the tetrasaccharides or disaccharide such as cellobiose. There are two main types of exocellulases (or cellobiohydrolases, abbreviate CBH) - one type working processively from the reducing end, and one type working processively from the non-reducing end of cellulose; cellobiose or beta-glucosidase hydrolyses the exocellulase product into individual monosaccharides; oxidative cellulases that depolymerize cellulose by radical reactions, as for instance cellobiose dehydrogenase (acceptor); cellulose phosphorylases that depolymerize cellulose using phosphates instead of water. In the most familiar case of cellulase activity, the enzyme complex breaks down cellulose to beta-glucose. A "cellulase" can be any enzyme involved in cellulose digestion, metabolism and/or hydrolysis, including, for example, an endoglucanase, glucosidase, cellobiohydrolase, xylanase, glucanase, xylosidase, xylan esterase, arabinofuranosidase, galactosidase, cellobiose phosphorylase, cellodextrin phosphorylase, mannanase, mannosidase, xyloglucanase, endoxylanase, glucuronidase, acetylxylanesterase, arabinofuranohydrolase, swollenin, glucuronyl esterase, expansin, pectinase, and feruoyl esterase protein.

[0085] This full suite of cellulase enzymes that can be used in host cells and methods of the present invention contains activities beyond those identified previously for expression in yeast: CBH1, CBH2, EG, and BGL (as disclosed *e.g.* in PCT Application No. PCT/US2009/065571). In some embodiments, the present invention relates to a yeast cell that expresses one or more gene products of the genes: *Aspergillus fumigatus* Endoglucanase (Accession No. XP_747897); *Neosartorya fischeri* Endoglucanase (Accession No. XP_001257357); *Aspergillus clavatus* Endoglucanase (Accession No. XP_001270378); *Aspergillus terreus* Endoglucanase (Accession No. XP_001217291); *Penicillium marneffei* Endoglucanase (Accession No. XP_002152969); *Chaetomium*

globosum Endoglucanase (Accession No. XP_001229968); *Neurospora crassa* Endoglucanase (Accession No. XP_956431); *Aspergillus oryzae* Endoglucanase (Accession No. BAA22589); *Thielavia heterothallica* Endoglucanase (Accession No. AAE25067); *Fusarium oxysporum* Endoglucanase (Accession No. AAG09047); *Humicola insolens* Endoglucanase (Accession No. 1DYM_A); *Pyrenophora tritici-repentis* Endoglucanase (Accession No. XP_001935476); *Magnaporthe grisea Endoglucanase* (Accession No. XP_370166); *Fusarium graminearum* Endoglucanase (Accession No. XP_388429); *Chrysosporium lucknowense* Endoglucanase; *Polyporus arcularius* Endoglucanase (Accession No. BAF75943.1); *Aspergillus kawachii* Endoglucanase (Accession No. BAB62317.1); *Heterodera schachtii* Endoglucanase (Accession No. CAC12958.1); *Orpinomyces sp.* Endoglucanase (Accession No. AAD04193.1); *Irpex lacteus* Endoglucanase (Accession No. BAD67544.1); *Chaetomium globosum* Endoglucanase (Accession No. XP_001220409.1); *Aspergillus niger* Endoglucanase (Accession No. XP_001397982.1); *Penicillium decumbens* Endoglucanase (Accession No. ABY28340.1); *Phanerochaete chrysosporium* Endoglucanase (Accession No. AAU12276); *Stachybotrys echinata* Endoglucanase (Accession No. AAM77710); *Neosartorya fischeri* Endoglucanase (Accession No. XP_001261563); *Chaetomium brasiliense* Endoglucanase (Accession No. AAM77701); *Chaetomium globosum* Endoglucanase (Accession No. EAQ86340); *Aspergillus fumigatus* Endoglucanase (Accession No. CAF31975); *Humicola insolens* Endoglucanase (Accession No. CAG27577); *Neosartorya fischeri* Endoglucanase (Accession No. XP_001267517); *Thielavia terrestris* Endoglucanase (Accession No. ACE10231); *Chrysosporium lucknowense* Endoglucanase (Accession No. ACH15008); *Chaetomium globosum* Endoglucanase (Accession No. XP_001226436); *Acremonium thermophilum* Endoglucanase (Accession No. ACE10216); *Humicola insolens* Endoglucanase (Accession No. CAB42307); *Thielavia terrestris* Endoglucanase (Accession No. CAH03187); *Chrysosporium lucknowense* Endoglucanase (Accession No. AAQ38151); *Magnaporthe grisea* Endoglucanase (Accession No. EDJ97375); *Chaetomium globosum* Endoglucanase (Accession No. EAQ84577); *Humicola insolens* Endoglucanase 1DYS_B; *Neurospera crassa* Endoglucanase (Accession No. XP_957415); *Trichoderma reesei* Xyloglucanase (Accession No. AAP57752); *Aspergillus niger* Xyloglucanase (Accession No. AAK77227); *Aspergillus aculeatus* Xyloglucanase (Accession No. BAA29031);

Neosartorya fischeri Xyloglucanase (Accession No. XP_001261776); *Chaetomium thermophilum* Endoxylanase (Accession No. CAD48749); *Trichoderma reesei* Endoxylanase (Accession No. ABK59833); *Chrysosporium lucknowense* Endoxylanase (Accession No. AAQ38147); *Aureobasidium pullulans* Endoxylanase (Accession No. BAE71410); *Aspergillus nidulans* beta-xylosidase (Accession No. CAA73902; *Cochliobolus carbonum* beta-xylosidase (Accession No. AAC67554); *Penicillium herquei* beta-xylosidase (Accession No. BAC75546); *Pyrenophora tritici-repentis* beta-xylosidase (Accession No. XP_001940956); *Aspergillus niger* beta-mannosidase (Accession No. Q9UUZ3); *Aspergillus aculeatus* beta-mannosidase (Accession No. BAA29029); *Neosartorya fischeri* beta-mannosidase (Accession No. XP_001258000); *Trichoderma reesei* alpha-glucuronidase (Accession No. CAA92949); *Aspergillus niger* alpha-glucuronidase (Accession No. CAC38119); *Talaromyces emersonii* alpha-glucuronidase (Accession No. AAL33576); *Aspergillus niger* acetylxylanesterase (Accession No. XP_001395572); *Trichoderma reesei* acetylxylanesterase (Accession No. Q99034); *Neosartorya fischeri* acetylxylanesterase (Accession No. XP_001262186); *Trichoderma reesei* arabinofuranosidase, 1,4-beta-D-arabinoxylan arabinofuranohydrolase (Accession No. AAP57750); *Chaetomium globosum* arabinofuranosidase, 1,4-beta-D-arabinoxylan arabinofuranohydrolase (Accession No. XP_001223478); *Aspergillus niger* arabinofuranosidase, 1,4-beta-D-arabinoxylan arabinofuranohydrolase (Accession No. XP_001389998); *Penicillium decumbens* Swollenin (Accession No. ACH57439); *Neosartorya fischeri* Swollenin (Accession No. XP_001257521); *Talaromyces stipitatus* Swollenin (Accession No EED19018); *Trichoderma reesei* (Accession No. AAP57751); *Chaetomium globosum* (Accession No. XP_001228455); *Magnaporthe grisea* (Accession No. XP_365869); *Trichoderma reesei* glucuronyl esterase (Accession No. AAP57749); *Chaetomium globosum* glucuronyl esterase (Accession No. XP_001226041); *Aspergillus fumigatus* glucuronyl esterase (Accession No. XP_751313); *Populus alba* alpha-expansin (Accession No. BAB39482); *Vitis lubrusca* alpha-expansin (Accession No. BAC66697); *Triticum aestivum* beta-expansin (Accession No. AAS48881); *Eucalyptus globulus* beta-expansin (Accession No. AAZ08315); *Aspergillus niger* Feruoyl esterase (Accession No. XP_001393337); *Aspergillus terreus* Feruoyl esterase (Accession No. XP_001211092); *Talaromyces stipitatus* Feruoyl esterase (Accession No. EED17739); *Chaetomium globosum* Feruoyl

esterase (Accession No. XP_001228412) *Streptomyces avermitilis* 1,4-beta-cellobiosidase guxA1 (Accession No. NP_821732.1); *Streptomyces avermitilis* 1,4-beta-cellobiosidase guxA2 (Accession No. NP_823029.1); *Streptomyces avermitilis* 1,4-beta-cellobiosidase guxA3 (Accession No. NP_823031.1); *Streptomyces avermitilis* Endo-1,4-beta-glucanase celA1 (Accession No. NP_821730.1); *Streptomyces avermitilis* Endo-1,4-beta-glucanase celA2 (Accession No. NP_823030.1); *Streptomyces avermitilis* Endo-1,4-beta-glucanase celA3 (Accession No. NP_823032.1); *Streptomyces avermitilis* Endo-1,4-beta-glucanase celA4 (Accession No. NP_823744.1); *Streptomyces avermitilis* Endo-1,4-beta-glucanase (Accession No. NP_826394.1); *Streptomyces avermitilis* Endo-1,4-beta-glucanase celA5 (Accession No. NP_828072.1); *Streptomyces avermitilis* Beta-1,4-xylanase (Accession No. NP_823272.1); *Streptomyces avermitilis* Beta-1,4-xylanase (Accession No. NP_826161.1); *Streptomyces avermitilis* Xylanase (Accession No. NP_827548.1); *Streptomyces avermitilis* Endo-1,4-beta-xylanase xynD (Accession No. NP_827557.1); *Streptomyces avermitilis* 1,4-beta-xylosidase xynB1 (Accession No. NP_822628.1); *Streptomyces avermitilis* Beta-xylosidase (Accession No. NP_823285.1); *Streptomyces avermitilis* 1,4-beta-xylosidase xynB2 (Accession No. NP_826159.1); *Streptomyces avermitilis* 1,4-beta-xylosidase xynB3 (Accession No. NP_827745.1); *Streptomyces avermitilis* Beta-glucosidase bglC1 (Accession No. NP_822977.1); *Streptomyces avermitilis* Beta-glucosidase bglC2 (Accession No. NP_826430.1); *Streptomyces avermitilis* Beta-glucosidase bglC3 (Accession No. NP_826775.1); *Streptomyces avermitilis* AXE1 (Accession No. NP_822477.1); *Streptomyces avermitilis* AXE1 (Accession No. NP_822632.1); *Streptomyces avermitilis* abfA (Accession No. NP_822218.1); *Streptomyces avermitilis* abfB (Accession No. NP_822290.1); *Streptomyces avermitilis* abfA (Accession No. NP_826920.1); *Streptomyces avermitilis* abfB (Accession No. BAC74043.1); *Streptomyces avermitilis* SAV_6756 (Accession No. BAC74467.1); *Streptomyces avermitilis* agaA1 (Accession No. BAC68338.1); *Streptomyces avermitilis* agaA3 (Accession No. BAC68787.1); *Streptomyces avermitilis* agaB2 (Accession No. BAC69185.1); *Saccharophagus degradans* 2-40 Sde_2993 (Accession No. YP_528462.1); *Saccharophagus degradans* 2-40 Sde_2996 (Accession No. YP_528465.1); *Saccharophagus degradans* 2-40 Sde_3023 (Accession No. YP_528492.1); *Saccharophagus degradans* 2-40 cel5A (Accession No. ABD82260.1); *Saccharophagus degradans* 2-40 cel5E (Accession No. ABD82186.1); *Saccharophagus*

degradans 2-40 cel5F (Accession No. ABD80834.1); *Saccharophagus degradans* 2-40 cel5J (Accession No. ABD81754.1; *Saccharophagus degradans* 2-40 cel9A (Accession No. ABD79898.1); *Saccharophagus degradans* 2-40 ced3A (Accession No. ABD81757.1); *Saccharophagus degradans* 2-40 ced3B (Accession No. ABD79509.1); *Saccharophagus degradans* 2-40 bgl1A (Accession No. ABD82858.1); *Saccharophagus degradans* 2-40 bgl1B (Accession No. ABD80656.1); *Saccharophagus degradans* 2-40 Cep94A (Accession No. ABD80580.1); *Saccharophagus degradans* 2-40 Cep94B (Accession No. ABD80168.1); *Saccharophagus degradans* 2-40 Sde_0509 (Accession No. YP_525985.1); *Saccharophagus degradans* 2-40 Sde_0169 (Accession No. YP_525645.1); *Bacillus subtilis* Expansin exlX (Accession No. CAB13755.1); *Bacillus subtilis* Endo-1,4-beta-glucanase eglS (Accession No. CAB13696.2); *Bacillus subtilis* Endo-xylanase xynC (Accession No. CAB13698.1); *Bacillus subtilis* Endo-1,4-beta-xylanase xynD (Accession No. CAB13699.1); *Bacillus subtilis* Endo-1,4-beta-xylanase xynA (Accession No. CAB13776.1); *Bacillus subtilis* Xylan beta-1,4-xylosidase xynB (Accession No. CAB13642.2); *Clostridium phytofermentans* Cphy_3367 (Accession No. YP_001560459.1); *Clostridium phytofermentans* Cphy_3368 (Accession No. YP_001560460.1); *Clostridium phytofermentans* Cphy_2058 (Accession No. YP_001559165.1); *Clostridium phytofermentans* Cphy_3202 cellulase B (Accession No. YP_001560295.1); *Clostridium phytofermentans* Cphy_1163 (Accession No. YP_001558280.1); *Clostridium phytofermentans* Cphy_3329 (Accession No. YP_001560421.1); *Clostridium phytofermentans* Cphy_1125 (Accession No. YP_001558242.1); *Clostridium phytofermentans* Cphy_1510 (Accession No. YP_001558623.1); *Clostridium phytofermentans* Cphy_0624 (Accession No. YP_001557750.1); *Clostridium phytofermentans* Cphy_2105 XynA (Accession No. YP_001559210.1); *Clostridium phytofermentans* Cphy_2108 (Accession No. YP_001559213.1); Clostridium phytofermentans Cphy_3207 Y (Accession No. YP_001560300.1); *Clostridium phytofermentans* Cphy_0191 (Accession No. YP_001557317.1); *Clostridium phytofermentans* Cphy_0875 (Accession No. YP_001558000.1); Clostridium phytofermentans Cphy_1169 (Accession No. YP_001558286.1); *Clostridium phytofermentans* Cphy_1071 (Accession No. YP_001558190.1); *Clostridium phytofermentans* Cphy_2128 (Accession No. YP_001559233.1); *Clostridium phytofermentans* Cphy_2276 (Accession No.

YP_001559376.1); *Clostridium phytofermentans* Cphy_1936 (Accession No. YP_001559043.1); *Clostridium cellulolyticum* cel5I (Accession No. AAL79562.1); *Clostridium cellulolyticum* CelCCF (dockerin) Cel48F-yeast CO template pMU914 (Accession No. AAB41452.1); *Clostridium cellulolyticum* Ccel_1259 (Accession No. YP_002505595); *Clostridium cellulolyticum* Ccel_2226 (Accession No. YP_002506548.1); *Clostridium cellulolyticum* Ccel_0732 (dockerin) Cel9E-yeast CO template pMU913 (Accession No. YP_002505091.1); *Clostridium cellulolyticum* Ccel_1099 (dockerin) Cel5A-yeast CO template pMU967 (Accession No. YP_002505438.1); *Clostridium cellulolyticum* Ccel_2392 (dockerin) (Accession No. YP_002506705.1); *Clostridium cellulolyticum* Ccel_0731 (dockerin) Cel9G-yeast CO template pMU892 (Accession No. YP_002505090.1); *Clostridium cellulolyticum* Ccel_0840 (dockerin) Cel5D-yeast CO template pMU891 (Accession No. YP_002505196.1); *Clostridium cellulolyticum* CelCCC (dockerin) Cel8C-yeast CO template pMU969 (Accession No. AAA73867.1); *Thermobifida fusca* endo-1,4-beta xylanase (Accession No. ABL73883.1); *Thermobifida fusca* endo-1,4-beta-D-xylanase (xyl11) (Accession No. AAV64879.1); *Thermobifida fusca* Endoglucanase (Accession No. AAZ55112.1); *Thermobifida fusca* cellulase (Accession No. AAZ56745.1); *Thermobifida fusca* exo-1,4-beta-glucosidase (Accession No. AAZ55642.1); *Thermobifida fusca* beta-glucosidase (Accession No. AAZ55664.1); *Thermobifida fusca* cellulose 1,4-beta-cellobiosidase (Accession No. YP_290015.1); *Thermobifida fusca* CBD E8 (Accession No. AAZ55700.1); *Thermobifida fusca* celC (E3) (Accession No. YP_288681.1); *Thermobifida fusca* celE (E5) (Accession No. YP_288962.1); *Thermobifida fusca* cel5B (Endoglucanase) (Accession No. AAP56348.1); *Thermobifida fusca* celA (E1) (Accession No. AAC06387.1); *Thermobifida fusca* celB (E2) (Accession No. YP_289135.1); *Thermobifida fusca* Tfu_1627 (1,4-beta-cellobiosidase) (Accession No. YP_289685.1); *Clostridium thermocellum* celA (dockerin) (Accession No. YP_001036701.1); *Clostridium thermocellum* celY (cel48Y) (Accession No. CAI06105.1); *Clostridium thermocellum* Cthe_0625 (dockerin) (Accession No. YP_001037053.1); *Clostridium thermocellum* celC (Accession No. CAC27410.1); *Clostridium thermocellum* (Accession No. YP_001037893.1); *Clostridium thermocellum* (Accession No. YP_001038519.1); *Clostridium thermocellum* bglA (Accession No. CAA42814.1); *Clostridium thermocellum* bglB (Accession No. CAA33665.1);

Clostridium thermocellum Cthe_2548 (Accession No. YP_001038942.1); *Clostridium thermocellum* Cthe_1273 (Accession No. YP_001037698.1); *Clostridium thermocellum* Cthe_0040 (Cel9I) (Accession No. YP_001036474.1); *Clostridium thermocellum* Cthe_0412 (dockerin) (Accession No. YP_001036843.1); *Clostridium thermocellum* Cthe_0825 (dockerin) (Accession No. YP_001037253.1); *Clostridium stercorarium* xynA (Accession No. CAD48307); *Clostridium stercorarium* xynB (CelW - celloxylanase) (Accession No. CAD48313); *Clostridium stercorarium* xynC (CelX - celloxylanase) (Accession No. CAD48314); *Clostridium stercorarium* bxlB (b-Xylosidase B) (Accession No. AJ508405); *Clostridium stercorarium* bxlA (b-Xylosidase A) (Accession No. AJ508404) ; *Clostridium stercorarium* bglZ (beta-glucosidase) (Accession No. CAB08072); *Clostridium stercorarium* arfA (alpha-arabinofuranosidaseA) (Accession No. AJ508406); *Clostridium stercorarium* arfB (alpha-arabinofuranosidaseB) (Accession No. AAC28125); *Clostridium stercorarium* celZ (Cs-Cel9Z - Avicellase I) (Accession No CAA39010); *Clostridium stercorarium* celY (Cs-Cel48Y - Avicellase II) (Accession No. CAA93280); *Anaerocellum thermophilum* celA (1,4-beta-glucanase) (Accession No. CAB06786); *Anaerocellum thermophilum* celD (EG) (Accession No. CAB01405); *Anaerocellum thermophilum* xynA (1,4-beta-D-xylan xylanhydrolase) (Accession No. CAA93627); *Anaerocellum thermophilum* celB (EG5) (Accession No. Z86104); *Anaerocellum thermophilum* Athe_1866 (endo-1,4-beta-mannosidase) (Accession No. YP_002573059); *Anaerocellum thermophilum* Athe_0594 ("cellulase") (Accession No. YP_002572493).

[0086] Additionally, host cells of the invention may be used in co-culture with other host cells that are capable of performing some beneficial function. For example, cells capable of breaking down cellulose can be co-cultured with the arabinose-utilizing cells of the invention. In such embodiments, co-cultured cells expresse cellulases and can release sugars into the media. As used herein, "co-culture" refers to growing two different strains or species of host cells together in the same vessel either contemporaneously or serially. Additionally, "co-culture" can mean that different strains or species of host cell are in fluid communication with each other, but in different containers.

[0087] Introduction of a polynucleotide encoding one or more heterologous arabinose-utilizing enzymes into a host cell can be done by methods known in the art. Introduction of polynucleotides encoding heterologous xylose metabolizing enzymes into, for example

yeast host cells, can be effected by lithium acetate transformation, spheroplast transformation, or transformation by electroporation, as described in *Current Protocols in Molecular Biology*, 13.7.1-13.7.10. Introduction of the construct in other host cells can be effected by calcium phosphate transfection, DEAE-Dextran mediated transfection, or electroporation. (Davis, L., *et al., Basic Methods in Molecular Biology*, (1986)).

[0088] In certain embodiments, arabinose-utilizing gene donors may include microorganisms that confer to the host cell, the ability to metabolize hexose and pentose sugars. In some embodiments, such microorganisms are *Thermoanaerobacterium saccharolyticum, C. cellulolyticum, Caldicellulosiruptor kristjanssonii, C. phytofermentans, C. stercorarium, Pyromyces spp.* and *B. thetaiotamicron.*

[0089] Accordingly, it is an embodiment of the invention to modify one or more microorganism strains so as to optimize sugar utilization capability by, for example, introducing genes for one or more enzymes required for the production of a fermentation product from biomass-derived pentoses, *e.g.*, D-xylose or L-arabinose metabolism. Promoters, including native promoters of the host cell may be used to express these genes. Such promoters include, for example, the ADH1 and the ENO1 promoter of *S. cerevisiae*. Suitable yeast promoters are well known in the art. Promoters of the invention may be constitutive or inducible. *See e.g.* Table 1.

[0090] Similarly, terminator sequences normally endogenous to the host cell can also be used to express the genes of the invention. Such terminator sequences include, for example, the PDC1 the ENO1 and the PDC1 terminators from *S. cerevisiae.* Once the gene or genes have been cloned, codon optimization may be performed before expression. Cassettes containing, for example, the native promoter, one or more arabinose-utilization genes and a selectable marker may then be used to transform the host cell and select for successful transformants.

[0091] In additional embodiments, the transformed host cells or cell cultures are assayed for ethanol production. Ethanol production can be measured by techniques known to one of ordinary skill in the art, *e.g.*, by a standard HPLC refractive index method.

Heterologous Arabinose-Metabolizing Enzymes

[0092] According to one aspect of the present invention, the expression of heterologous arabinose-metabolizing enzymes in a host cell can be used advantageously to produce products such as ethanol from the arabinose portion of cellulosic sources. Arabinose-

metabolizing enzymes from a variety of sources can be heterologously expressed to successfully increase efficiency of ethanol production, for example. The arabinose-metabolizing enzymes can be from fungi, bacteria, plants, and protozoan or termite sources. In some embodiments, the arabinose- metabolizing enzyme is a *Thermoanaerobacterium saccharolyticum, H. grisea, T. aurantiacus, T. emersonii, T. reesei, C. lacteus, C. formosanus, N. takasagoensis, C. acinaciformis, M. darwinensis, N. walkeri, S. fibuligera, C. luckowense R. speratus, Thermobfida fusca, Clostridium cellulolyticum, Clostridum josui, Bacillus pumilis, Cellulomonas fimi, Saccharophagus degradans, Piromyces equii, Neocallimastix patricarum* or *Arabidopsis thaliana* arabinose-metabolizing enzyme. In some embodiments, the arabinose-metabolizing enzyme is a *B. thetaiotamicron* arabinose-metabolizing enzyme. In some embodiments, the arabinose-metabolizing enzyme of the invention is any arabinose-metabolizing enzyme known in the art. In a specific embodiment, the arabinose-metabolizing enzyme of the invention is an enzyme disclosed in Table 2. In some embodiments, the arabinose-metabolizing enzyme is encoded by a nucleic acid sequence at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% identical to any one of SEQ ID NOs: 1-5. In some embodiments, the arabinose-metabolizing enzyme has an amino acid sequence that is at least about 80%, at least about 85%, at least about 90%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, at least about 99%, or 100% identical to any one of SEQ ID NOs: 6-10. In some embodiments, the arabinose-metabolizing enzyme of the invention is any arabinose-metabolizing enzyme suitable for expression in an appropriate host cell.

[0093] In some embodiments of the invention, multiple arabinose-metabolizing enzymes from a single organism are co-expressed in the same host cell. In some embodiments of the invention, multiple arabinose-metabolizing enzymes from different organisms are co-expressed in the same host cell. In particular, arabinose-metabolizing enzymes from two, three, four, five, six, seven, eight, nine or more organisms can be co-expressed in the same host cell. Similarly, the invention can encompass co-cultures of microorganism strains, wherein the microorganism strains express different arabinose-metabolizing enzymes. Co-cultures can include microorganism strains expressing heterologous arabinose-metabolizing enzymes from the same organism or from different organisms.

Co-cultures can include microorganism strains expressing arabinose-metabolizing enzymes from two, three, four, five, six, seven, eight, nine or more microorganisms.

[0094] In some aspects of the invention, one or more of the native enzymes in the engineered metabolic pathways are down-regulated or deleted. In certain embodiments, the downregulated or deleted native enzyme is an enzyme involved in central metabolism. In some embodiments, the downregulated or deleted native enzyme is selected from the group consisting of a pyruvate kinase; a hydrogenase; a lactate dehydrogenase; a phosphotransacetylase; an acetate kinase; an acetaldehyde dehydrogenase; a glyceraldehyde phosphate dehydrogenase; pyruvate formate lyase, an aldose reductase; an alcohol dehydrogenase; a pyruvate formate lyase; a pyruvate decarboxylase; and combinations thereof.

[0095] In certain embodiments of the invention, the arabinose-metabolizing enzyme can be an arabinose transporter (AraT), an arabinose isomerase (AI), a ribulokinanse (RK), and a ribulose 5-phosphate epimerase (R5PE). Cells of the invention may additionally express other enzymes associated with pentose utilization including a xylose isomerase, a transketolase, and a transaldolase or other enzymes of the pentose phosphate pathway.

[0096] As a practical matter, whether any polypeptide is at least 70%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 100% identical to a polypeptide of the present invention can be determined conventionally using known computer programs. Methods for determining percent identity, as discussed in more detail below in relation to polynucleotide identity, are also relevant for evaluating polypeptide sequence identity.

[0097] In some particular embodiments of the invention, amino acid and nucleic acid sequences are readily determined for a gene, protein or other element by an accession number upon consulting the proper database, for example Genebank. However, sequences for exemplary genes and proteins of the present invention are also disclosed herein (SEQ ID NOs: 1-26).

[0098] Some embodiments of the invention encompass a polypeptide comprising at least 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, or 500 or more consecutive amino acids of any of SEQ ID NOs: 6-23 & 25, or domains, fragments, variants, or derivatives.

[0099] In certain aspects of the invention, the polypeptides and polynucleotides of the present invention are provided in an isolated form, *e.g.*, purified to homogeneity.

[00100] The present invention also encompasses polypeptides which comprise, or alternatively consist of, an amino acid sequence which is at least about 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% similar to the polypeptide of any of SEQ ID NOs: 6-23 & 25 and to portions of such polypeptide with such portion of the polypeptide generally containing at least 30 amino acids and more preferably at least 50 amino acids.

[0100] As known in the art "similarity" between two polypeptides is determined by comparing the amino acid sequence and conserved amino acid substitutes thereto of the polypeptide to the sequence of a second polypeptide.

[0101] The present invention further relates to a domain, fragment, variant, derivative, or analog of the polypeptide of any of SEQ ID NOs: 6-23 & 25.

[0102] Fragments or portions of the polypeptides of the present invention can be employed for producing the corresponding full-length polypeptide by peptide synthesis. Therefore, the fragments can be employed as intermediates for producing the full-length polypeptides.

[0103] Fragments of arabinose-metabolizing enzymes of the invention encompass domains, proteolytic fragments, deletion fragments and fragments of any of the genes which retain any specific biological activity of the native enzyme.

[0104] The variant, derivative or analog of the polypeptide of arabinose-metabolizing enzymes of the invention may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or (iv) one in which the additional amino acids are fused to the mature polypeptide for purification of the polypeptide. Such variants, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

[0105] The polypeptides of the present invention further include variants of the polypeptides. A "variant" of the polypeptide can be a conservative variant, or an allelic variant. As used herein, a "conservative variant" refers to alterations in the amino acid sequence that do not adversely affect the biological functions of the protein. A

substitution, insertion or deletion is said to adversely affect the protein when the altered sequence prevents or disrupts a biological function associated with the protein. For example, the overall charge, structure or hydrophobic-hydrophilic properties of the protein can be altered without adversely affecting a biological activity. Accordingly, the amino acid sequence can be altered, for example to render the peptide more hydrophobic or hydrophilic, without adversely affecting the biological activities of the protein.

[0106] An "allelic variant" as used herein, is intended to designate alternate forms of a gene occupying a given locus on a chromosome of an organism. *Genes II*, Lewin, B., ed., John Wiley & Sons, New York (1985). Non-naturally occurring variants may be produced using art-known mutagenesis techniques. Allelic variants, though possessing a slightly different amino acid sequence than those recited above, will still have the same or similar biological functions associated with the arabinose-metabolizing enzymes of the invention. The allelic variants, the conservative substitution variants, and members of the arabinose-metabolizing enzymes can have an amino acid sequence having at least 75%, at least 80%, at least 90%, at least 95% amino acid sequence identity with the arabinose-metabolizing enzymes of the invention, and, particularly, with the amino acid sequence set forth in any one of SEQ ID NOs: 6-23 & 25. Identity or homology with respect to such sequences is defined herein as the percentage of amino acid residues in the candidate sequence that are identical with the known peptides, after aligning the sequences and introducing gaps, if necessary, to achieve the maximum percent homology, and not considering any conservative substitutions as part of the sequence identity. N-terminal, C-terminal or internal extensions, deletions, or insertions into the peptide sequence shall not be construed as affecting homology.

[0107] Thus, in one aspect the proteins and peptides of the present invention include molecules comprising the amino acid sequence of SEQ ID NOs: 6-23 & 25 or fragments thereof having a consecutive sequence of at least about 3, 4, 5, 6, 10, 15, 20, 25, 30, 35 or more amino acid residues of the arabinose transporters (AraT) , arabinose isomerase (AI) , ribulokinanse (RK), or the ribulose 5-phosphate epimerase (R5PE; amino acid sequence variants of such sequences wherein at least one amino acid residue has been inserted N- or C- terminal to, or within, the disclosed sequence; amino acid sequence variants of the disclosed sequences, or their fragments as defined above, that have been substituted by another residue. Contemplated variants further include those containing predetermined

mutations by, *e.g.*, homologous recombination, site-directed or PCR mutagenesis, and the corresponding proteins of other species, including, but not limited to bacterial, fungal, and insect.

[0108] Using known methods of protein engineering and recombinant DNA technology, variants may be generated to improve or alter the characteristics of the polypeptides encoding the arabinose-metabolizing enzymes. For instance, one or more amino acids can be deleted from the N-terminus or C-terminus of a secreted protein without substantial loss of biological function.

[0109] Thus, in another aspect the invention further includes arabinose transporters (AraT), arabinose isomerase (AI), ribulokinanse (RK), and ribulose 5-phosphate epimerase (R5PE) polypeptide variants which show substantial biological activity. Such variants include deletions, insertions, inversions, repeats, and substitutions selected according to general rules known in the art so as have little effect on activity.

[0110] The skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly effect protein function (*e.g.*, replacing one aliphatic amino acid with a second aliphatic amino acid), as further described below.

[0111] For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided in Bowie *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247*:1306-1310 (1990), wherein the authors indicate that there are two main strategies for studying the tolerance of an amino acid sequence to change.

[0112] The first strategy exploits the tolerance of amino acid substitutions by natural selection during the process of evolution. By comparing amino acid sequences in different species, conserved amino acids can be identified. These conserved amino acids are likely important for protein function. In contrast, the amino acid positions where substitutions have been tolerated by natural selection indicates that these positions are not critical for protein function. Thus, positions tolerating amino acid substitution could be modified while still maintaining biological activity of the protein.

[0113] The second strategy uses genetic engineering to introduce amino acid changes at specific positions of a cloned gene to identify regions critical for protein function. For example, site directed mutagenesis or alanine-scanning mutagenesis (introduction of single alanine mutations at every residue in the molecule) can be used. (Cunningham and

Wells, *Science 244*:1081-1085 (1989).) The resulting mutant molecules can then be tested for biological activity.

[0114] These two strategies have revealed that proteins are often surprisingly tolerant of amino acid substitutions. The authors further indicate which amino acid changes are likely to be permissive at certain amino acid positions in the protein. For example, most buried (within the tertiary structure of the protein) amino acid residues require nonpolar side chains, whereas few features of surface side chains are generally conserved. Moreover, tolerated conservative amino acid substitutions involve replacement of the aliphatic or hydrophobic amino acids Ala, Val, Leu and Ile; replacement of the hydroxyl residues Ser and Thr; replacement of the acidic residues Asp and Glu; replacement of the amide residues Asn and Gln, replacement of the basic residues Lys, Arg, and His; replacement of the aromatic residues Phe, Tyr, and Trp, and replacement of the small-sized amino acids Ala, Ser, Thr, Met, and Gly.

[0115] The terms "derivative" and "analog" refer to a polypeptide differing from the arabinose-metabolizing enzymes of the invention, but retaining essential properties thereof. Generally, derivatives and analogs are overall closely similar, and, in many regions, identical to the arabinose-metabolizing enzymes of the invention. The terms "derived-from", "derivative" and "analog" when referring to arabinose transporters (AraT), arabinose isomerases (AI), ribulokinanses (RK), and ribulose 5-phosphate epimerases (R5PE) of the invention include any polypeptides which retain at least some of the activity of the corresponding native polypeptide, *e.g.*, the arabinose isomerase activity, or the activity of the its catalytic domain.

[0116] Derivatives of the arabinose-metabolizing, xylose-metabolizing and cellulase enzymes disclosed herein are polypeptides which may have been altered so as to exhibit features not found on the native polypeptide. Derivatives can be covalently modified by substitution (*e.g.* amino acid substitution), chemical, enzymatic, or other appropriate means with a moiety other than a naturally occurring amino acid (*e.g.*, a detectable moiety such as an enzyme or radioisotope). Examples of derivatives include fusion proteins, or proteins which are based on a naturally occurring protein sequence, but which have been altered. For example, proteins can be designed by knowledge of a particular amino acid sequence, and/or a particular secondary, tertiary, and/or quaternary structure. Derivatives include proteins that are modified based on the knowledge of a previous

sequence, natural or synthetic, which is then optionally modified, often, but not necessarily to confer some improved function. These sequences, or proteins, are then said to be derived from a particular protein or amino acid sequence. In some embodiments of the invention, a derivative must retain at least about 50% identity, at least about 60% identity, at least about 70% identity, at least about 80% identity, at least about 90% identity, at least about 95% identity, at least about 97% identity, or at least about 99% identity to the sequence the derivative is "derived-from." In some embodiments of the invention, an arabinose-metabolizing, xylose-metabolizing or cellulase enzyme is said to be derived-from an enzyme naturally found in a particular species if, using molecular genetic techniques, the DNA sequence for part or all of the enzyme is amplified and placed into a new host cell.

[0117] An "analog" is another form of a arabinose transporters (AraT), arabinose isomerase (AI), ribulokinanse (RK), and a ribulose 5-phosphate epimerase (R5PE) polypeptide of the present invention. An analog also retains substantially the same biological function or activity as the polypeptide of interest, *e.g.*, functions as an arabinose isomerase. An analog includes a proprotein which can be activated by cleavage of the proprotein portion to produce an active mature polypeptide.

[0118] The polypeptide of the present invention may be a recombinant polypeptide, a natural polypeptide or a synthetic polypeptide. In some particular embodiments, the polypeptide is a recombinant polypeptide.

Combinations of Arabinose-Metabolizing Enzymes

[0119] In some embodiments of the present invention, the host cell expresses a combination of heterologous arabinose-metabolizing enzymes. For example, the host cell can contain at least two heterologous arabinose-metabolizing enzymes, at least three heterologous arabinose-metabolizing enzymes, at least four heterologous arabinose-metabolizing enzymes, at least five heterologous arabinose-metabolizing enzymes, at least six heterologous arabinose-metabolizing enzymes, at least seven heterologous arabinose-metabolizing enzymes, or at least eight heterologous arabinose-metabolizing enzymes. The heterologous arabinose-metabolizing enzymes in the host cell can be from the same or from different species (*e.g.* one from a bacterial species and one from a eukaryotic species). In one embodiment, the one or more heterologous arabinose-metabolizing enzymes are contained in an operon.

Fusion Proteins Comprising Arabinose-Metabolizing Enzymes

[0120] The present invention also encompasses fusion proteins. For example, the fusion proteins can be a fusion of a heterologous arabinose-metabolizing enzyme and a second peptide. The heterologous arabinose-metabolizing enzyme and the second peptide can be fused directly or indirectly, for example, through a linker sequence. The fusion protein can comprise for example, a second peptide that is N-terminal to the heterologous arabinose-metabolizing enzyme and/or a second peptide that is C-terminal to the heterologous arabinose-metabolizing enzyme. Thus, in certain embodiments, the polypeptide of the present invention comprises a first polypeptide and a second polypeptide, wherein the first polypeptide comprises a heterologous arabinose-metabolizing enzyme.

[0121] According to one aspect of the present invention, the fusion protein can comprise a first and second polypeptide wherein the first polypeptide comprises a heterologous arabinose-metabolizing enzyme and the second polypeptide comprises a signal sequence. According to another embodiment, the fusion protein can comprise a first and second polypeptide, wherein the first polypeptide comprises a heterologous arabinose-metabolizing enzyme and the second polypeptide comprises a polypeptide used to facilitate purification or identification or a reporter peptide. The polypeptide used to facilitate purification or identification or the reporter peptide can be, for example, a HIS-tag, a GST-tag, an HA-tag, a FLAG-tag, or a MYC-tag.

[0122] According to another embodiment, the fusion protein can comprise a first and second polypeptide, wherein the first polypeptide comprises a heterologous arabinose-metabolizing enzyme and the second polypeptide comprises a fluorescent protein. In one aspect, the fluorescent protein is used to detect the heterologous arabinose-metabolizing enzyme fusion protein.

[0123] According to yet another embodiment, the fusion protein can comprise a first and second polypeptide, wherein the first polypeptide comprises a heterologous arabinose-metabolizing enzyme and the second polypeptide comprises an anchoring peptide.

[0124] According to still another embodiment, the fusion protein can comprise a first and second polypeptide, wherein the first polypeptide comprises a heterologous arabinose-metabolizing enzyme and the second polypeptide comprises a cellulose binding module (CBM).

[0125] In certain other embodiments, the first polypeptide and the second polypeptide are fused via a linker sequence. The linker sequence can, in some embodiments, be encoded by a codon-optimized polynucelotide. (Codon-optimized polynucleotides are described in more detail below).

Co-Cultures

[0126] In another aspect, the present invention is directed to co-cultures comprising at least two host cells wherein the at least two host cells each comprise an isolated polynucleotide encoding a heterologous xylose metabolizing enzyme. In one embodiment, the co-culture can comprise two or more strains of host cells and the heterologous arabinose-utilizing enzymes can be expressed in any combination in the two or more strains of host cells.

[0127] The various host cell strains in the co-culture can be present in equal numbers, or one strain or species of host cell can significantly outnumber another second strain or species of host cells. For example, in a co-culture comprising two strains or species of host cells the ratio of one host cell to another can be about 1:1, 1:2, 1:3, 1:4, 1:5, 1:10, 1:100, 1:500 or 1:1000. Similarly, in a co-culture comprising three or more strains or species of host cells, the strains or species of host cells may be present in equal or unequal numbers.

Polynucleotides Encoding Heterologous Arabinose-Metabolizing Enzymes

[0128] In another aspect, the present invention includes isolated polynucleotides encoding arabinose-metabolizing enzymes of the present invention. The polynucleotides can encode an arabinose transporter (AraT), arabinose isomerase (AI), ribulokinanse (RK), and/or ribulose 5-phosphate epimerase (R5PE).

[0129] The present invention also encompasses an isolated polynucleotide comprising a nucleic acid that is at least about 70%, 75%, or at least about 80% identical, at least about 90% to about 95% identical, or at least about 96%, 97%, 98%, 99% or 100% identical to a nucleic acid encoding an arabinose transporter (AraT), arabinose isomerase (AI), ribulokinanse (RK), and ribulose 5-phosphate epimerase (R5PE).

[0130] The present invention also encompasses variants of the arabinose-metabolizing enzyme genes. Variants may contain alterations in the coding regions, non-coding regions, or both. Examples are polynucleotide variants containing alterations which

produce silent substitutions, additions, or deletions, but do not alter the properties or activities of the encoded polypeptide. In certain embodiments, nucleotide variants are produced by silent substitutions due to the degeneracy of the genetic code. In further embodiments the arabinose transporter (AraT), arabinose isomerase (AI), ribulokinanse (RK), and ribulose 5-phosphate epimerase (R5PE) polynucleotide variants can be produced for a variety of reasons, *e.g.*, to optimize codon expression for a particular host. Codon-optimized polynucleotides of the present invention are discussed further below.

[0131] The present invention also encompasses an isolated polynucleotide encoding a fusion protein. In further embodiments, the first and second polynucleotides are in the same orientation, or the second polynucleotide is in the reverse orientation of the first polynucleotide. In additional embodiments, the first polynucleotide encodes a polypeptide that is either N-terminal or C-terminal to the polypeptide encoded by the second polynucleotide. In certain other embodiments, the first polynucleotide and/or the second polynucleotide are encoded by codon-optimized polynucleotides, for example, polynucleotides codon-optimized for *S. cerevisiae.*

[0132] Also provided in the present invention are allelic variants, orthologs, and/or species homologs. Procedures known in the art can be used to obtain full-length genes, allelic variants, splice variants, full-length coding portions, orthologs, and/or species homologs of genes corresponding to any of SEQ ID NOs: 1-5, or other arabinose-metabolizing enzymes using information from the sequences disclosed herein or the clones deposited with the ATCC or otherwise publically available. For example, allelic variants and/or species homologs may be isolated and identified by making suitable probes or primers from the sequences provided herein and screening a suitable nucleic acid source for allelic variants and/or the desired homolog.

[0133] By a nucleic acid having a nucleotide sequence at least, for example, 95% "identical" to a reference nucleotide sequence of the present invention, it is intended that the nucleotide sequence of the nucleic acid is identical to the reference sequence except that the nucleotide sequence may include up to five point mutations per each 100 nucleotides of the reference nucleotide sequence encoding the particular polypeptide. In other words, to obtain a nucleic acid having a nucleotide sequence at least 95% identical to a reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to

5% of the total nucleotides in the reference sequence may be inserted into the reference sequence. In one embodiment, the query sequence may be an entire sequence shown of any of SEQ ID NOs: 1-5 or any fragment or domain specified as described herein.

[0134] As a practical matter, whether any particular nucleic acid molecule or polypeptide is at least 80%, 85%, 90%, 95%, 96%, 97%, 98% or 99% identical to a nucleotide sequence or polypeptide of the present invention can be determined conventionally using known computer programs. A method for determining the best overall match between a query sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag *et al.* (*Comp. App. Biosci.* (1990) 6:237-245.) In a sequence alignment the query and subject sequences are both DNA sequences. An RNA sequence can be compared by converting U's to T's. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB alignment of DNA sequences to calculate percent identity are: Matrix=Unitary, k-tuple=4, Mismatch Penalty=1, Joining Penalty=30, Randomization Group Length=0, Cutoff Score=1, Gap Penalty=5, Gap Size Penalty 0.05, Window Size=500 or the length of the subject nucleotide sequence, whichever is shorter.

[0135] If the subject sequence is shorter than the query sequence because of 5' or 3' deletions, not because of internal deletions, a manual correction must be made to the results. This is because the FASTDB program does not account for 5' and 3' truncations of the subject sequence when calculating percent identity. For subject sequences truncated at the 5' or 3' ends, relative to the query sequence, the percent identity is corrected by calculating the number of bases of the query sequence that are 5' and 3' of the subject sequence, which are not matched/aligned, as a percent of the total bases of the query sequence. Whether a nucleotide is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using the specified parameters, to arrive at a final percent identity score. This corrected score is what is used for the purposes of the present invention. Only bases outside the 5' and 3' bases of the subject sequence, as displayed by the FASTDB alignment, which are not matched/aligned with the query sequence, are calculated for the purposes of manually adjusting the percent identity score.

[0136] For example, a 90 base subject sequence is aligned to a 100 base query sequence to determine percent identity. The deletions occur at the 5' end of the subject sequence and therefore, the FASTDB alignment does not show a matched/alignment of the first 10 bases at 5' end. The 10 unpaired bases represent 10% of the sequence (number of bases at the 5' and 3' ends not matched/total number of bases in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 bases were perfectly matched the final percent identity would be 90%. In another example, a 90 base subject sequence is compared with a 100 base query sequence. This time the deletions are internal deletions so that there are no bases on the 5' or 3' of the subject sequence which are not matched/aligned with the query. In this case the percent identity calculated by FASTDB is not manually corrected. Once again, only bases 5' and 3' of the subject sequence which are not matched/aligned with the query sequence are manually corrected for. No other manual corrections are to be made for the purposes of the present invention.

[0137] Some embodiments of the invention encompass a nucleic acid molecule comprising at least 10, 20, 30, 35, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, or 800 consecutive nucleotides or more of any of SEQ ID NOs: 1-5, or domains, fragments, variants, or derivatives thereof.

[0138] The polynucleotide of the present invention may be in the form of RNA or in the form of DNA, which DNA includes cDNA, genomic DNA, and synthetic DNA. The DNA may be double stranded or single-stranded, and if single stranded can be the coding strand or non-coding (anti-sense) strand. In one embodiment, the coding sequence which encodes the mature polypeptide can be identical to the coding sequence encoding SEQ ID NO: 1-5, or may be a different coding sequence which coding sequence, as a result of the redundancy or degeneracy of the genetic code, encodes the same mature polypeptide as the nucleic acid sequences of any one of SEQ ID NOs: 1-5.

[0139] In certain embodiments, the present invention provides an isolated polynucleotide comprising a nucleic acid fragment which encodes at least 10, at least 20, at least 30, at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 95, or at least 100 or more contiguous amino acids of SEQ ID NOs: 6-23 & 25.

[0140] The polynucleotide encoding the mature polypeptide of SEQ ID NOs: 6-23 & 25 may include: only the coding sequence for the mature polypeptide; the coding sequence

of any domain of the mature polypeptide; and the coding sequence for the mature polypeptide (or domain-encoding sequence) together with non coding sequence, such as introns or non-coding sequence 5' and/or 3' of the coding sequence for the mature polypeptide.

[0141] Thus, the term "polynucleotide encoding a polypeptide" encompasses a polynucleotide which includes only sequences encoding for the polypeptide as well as a polynucleotide which includes additional coding and/or non-coding sequences.

[0142] Due to the degeneracy of the genetic code, one of ordinary skill in the art will immediately recognize that a large portion of the nucleic acid molecules having a sequence at least 90%, 95%, 96%, 97%, 98%, or 99% identical to the nucleic acid sequence of any of SEQ ID NOs: 1-5, or fragments thereof, will encode polypeptides having functional activity. In fact, since degenerate variants of any of these nucleotide sequences encode the same polypeptide, in many instances, this will be clear to the skilled artisan even without performing the above described comparison assay. It will be further recognized in the art that, for such nucleic acid molecules that are not degenerate variants, a reasonable number will also encode a polypeptide having functional activity.

[0143] The polynucleotides of the present invention also comprise nucleic acids encoding arabinose-metabolizing enzyme or domain, fragment, variant, or derivative thereof, fused to a polynucleotide encoding a marker sequence which allows for detection of the polynucleotide of the present invention. In one embodiment of the invention, expression of the marker is independent from expression of the arabinose-metabolizing enzyme.

[0144] In one embodiment, the one or more polynucleotides of the present invention are stably integrated into the genome of the host cell. In one aspect, the polynucleotides are randomly integrated into the genome of the host cell. In another aspect, multiple copies of polynucleotides are randomly integrated into the genome of the host cell. In one aspect, at least two copies of polynucleotides are randomly integrated into the genome of the host cell.

[0145] In another embodiment, the one or more polynucleotides are not integrated into the genome of the host cell. In one aspect, the one or more polynucleotides are present in the host cell in an extra chromosomal plasmid.

[0146] In one embodiment, one or more polynucleotides of the present invention are stably integrated at a specific site in the genome of the host cell. In one aspect, the one or

more polynucleotides are stably integrated at the site of one or more specific genes in the genome of the host cell. In one embodiment, the one or more specific genes are disrupted as a result of the one or more integration events. In another aspect, the one or more specific genes are deleted as a result of the one or more integration events. In one embodiment, the host cell cannot make the protein product(s) of the one or more specific disrupted genes. In another aspect, the host cell cannot make the protein product(s) of the one or more specific deleted genes. In another embodiment, the one or more polynucleotides are stably integrated at the site of the rDNA in the genome of the host cell.

[0147] In one embodiment, the start codon of a polynucleotide of the present invention is integrated in frame with the promoter of a specific gene in the genome of the host cell. In another embodiment, the stop codon of a polynucleotide of the invention is integrated in frame with the terminator of a specific gene in the genome of the host cell. In one embodiment, the start codon of a polynucleotides is integrated in frame with the promoter of a specific gene in the genome of the host cell, and the terminator of the same polynucleotide is also integrated in frame with the terminator of the specific gene.

[0148] In one embodiment, the one or more polynucleotides are part of an operon. In one aspect, the start codon of the first polynucleotides in the operon is integrated in frame with the promoter of a specific gene in the genome of the host cell. In another aspect, the stop codon of the last polynucleotides in the operon is integrated in frame with the terminator of a specific gene in the genome of the host cell. In one embodiment, the start codon of the first polynucleotide in the operon is integrated in frame with the promoter of a specific gene in the genome of the host cell, and the stop codon of the last polynucleotide in the operon is integrated in frame with the terminator of the specific gene.

Codon Optimized Polynucleotides

[0149] The polynucleotides of the invention can be codon-optimized. As used herein the term "codon-optimized coding region" means a nucleic acid coding region that has been adapted for expression in the cells of a given organism by replacing at least one, or more than one, or a significant number, of codons with one or more codons that are more frequently used in the genes of that organism.

[0150] In general, highly expressed genes in an organism are biased towards codons that are recognized by the most abundant tRNA species in that organism. One measure of this bias is the "codon adaptation index" or "CAI," which measures the extent to which the codons used to encode each amino acid in a particular gene are those which occur most frequently in a reference set of highly expressed genes from an organism.

[0151] The CAI of codon optimized sequences of the present invention corresponds to between about 0.8 and 1.0, between about 0.8 and 0.9, or about 1.0. A codon optimized sequence may be further modified for expression in a particular organism, depending on that organism's biological constraints. For example, large runs of "As" or "Ts" (*e.g.*, runs greater than 3, 4, 5, 6, 7, 8, 9, or 10 consecutive bases) can be removed from the sequences if these are known to effect transcription negatively. Furthermore, specific restriction enzyme sites may be removed for molecular cloning purposes. Examples of such restriction enzyme sites include PacI, AscI, BamHI, BglII, EcoRI and XhoI. Additionally, the DNA sequence can be checked for direct repeats, inverted repeats and mirror repeats with lengths of ten bases or longer, which can be modified manually by replacing codons with "second best" codons, *i.e.*, codons that occur at the second highest frequency within the particular organism for which the sequence is being optimized.

[0152] Deviations in the nucleotide sequence that comprise the codons encoding the amino acids of any polypeptide chain allow for variations in the sequence coding for the gene. Since each codon consists of three nucleotides, and the nucleotides comprising DNA are restricted to four specific bases, there are 64 possible combinations of nucleotides, 61 of which encode amino acids (the remaining three codons encode signals ending translation). The "genetic code" which shows which codons encode which amino acids is reproduced herein as Table 4. As a result, many amino acids are designated by more than one codon. For example, the amino acids alanine and proline are coded for by four triplets, serine and arginine by six, whereas tryptophan and methionine are coded by just one triplet. This degeneracy allows for DNA base composition to vary over a wide range without altering the amino acid sequence of the proteins encoded by the DNA.

TABLE 4: The Standard Genetic Code

	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC " TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

[0153] Many organisms display a bias for use of particular codons to code for insertion of a particular amino acid in a growing peptide chain. Codon preference or codon bias, differences in codon usage between organisms, is afforded by degeneracy of the genetic code, and is well documented among many organisms. Codon bias often correlates with the efficiency of translation of messenger RNA (mRNA), which is in turn believed to be dependent on, *inter alia*, the properties of the codons being translated and the availability of particular transfer RNA (tRNA) molecules. The predominance of selected tRNAs in a cell is generally a reflection of the codons used most frequently in peptide synthesis. Accordingly, genes can be tailored for optimal gene expression in a given organism based on codon optimization.

[0154] Given the large number of gene sequences available for a wide variety of animal, plant and microbial species, it is possible to calculate the relative frequencies of codon usage. Codon usage tables are readily available, for example, at http://www.kazusa.or.jp/codon/ (visited October 5, 2011), and these tables can be adapted in a number of ways. *See* Nakamura, Y., *et al.* "Codon usage tabulated from the international DNA sequence databases: status for the year 2000," *Nucl. Acids Res. 28*:292

(2000). Codon usage tables for yeast, calculated from GenBank Release 128.0 [15 February 2002], are reproduced below as Table 5. This table uses mRNA nomenclature, and so instead of thymine (T) which is found in DNA, the table uses uracil (U) which is found in RNA. The table has been adapted so that frequencies are calculated for each amino acid, rather than for all 64 codons.

TABLE 5: Codon Usage Table for *Saccharomyces cerevisiae* Genes

Amino Acid	Codon	Number	Frequency per hundred
Phe	UUU	170666	26.1
Phe	UUC	120510	18.4
Leu	UUA	170884	26.2
Leu	UUG	177573	27.2
Leu	CUU	80076	12.3
Leu	CUC	35545	5.4
Leu	CUA	87619	13.4
Leu	CUG	68494	10.5
Ile	AUU	196893	30.1
Ile	AUC	112176	17.2
Ile	AUA	116254	17.8
Met	AUG	136805	20.9
Val	GUU	144243	22.1
Val	GUC	76947	11.8
Val	GUA	76927	11.8
Val	GUG	70337	10.8
Ser	UCU	153557	23.5
Ser	UCC	92923	14.2
Ser	UCA	122028	18.7
Ser	UCG	55951	8.6
Ser	AGU	92466	14.2
Ser	AGC	63726	9.8
Pro	CCU	88263	13.5
Pro	CCC	44309	6.8
Pro	CCA	119641	18.3
Pro	CCG	34597	5.3
Thr	ACU	132522	20.3
Thr	ACC	83207	12.7

Amino Acid	Codon	Number	Frequency per hundred
Thr	ACA	116084	17.8
Thr	ACG	52045	8.0
Ala	GCU	138358	21.2
Ala	GCC	82357	12.6
Ala	GCA	105910	16.2
Ala	GCG	40358	6.2
Tyr	UAU	122728	18.8
Tyr	UAC	96596	14.8
His	CAU	89007	13.6
His	CAC	50785	7.8
Gln	CAA	178251	27.3
Gln	CAG	79121	12.1
Asn	AAU	233124	35.7
Asn	AAC	162199	24.8
Lys	AAA	273618	41.9
Lys	AAG	201361	30.8
Asp	GAU	245641	37.6
Asp	GAC	132048	20.2
Glu	GAA	297944	45.6
Glu	GAG	125717	19.2
Cys	UGU	52903	8.1
Cys	UGC	31095	4.8
Trp	UGG	67789	10.4
Arg	CGU	41791	6.4
Arg	CGC	16993	2.6
Arg	CGA	19562	3.0
Arg	CGG	11351	1.7
Arg	AGA	139081	21.3
Arg	AGG	60289	9.2
Gly	GGU	156109	23.9
Gly	GGC	63903	9.8
Gly	GGA	71216	10.9

Amino Acid	Codon	Number	Frequency per hundred
Gly	GGG	39359	6.0
Stop	UAA	6913	1.1
Stop	UAG	3312	0.5
Stop	UGA	4447	0.7

[0155] By utilizing this or similar tables, one of ordinary skill in the art can apply the frequencies to any given polypeptide sequence, and produce a nucleic acid fragment of a codon-optimized coding region which encodes the polypeptide, but which uses codons optimal for a given species. Codon-optimized coding regions can be designed by various different methods.

[0156] In one method, a codon usage table is used to find the single most frequent codon used for any given amino acid, and that codon is used each time that particular amino acid appears in the polypeptide sequence. For example, referring to Table 5 above, for leucine, the most frequent codon is UUG, which is used 27.2% of the time. Thus all the leucine residues in a given amino acid sequence would be assigned the codon UUG.

[0157] In another method, the actual frequencies of the codons are distributed randomly throughout the coding sequence. Thus, using this method for optimization, if a hypothetical polypeptide sequence had 100 leucine residues, referring to Table 5 for frequency of usage in the *S. cerevisiae*, about 5, or 5% of the leucine codons would be CUC, about 11, or 11% of the leucine codons would be CUG, about 12, or 12% of the leucine codons would be CUU, about 13, or 13% of the leucine codons would be CUA, about 26, or 26% of the leucine codons would be UUA, and about 27, or 27% of the leucine codons would be UUG.

[0158] These frequencies would be distributed randomly throughout the leucine codons in the coding region encoding the hypothetical polypeptide. As will be understood by those of ordinary skill in the art, the distribution of codons in the sequence can vary significantly using this method; however, the sequence always encodes the same polypeptide.

[0159] When using the methods above, the term "about" is used precisely to account for fractional percentages of codon frequencies for a given amino acid. As used herein, "about" is defined as one amino acid more or one amino acid less than the value given.

The whole number value of amino acids is rounded up if the fractional frequency of usage is 0.50 or greater, and is rounded down if the fractional frequency of use is 0.49 or less. Using again the example of the frequency of usage of leucine in human genes for a hypothetical polypeptide having 62 leucine residues, the fractional frequency of codon usage would be calculated by multiplying 62 by the frequencies for the various codons. Thus, 7.28 percent of 62 equals 4.51 UUA codons, or "about 5," *i.e.*, 4, 5, or 6 UUA codons, 12.66 percent of 62 equals 7.85 UUG codons or "about 8," *i.e.*, 7, 8, or 9 UUG codons, 12.87 percent of 62 equals 7.98 CUU codons, or "about 8," *i.e.*, 7, 8, or 9 CUU codons, 19.56 percent of 62 equals 12.13 CUC codons or "about 12," *i.e.*, 11, 12, or 13 CUC codons, 7.00 percent of 62 equals 4.34 CUA codons or "about 4," *i.e.*, 3, 4, or 5 CUA codons, and 40.62 percent of 62 equals 25.19 CUG codons, or "about 25," *i.e.*, 24, 25, or 26 CUG codons.

[0160] Randomly assigning codons at an optimized frequency to encode a given polypeptide sequence, can be done manually by calculating codon frequencies for each amino acid, and then assigning the codons to the polypeptide sequence randomly. Additionally, various algorithms and computer software programs are readily available to those of ordinary skill in the art. For example, the "EditSeq" function in the Lasergene Package, available from DNAstar, Inc., Madison, WI, the backtranslation function in the VectorNTI Suite, available from InforMax, Inc., Bethesda, MD, and the "backtranslate" function in the GCG--Wisconsin Package, available from Accelrys, Inc., San Diego, CA. In addition, various resources are publicly available to codon-optimize coding region sequences, *e.g.*, the "backtranslation" function at http://www.entelechon.com/bioinformatics/backtranslation.php?lang=eng (visited December 18, 2009) and the "backtranseq" function available at http://emboss.bioinformatics.nl/cgi-bin/emboss/backtranseq (visited October 5, 2011). Constructing a rudimentary algorithm to assign codons based on a given frequency can also easily be accomplished with basic mathematical functions by one of ordinary skill in the art.

[0161] A number of options are available for synthesizing codon optimized coding regions designed by any of the methods described above, using standard and routine molecular biological manipulations well known to those of ordinary skill in the art. In one approach, a series of complementary oligonucleotide pairs of 80-90 nucleotides each

in length and spanning the length of the desired sequence is synthesized by standard methods. These oligonucleotide pairs are synthesized such that upon annealing, they form double stranded fragments of 80-90 base pairs, containing cohesive ends, *e.g.*, each oligonucleotide in the pair is synthesized to extend 3, 4, 5, 6, 7, 8, 9, 10, or more bases beyond the region that is complementary to the other oligonucleotide in the pair. The single-stranded ends of each pair of oligonucleotides is designed to anneal with the single-stranded end of another pair of oligonucleotides. The oligonucleotide pairs are allowed to anneal, and approximately five to six of these double-stranded fragments are then allowed to anneal together via the cohesive single stranded ends, and then they ligated together and cloned into a standard bacterial cloning vector, for example, a TOPO® vector available from Invitrogen Corporation, Carlsbad, CA. The construct is then sequenced by standard methods. Several of these constructs consisting of 5 to 6 fragments of 80 to 90 base pair fragments ligated together, *i.e.*, fragments of about 500 base pairs, are prepared, such that the entire desired sequence is represented in a series of plasmid constructs. The inserts of these plasmids are then cut with appropriate restriction enzymes and ligated together to form the final construct. The final construct is then cloned into a standard bacterial cloning vector, and sequenced. Additional methods would be immediately apparent to the skilled artisan. In addition, gene synthesis is readily available commercially.

[0162] In additional embodiments, a full-length polypeptide sequence is codon-optimized for a given species resulting in a codon-optimized coding region encoding the entire polypeptide, and then nucleic acid fragments of the codon-optimized coding region, which encode fragments, variants, and derivatives of the polypeptide are made from the original codon-optimized coding region. As would be well understood by those of ordinary skill in the art, if codons have been randomly assigned to the full-length coding region based on their frequency of use in a given species, nucleic acid fragments encoding fragments, variants, and derivatives would not necessarily be fully codon optimized for the given species. However, such sequences are still much closer to the codon usage of the desired species than the native codon usage. The advantage of this approach is that synthesizing codon-optimized nucleic acid fragments encoding each fragment, variant, and derivative of a given polypeptide, although routine would be time consuming and would result in significant expense.

[0163] The codon-optimized coding regions can be, for example, versions encoding a xylose or arabinose metabolizing enzymes of the invention, or domains, fragments, variants, or derivatives thereof.

[0164] Codon optimization is carried out for a particular species by methods described herein, for example, in certain embodiments, codon-optimized coding regions encoding polypeptides disclosed in the present application or domains, fragments, variants, or derivatives thereof are optimized according to codon usage in yeast (*e.g. Saccharomyces cerevisiae*). In certain embodiments described herein, a codon-optimized coding region encoding any of SEQ ID NOs: 1-5 or domain, fragment, variant, or derivative thereof, is optimized according to codon usage in yeast (*e.g. Saccharomyces cerevisiae*). In some embodiments, the sequences are codon-optimized specifically for expression in *Saccharomyces cerevisiae*. Alternatively, a codon-optimized coding region encoding any of SEQ ID NOs: 1-5 may be optimized according to codon usage in any plant, animal, or microbial species.

[0165] Also provided are polynucleotides, vectors, and other expression constructs comprising codon-optimized coding regions encoding polypeptides disclosed herein, or domains, fragments, variants, or derivatives thereof, and various methods of using such polynucleotides, vectors and other expression constructs.

Vectors and Methods of Using Vectors in Host Cells

[0166] In another aspect, the present invention relates to vectors which include polynucleotides of the present invention, host cells which are genetically engineered with vectors of the invention and the production of polypeptides of the invention by recombinant techniques.

[0167] Host cells are genetically engineered (transduced or transformed or transfected) with the vectors of this invention which may be, for example, a cloning vector or an expression vector. The vector may be, for example, in the form of a plasmid, a viral particle, a phage, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants or amplifying the genes of the present invention. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to the ordinarily skilled artisan.

[0168] The polynucleotides of the present invention can be employed for producing polypeptides by recombinant techniques. Thus, for example, the polynucleotide may be included in any one of a variety of expression vectors for expressing a polypeptide. Such vectors include chromosomal, nonchromosomal and synthetic DNA sequences, *e.g.*, derivatives of SV40; bacterial plasmids; and yeast plasmids. However, any other vector may be used as long as it is replicable and viable in the host.

[0169] The appropriate DNA sequence can be inserted into the vector by a variety of procedures. In general, the DNA sequence is inserted into an appropriate restriction endonuclease site(s) by procedures known in the art. Such procedures and others are deemed to be within the scope of those skilled in the art.

[0170] The DNA sequence in the expression vector is operatively associated with an appropriate expression control sequence(s) (promoter) to direct mRNA synthesis. Any suitable promoter to drive gene expression in the host cells of the invention may be used.

[0171] In addition, the expression vectors may contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as *URA3*, *HIS3*, *LEU2*, *TRP1*, *LYS2* or *ADE2,* dihydrofolate reductase, neomycin (G418) resistance or zeocin resistance for eukaryotic cell culture, or tetracycline or ampicillin resistance in prokaryotic cell culture, *e.g.*, *Clostridium thermocellum.*

[0172] The expression vector may also contain a ribosome binding site for translation initiation and/or a transcription terminator. The vector may also include appropriate sequences for amplifying expression, or may include additional regulatory regions.

[0173] The vector containing the appropriate DNA sequence as herein, as well as an appropriate promoter or control sequence, may be employed to transform an appropriate host to permit the host to express the protein.

[0174] Thus, in certain aspects, the present invention relates to host cells containing the above-described constructs. The host cell can be a host cell as described elsewhere in the application. The host cell can be, for example, a lower eukaryotic cell, such as a yeast cell, *e.g.*, *Saccharomyces cerevisiae* or *Kluyveromyces*, or the host cell can be a prokaryotic cell, such as a bacterial cell, *e.g.*, *Clostridium thermocellum.*

[0175] The selection of an appropriate host is deemed to be within the scope of those skilled in the art from the teachings herein. In one embodiment, the vector is integrated

into the genome of the host cell. In another embodiment, the vector is present in the host cell as an extrachromosomal plasmid.

[0176] To select for foreign DNA that has entered a host it is preferable that the DNA be stably maintained in the organism of interest. With regard to plasmids, there are two processes by which this can occur. One is through the use of replicative plasmids. These plasmids have origins of replication that are recognized by the host and allow the plasmids to replicate as stable, autonomous, extra chromosomal elements that are partitioned during cell division into daughter cells. The second process occurs through the integration of a plasmid onto the chromosome. This predominately happens by homologous recombination and results in the insertion of the entire plasmid, or parts of the plasmid, into the host chromosome. Thus, the plasmid and selectable marker(s) are replicated as an integral piece of the chromosome and segregated into daughter cells. Therefore, to ascertain if plasmid DNA is entering a cell during a transformation event through the use of selectable markers requires the use of a replicative plasmid or the ability to recombine the plasmid onto the chromosome. These qualifiers cannot always be met, especially when handling organisms that do not have a suite of genetic tools.

[0177] One way to avoid issues regarding plasmid-associated markers is through the use of transposons. A transposon is a mobile DNA element, defined by mosaic DNA sequences that are recognized by enzymatic machinery referred to as a transposase. The function of the transposase is to randomly insert the transposon DNA into host or target DNA. A selectable marker can be cloned onto a transposon by standard genetic engineering. The resulting DNA fragment can be coupled to the transposase machinery in an *in vitro* reaction and the complex can be introduced into target cells by electroporation. Stable insertion of the marker onto the chromosome requires only the function of the transposase machinery and alleviates the need for homologous recombination or replicative plasmids.

[0178] The random nature associated with the integration of transposons has the added advantage of acting as a form of mutagenesis. Libraries can be created that comprise amalgamations of transposon mutants. These libraries can be used in screens or selections to produce mutants with desired phenotypes. For instance, a transposon library of a CBP organism could be screened for the ability to produce less ethanol, or more lactic acid and/or more acetate.

Methods of Using Host Cells to Produce Ethanol or Other Fermentation Products

[0179] Microorganisms produce a diverse array of fermentation products, including organic acids, such as lactate (the salt form of lactic acid), acetate (the salt form of acetic acid), pyruvate, succinate, and butyrate, and neutral products, such as ethanol, butanol, acetone, and butanediol. End products of fermentation share to varying degrees several fundamental features, including: they are relatively nontoxic under the conditions in which they are initially produced, but become more toxic upon accumulation.

[0180] In one aspect, the present invention is directed to use of host cells and co-cultures to produce ethanol or other products from the xylose and/or the arabinose portion of lignocellulosic substrates. Such methods can be accomplished, for example, by contacting a pentose-containing lignocellulosic substrate with a host cell or a co-culture of the present invention. Fermentation products include, but are not limited to products such as ethanol, propanol, isoamyl alcohol, butanol, acetate, amino acids, and vitamins.

[0181] In one embodiment, the end products of pentose fermentation by the host strain comprise pyruvate, acetate, and ethanol. In another embodiment, the end products of pentose fermentation by the host strain comprise acetate, and ethanol. In one aspect, the ratio of acetate to ethanol formed can be at least about 10:1, at least about 5:1, at least about 2:1, at least about 1:1, at least about 1:2, at least about 1:5, or at least about 1:10. In one embodiment, the host cell is further engineered in order to increase ethanol production from pentose fermentation by the host cell. In one embodiment, the PTA gene is deleted in order to increase ethanol production from pentose fermentation by the host cell. In one aspect, the deletion of the PTA gene results in ethanol being the major end product of xylose fermentation by the host cell. In another aspect, the deletion of the PTA gene results in ethanol being the only end product of pentose fermentation by the host cell.

[0182] The production of ethanol can, according to the present invention, be performed at temperatures of at least about 25° C, about 28 ° C, about 30° C, about 31° C, about 32° C, about 33° C, about 34° C, about 35° C, about 36° C, about 37° C, about 38° C, about 39° C, about 40° C, about 41° C, about 42° C, or about 50° C. In some embodiments of the present invention, the thermotolerant host cell can produce ethanol from an arabinose-containing cellulosic substrate at temperatures above about 30° C, about 31° C, about 32° C, about 33° C, about 34° C, about 35° C, about 36° C, about 37° C, about 38° C, about

39° C, about 40° C, about 41° C, about 42° C, or about 50° C. In some embodiments of the present invention, the thermotolerant host cell can produce ethanol from an arabinose-containing cellulosic substrate at temperatures from about 30° C to 60° C, about 30° C to 55° C, about 30° C to 50° C, about 40° C to 60° C, about 40° C to 55° C or about 40° C to 50° C.

[0183] In some embodiments, methods of producing ethanol can comprise contacting an arabinose-containing lignocellulosic substrate with a host cell or co-culture of the invention and additionally contacting the arabinose-containing lignocellulosic substrate with externally produced arabinose-metabolizing enzymes. Exemplary externally produced arabinose metabolizing enzymes are commercially available and are known to those of skill in the art.

[0184] The invention is also directed to methods of reducing the amount of externally produced arabinose-metabolizing enzymes required to produce a given amount of ethanol from a arabinose-containing cellulosic substrate comprising contacting the arabinose-containing cellulosic substrate with externally produced arabinose-metabolizing enzymes and with a host cell or co-culture of the invention. In some embodiments, the same amount of ethanol production can be achieved using at least about 5%, 10%, 15%, 20%, 25%, 30%, or 50% fewer externally produced arabinose-metabolizing enzymes. In other embodiments, ethanol production can be achieved without the addition of externally produced arabinose-metabolizing enzymes.

[0185] In some embodiments, the methods comprise producing ethanol at a particular rate. For example, in some embodiments, ethanol is produced at a rate of at least about 0.1 mg per hour per liter, at least about 0.25 mg per hour per liter, at least about 0.5 mg per hour per liter, at least about 0.75 mg per hour per liter, at least about 1.0 mg per hour per liter, at least about 2.0 mg per hour per liter, at least about 5.0 mg per hour per liter, at least about 10 mg per hour per liter, at least about 15 mg per hour per liter, at least about 20.0 mg per hour per liter, at least about 25 mg per hour per liter, at least about 30 mg per hour per liter, at least about 50 mg per hour per liter, at least about 100 mg per hour per liter, at least about 200 mg per hour per liter, or at least about 500 mg per hour per liter.

[0186] In some embodiments, the host cells of the present invention can produce ethanol at a rate of at least about 0.1 mg per hour per liter, at least about 0.25 mg per hour per liter, at least about 0.5 mg per hour per liter, at least about 0.75 mg per hour per liter, at

least about 1.0 mg per hour per liter, at least about 2.0 mg per hour per liter, at least about 5.0 mg per hour per liter, at least about 10 mg per hour per liter, at least about 15 mg per hour per liter, at least about 20.0 mg per hour per liter, at least about 25 mg per hour per liter, at least about 30 mg per hour per liter, at least about 50 mg per hour per liter, at least about 100 mg per hour per liter, at least about 200 mg per hour per liter, or at least about 500 mg per hour per liter more than a control strain (lacking heterologous arabinose-metabolizing enzymes) and grown under the same conditions. In some embodiments, the ethanol can be produced in the absence of any externally added arabinose-metabolizing enzymes.

[0187] In some embodiments, a recombinant eukaryotic host cell of the invention produces an ethanol yield of at least about 5, at least about 7, at least about 10, at least about 13, at least about 15, or at least about 20 g/l ethanol after 24 hours of fermentation from a medium containing 20 g/l xylose and 21 g/l arabinose. In some embodiments, a recombinant eukaryotic host cell of the invention produces an ethanol yield of at least about 10, at least about 13, at least about 15, or at least about 20 g/l ethanol after 24 hours of fermentation from a medium containing 20 g/l glucose and 21 g/l arabinose. In some embodiments, a recombinant eukaryotic host cell of the invention produces an ethanol yield of at least about 10, at least about 13, at least about 15, or at least about 20 g/l ethanol after 24 hours of fermentation from a medium containing 10 g/l glucose, 10 g/l xylose and 21 g/l arabinose.

[0188] In some embodiments, organisms of the invention produce between about 5 and about 50, about 10 and about 50, about 20 and about 50, about 30 and about 50 g/l of ethanol after 24 hours of fermentation on arabinose-containing feedstock. In some embodiments, organisms of the invention produce between about 10 and about 50, about 15 and about 50, about 20 and about 50, about 25 and about 50, about 30 and about 50, about 35 and about 50 g/l of ethanol after 24 hours of fermentation on arabinose-containing feedstock. In some embodiments, the arabinose-containing feedstock contains about 2%, about 5%, about 10%, about 15%, about 20%, about 50%, or about 100% arabinose.

[0189] In some embodiments, cells of the invention are able to take up at least about 2, at least about 3, at least about 4, at least about 5, at least about 6 or at least about 7 g/l of arabinose from the external environment to the intracellular space per 24 hours. In some

embodiments, cells of the invention are able to take up at least about 2, at least about 3, at least about 4, at least about 5, at least about 6 or at least about 7 g/l of arabinose from the external environment to the intracellular space per 72 hours.

[0190] In some embodiments, cells of the invention are able to take up at least about 0.1 nmol mg dry mass^{-1} min^{-1}, about 0.2 nmol mg dry mass^{-1} min^{-1}, about 0.3 nmol mg dry mass^{-1} min^{-1}, about 0.4 nmol mg dry mass^{-1} min^{-1}, about 0.6 nmol mg dry mass^{-1} min^{-1}, about 0.8 nmol mg dry mass^{-1} min^{-1}, about 1.0 nmol mg dry mass^{-1} min^{-1}, about 1.5 nmol mg dry mass^{-1} min^{-1}, about 2 nmol mg dry mass^{-1} min^{-1}, about 3 nmol mg dry mass^{-1} min^{-1}, about 5 nmol mg dry mass^{-1} min^{-1}, about 7 nmol mg dry mass^{-1} min^{-1}, or at least about 10 nmol mg dry mass^{-1} min^{-1} of arabinose.

[0191] Ethanol production can be measured using any method known in the art. For example, the quantity of ethanol in fermentation samples can be assessed using HPLC analysis. Many ethanol assay kits are commercially available that use, for example, alcohol oxidase enzyme based assays. Methods of determining ethanol production are within the scope of those skilled in the art from the teachings herein.

[0192] The U.S. Department of Energy (DOE) provides a method for calculating theoretical ethanol yield. Accordingly, if the weight percentages are known of C6 sugars (*i.e.*, glucan, galactan, mannan), the theoretical yield of ethanol in gallons per dry ton of total C6 polymers can be determined by applying a conversion factor as follows:

(1.11 pounds of C6 sugar/pound of polymeric sugar) x (0.51 pounds of ethanol/pound of sugar) x (2000 pounds of ethanol/ton of C6 polymeric sugar) x (1 gallon of ethanol/6.55 pounds of ethanol) x (1/100%), *wherein the factor* (1 gallon of ethanol/6.55 pounds of ethanol) *is taken as the specific gravity of ethanol at 20°C.*

[0193] And if the weight percentages are known of C5 sugars (*i.e.*, xylan, arabinan), the theoretical yield of ethanol in gallons per dry ton of total C5 polymers can be determined by applying a conversion factor as follows:

(1.136 pounds of C5 sugar/pound of C5 polymeric sugar) x (0.51 pounds of ethanol/pound of sugar) x (2000 pounds of ethanol/ton of C5 polymeric sugar) x (1 gallon of ethanol/6.55 pounds of ethanol) x (1/100%), *wherein the factor* (1 gallon of ethanol/6.55 pounds of ethanol) *is taken as the specific gravity of ethanol at 20°C.*

[0194] It follows that by adding the theoretical yield of ethanol in gallons per dry ton of the total C6 polymers to the theoretical yield of ethanol in gallons per dry ton of the total

C5 polymers gives the total theoretical yield of ethanol in gallons per dry ton of feedstock.

Examples

Identification of Potential Arabinose Transporters

[0195] There are seven putative Arabinose transporters in the public and patent literature. The sequences of these seven proteins were aligned and determined to have only 10% identity with each other (Figure 1). This indicates arabinose transport function may be encoded by a divergent set of enzymes. To identify potential arabinose transporters several of these published transporters were blasted against the fungal protein database. The phylogenetic relationship of the top 7-10 hits from these blasts is shown in Figure 2. From this analysis, 15 enzymes were cloned into expression vectors and transformed into *S. cerevisiae*.

Arabinose Uptake Assay

[0196] To assay arabinose uptake an HPLC based assay was developed. Plasmid containing strains expressing arabinose transporters were grown overnight in YNB-uracil to maintain selection of the plasmid. These overnight cultures were subsequently diluted to normalize the optical density to OD1. From these, dilutions, 1ML of culture was spun down and washed 2X in distilled H20. Following these washes, 50ul of assay buffer (20g/l arabinose and 10g/l glucose dissolved in H_20) was added. After 72 hrs, 10ul aliquots of these reactions were prepped for HPLC with dilution into 85ul H_20 and 5ul 10% sulfuric acid. Results of this assay are shown in Figure 3. Most of the transporters selected appeared to function with the exception of the protein selected from *Z. rouxii*. The protein from *K. thermotolerans* appears to function better than the three other transporters tested.

Strain Construction and Fermentation Analysis

[0197] The base *S. cerevisiae* strain used to demonstrate arabinose utilization was M2874 which contains upregulation -by containing multiple copies- of the pentose phosphate pathway genes (*TAL1*, *TKL1* and *RKI1*), a deletion of *GRE3* and multiple copies of XI

and XKS. This strain has been shown to efficiently convert xylose to ethanol. To engineer M2874 with the arabinose pathway, an integration into *S. cerevisiae* rDNA sites was designed. A schematic illustrating the genomic integration design strategy known as arabinose assembly 1(AA1) is indicated in Figure 4. To create this assembly, purified PCR products amplified using primers listed in Table 3 were generated (Figure 5) and transformed into M2874 (M2874+ara). As a control, a separate transformation was performed without addition of AA1 amplicons (M2874-noDNA). Both transformations were plated on YMA (yeast nitrogen base plus 20 g/l arabinose). After 48 hours of incubation several hundred colonies were clearly visible on the plates containing M2874+ara and zero colonies were observed on the no-DNA control (Figure 6).

[0198] To confirm M2874+ara strains were able to convert arabinose into ethanol, several fermentations were conducted using either M2874 or a single colony isolated from the M2874+ara plate. Fermentations were run in 50 mls of medium which had been added to sealed aerobic pressure bottles.

Table 3. Primers used to amplify components of the arabinose-utilization construct.

promoter/gene	Primer combination	Template 1	Template 2	Template 3
ADHp / PDCt	X16757 /X17758	pMU2712 (araA)	pMU2713 (araB)	pMU2714 (araD)
HXT7p / PMA1t	X16759 / X16760	pMU2715 (araA)	pMU2716 (araB)	pMU2717 (araD)
TPIp / FBAt	X16761 / X16762	pMU2718 (araA)	pMU2719 (araB)	pMU2720 (araD)
ENOp / ENOt	X16763/ X16764	pMU3053	pMU3118	
rDNA 5' flank	X13185 / X13186	M2390 genomic DNA		
rDNA 3' flank	X13187 / X13188	M2390 genomic DNA		

[0199] To create pMU2712, pMU2713 and pMU2714 the *B. thetaiotamicron* araA, araB and araD were cloned into a vector containing the *S. cerevisiae* ADH1 promoter and PDC1 terminator. To create pMU2715, pMU2716, and pMU2717 the *B. thetaiotamicron* araA, araB and araD were cloned into a vector containing the *S. cerevisiae* HXT7p promoter and PMA1 terminator. To create pMU2718, pMU2719, and pMU2720 the *B. thetaiotamicron* araA, araB and araD were cloned into a vector containing the *S.*

cerevisiae TPI1 promoter and FBA1 terminator. To create pMU3053 and pMU3118, the *K. lactis* and *K. thermotolerans* AraT were cloned into a vector containing the ENO1 promoter and the ENO1 terminator.

Growth on Arabinose as the Sole Carbon Source (YP-21 g/l Arabinose)

[0200] The ability of the stain M2874 to convert arabinose into ethanol via fermentation was tested. ~9 g/l of arabinose was converted into ~4.0 g/l of ethanol in 48 hours. The parent strain is unable to consume any arabinose and no ethanol accumulation was observed.

Fermentation Analysis

[0201] To confirm M2874+ara strains were able to convert Arabinose into ethanol, several fermentations were conducted using either M2874 or a single colony isolated from the M2874+ara plate. M2874 and M2874 + ara were grown up overnight to prepare the inoculums. Each strain was adjusted to an OD of 1.0 of which 100ul was added to fermentation medium. Fermentations were run in sealed aerobic pressure bottles containing 50 mls of YP-medium containing either arabinose or a mixture of arabinose and xylose and/or glucose.

Growth on Arabinose as the Sole Carbon Source

[0202] Figure 7 shows that M2874 is able to convert ~9 g/l of arabinose into ~4.0 g/l of ethanol in 48 hours, giving a yield of 0.44 g/ethanol per g/arabinose consumed in a medium containing YP-21 g/l arabinose. The parent strain was unable to consume any arabinose and no ethanol accumulation was observed.

Growth on Arabinose and Xylose as the Sole Carbon Sources

[0203] Figure 8 shows that both M2874 and M2874+ara are able to consume all 20 grams of xylose by 48 hours when grown on YP-20g/l xylose/21 g/l arabinose. However, only M2874+ara is able to use arabinose, resulting in an extra 3.5 g/l ethanol from the ~7 g/l which was consumed, yielding ~ 0.5 g ethanol per gram of arabinose consumed.

Growth on Arabinose and Glucose

[0204] Figure 9 shows that both M2874 and M2874+ara are able to consume all 20g/l glucose by 48 hours when grown on YP-20g/l glucose/21 g/l arabinose media. However, only M2874+ara was able to use arabinose, resulting in an extra 6.5 g/l ethanol from the ~13 g/l arabinose consumed, which yielded 0.5 g ethanol per gram of arabinose consumed.

Arabinose Consumption in Yeast Containing *Piromyces* XI

[0205] Yeast expressing a *Pyromyces sp.* xylose isomerase were transformed with the arabinose utilization construct of the invention. Transformants were selected on YNB+arabinose plates. Two single colony isolates were tested on various media, with the untransformed parental strain used as a control. The media tested included: 20% washate (pH 6) with 20 g/l arabinose; YPA (20 g/l arabinose); YPAX (20 g/l arabinose, 20 g/l xylose); and, YPAXD (20 g/l arabinose, 10 g/l xylose, 10 g/l glucose).

[0206] The yeast strains were pregrown on YPX, washed, and used to inoculate 150 ml sealed bottles with 25 ml medium, which were flushed with N_2 and incubated at 35°C at 250 RPM. Cultures were sampled for HPLC at 0, 24, 48 and 72 hours. No growth was observed on YPA. 3.7 g l^{-1} arabinose was consumed for clone 1 in YPAXD. Over half of this (2.1 g l^{-1}) was consumed between 24 and 48 h, when xylose and glucose already had been depleted. Figures 11-13 depict the results of these assays.

Summary of Fermentation Data

[0207] The fermentation data depicted herein demonstrate that the arabinose assembly of the invention enables conversion of arabinose into ethanol by *S. cerevisiae*. The ethanol yields from arabinose are around 0.5 g/g in all fermentations in which combinations of sugars were tested. The ethanol yield was slightly lower when arabinose was used as the sole carbon source, which likely indicates an increase in biomass generated from consumption of arabinose.

Equivalents

[0208] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims. Additionally, all references cited herein are incorporated herein by reference as though they were reproduced herein in their entirety.

What is claimed is:

1. A recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (R5PE).

2. The recombinant eukaryotic host cell of claim 1, wherein the AraT is derived from an AraT of an organism selected from the group consisting of *Kluveromyces lactis, Kluveromyces thermotolerans, Zygosaccharomyces rouxii, Vanderwaltozyma polyspora, Debaryomyces hansenii, Aspergillus niger, Penicillium chrysogenum, Pichia guilermondii, Aspergillus flavus, Candida lusitnaea, Candida albicans (SC5314), Kluveromyces marxianus, Pichia stipites,* and *Candida arbinofermentans.*

3. The recombinant eukaryotic host cell of either claim 1 or 2, wherein the AraT comprises an amino acid sequence at least 80% identical to any one of the amino acid sequences of SEQ ID NOs: 9-22.

4. The recombinant eukaryotic host cell of any one of claims 1 to 3, wherein one or more of the AI, RK and R5PE is derived from an AI, RK and R5PE of *B. thetaiotamicron.*

5. A recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (R5PE), wherein one or more of the AI, RK and R5PE is derived from an AI, RK and R5PE of *B. thetaiotamicron.*

6. The recombinant eukaryotic host cell of either claim 4 or 5, wherein

a) the AI comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 6;

b) the RK comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 7; and

c) the R5PE comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 8.

7. The recombinant eukaryotic host cell of any one of claims 1 to 6, wherein expression of the heterologous polynucleotide confers an ability to ferment arabinose to the recombinant host cell.

8. The recombinant eukaryotic host cell of any one of claims 1 to 7 further comprising a heterologous polynucleotide encoding a xylose isomerase (XI).

9. The recombinant eukaryotic host cell of claim 8, wherein the XI is derived from an XI of *B. thetaiotamicron*.

10. The recombinant eukaryotic host cell of either of claims 8 or 9, wherein the XI comprises an amino acid sequence at least 80% identical to the amino acid sequence of SEQ ID NO: 24 or SEQ ID NO: 26.

11. The recombinant eukaryotic host cell of any one of claims 1-10, wherein the host cell is a yeast cell.

12. The recombinant eukaryotic host cell of claim 11, wherein the yeast cell is selected from the group consisting of *Saccharomyces cerevisiae*, *Schizzosaccharomyces pombe, Candida albicans, Pichia pastoris, Pichia stipitis, Yarrowia lipolytica, Hansenula polymorpha, Phaffia rhodozyma, Candida utilis, Arxula adeninivorans, Debaryomyces hansenii, Debaryomyces polymorphus, Schizosaccharomyces pombe* and *Schwanniomyces occidentalis*.

13. The recombinant eukaryotic host cell of either of claim 11 or 12, wherein the yeast cell comprises a heterologous sequence encoding a xylulokinase, ribulose 5-phosphate isomerase, ribulose 5-phophate epimerase, transketolase and transaldolase, and wherein the yeast cell does not express an aldose reductase that is capable of catalyzing the conversion of xylose to xylitol.

14. The recombinant eukaryotic host cell of any one of claims 1 to 13, wherein the host cell is capable of fermenting xylose, arabinose, or a combination thereof.

15. The recombinant eukaryotic host cell of any one of claims 1 to 14, wherein the host cell is capable of fermenting arabinose from a cellulosic substrate.

16. The recombinant eukaryotic host cell of either claim 14 or 15, wherein the fermentation product is selected from the group consisting of ethanol, lactic acid, hydrogen, butyric acid, acetone, and butanol.

17. The recombinant eukaryotic host cell of any one of the preceding claims, wherein the host cell is an industrial strain exhibiting high ethanol tolerance.

18. The recombinant eukaryotic host cell of claim 17, wherein the host cell further exhibits high temperature tolerance.

19. The recombinant eukaryotic host cell of any one of the preceding claims, wherein the host cell produces an ethanol yield of at least about 10 g/l ethanol after 24 hours of fermentation from a medium containing 20 g/l xylose and 21 g/l arabinose.

20. The recombinant eukaryotic host cell of any one of the preceding claims, wherein the host cell produces an ethanol yield of at least about 13 g/l ethanol after 24 hours of fermentation from a medium containing 20 g/l glucose and 21 g/l arabinose.

21. The recombinant eukaryotic host cell of any one of the preceding claims, wherein the host cell produces an ethanol yield of at least about 15 g/l ethanol after 24 hours of fermentation from a medium containing 10 g/l glucose, 10 g/l xylose and 21 g/l arabinose.

22. The recombinant eukaryotic host cell of any one of the preceding claims, wherein the host cell further comprises one or more heterologous polynucleotides encoding a cellulase.

23. The recombinant eukaryotic host cell of claim 22, wherein the one or more cellulases is selected from the group consisting of endoglucanases, exoglucanases, and β-glucosidases.

24. The recombinant eukaryotic host cell of claim 23, wherein the host cell comprises: (a) a first heterologous polynucleotide that encodes an endoglucanase; (b) a second heterologous polynucleotide that encodes a β-glucosidase; (c) a third heterologous polynucleotide that encodes

a first cellobiohydrolase; and, (d) a fourth heterologous polynucleotide that encodes a second cellobiohydrolase.

25. The recombinant eukaryotic host cell of claim 24, wherein (a) the first heterologous polynucleotide that encodes an endoglucanase is derived from *A. fumigatus*; (b) the second heterologous polynucleotide that encodes a β-glucosidase is derived from *S. fibuligera*; (c) the third heterologous polynucleotide that encodes a first cellobiohydrolase is derived from *T. emersonii*; and, (d) the fourth heterologous polynucleotide that encodes a second cellobiohydrolase is derived from *C. lucknowense*.

26. A recombinant yeast cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), wherein the yeast cell is capable of taking up at least about 5 g/l of arabinose in 24 hours.

27. The recombinant yeast cell of claim 26, wherein the AraT is derived from an AraT of an organism selected from the group consisting of *Kluveromyces lactis, Kluveromyces thermotolerans,* and *Aspergillus niger.*

28. The recombinant yeast cell of either claim 26 or 27, wherein the AraT comprises an amino acid sequence at least 80% identical to any one of the amino acid sequences of SEQ ID NOs: 9-22.

29. The recombinant host cell of any one of the preceding claims, wherein at least one of the heterologous polynucleotides is integrated into the genome of the host cell.

30. A method of producing a fermentation product comprising:

a) combining the recombinant eukaryotic host cell of any one of claims 1 to 29 with a substrate;

b) allowing the host cell to ferment the substrate; and,

c) recovering one or more products of the fermentation,

wherein the substrate is a cellulosic substrate and the yield of fermentation product is increased by virtue of the host cell's ability to ferment arabinose.

31. A composition comprising a carbon source and the recombinant eukaryotic host cell of any one of claims 1 to 29, wherein the carbon source is a cellulosic substrate that contains at least about 1% arabinose.

32. A media supernatant generated by incubating the recombinant eukaryotic host cell of any one of claims 1 to 29 with a medium containing a carbon source that contains at least about 1% arabinose.

33. A composition comprising at least a first and a second recombinant eukaryotic host cell according to any one of claims 1 to 29, wherein the first and second recombinant eukaryotic host cells are genetically different.

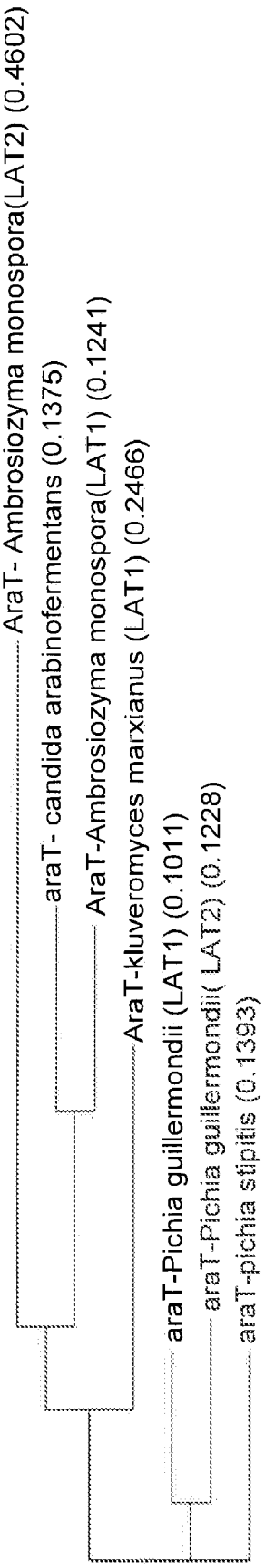
AraT- Ambrosiozyma monospora(LAT2) (0.4602)
araT- candida arabinofermentans (0.1375)
AraT-Ambrosiozyma monospora(LAT1) (0.1241)
AraT-kluveromyces marxianus (LAT1) (0.2466)
araT-Pichia guillermondii (LAT1) (0.1011)
araT-Pichia guillermondii(LAT2) (0.1228)
araT-pichia stipitis (0.1393)

FIG. 1

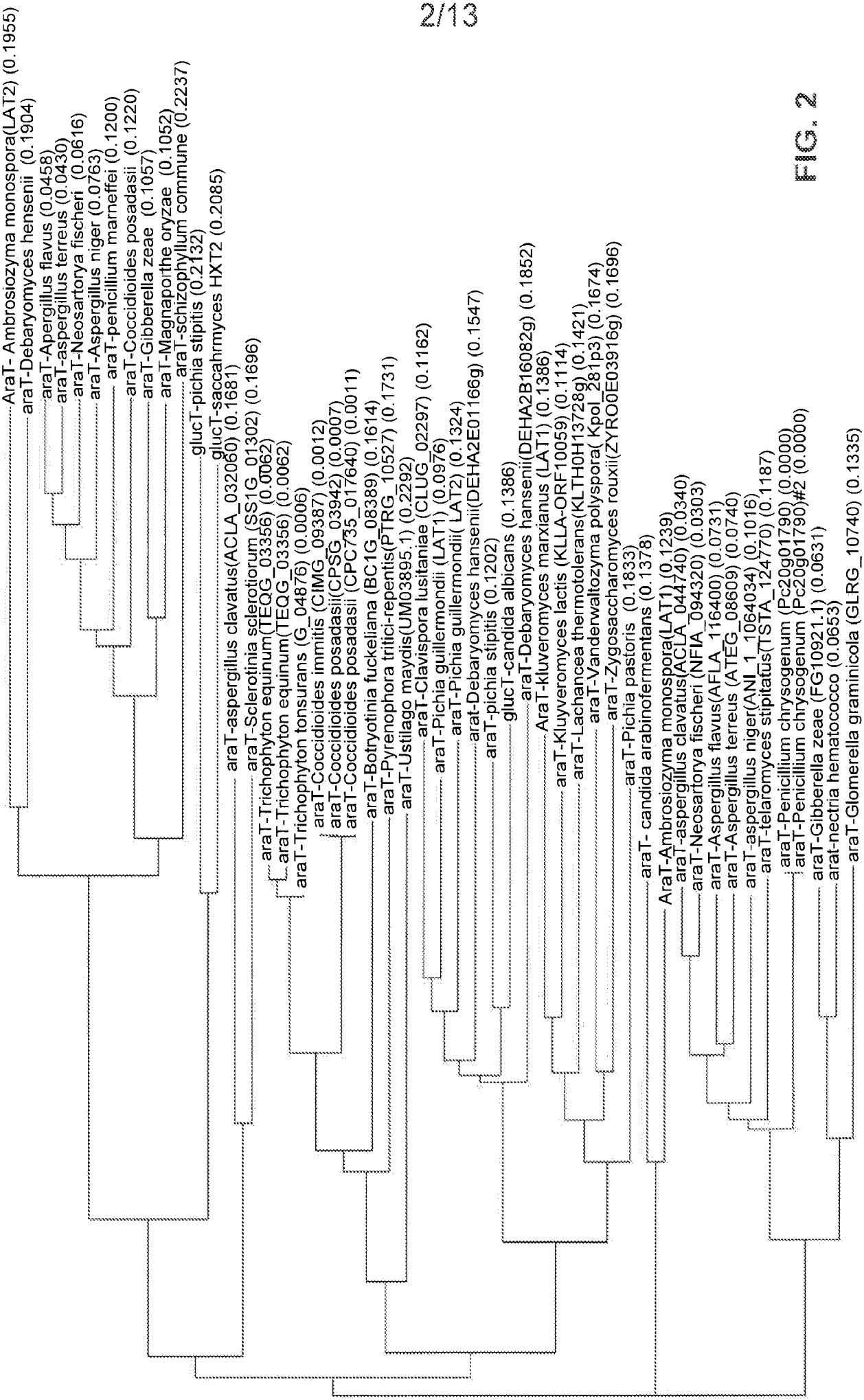
AraT- Ambrosiozyma monospora(LAT2) (0.1955)
araT-Debaryomyces hensenii (0.1904)
araT-Apergillus flavus (0.0458)
araT-aspergillus terreus (0.0430)
araT-Neosartorya fischeri (0.0616)
araT-Aspergillus niger (0.0763)
araT-penicillium marneffei (0.1200)
araT-Coccidioides posadasii (0.1220)
araT-Gibberella zeae (0.1057)
araT-Magnaporthe oryzae (0.1052)
araT-schizophyllum commune (0.2237)
glucT-pichia stipitis (0.2132)
glucT-saccahrmyces HXT2 (0.2085)
araT-aspergillus clavatus(ACLA_032060) (0.1681)
araT-Sclerotinia sclerotiorum (SS1G_01302) (0.1696)
araT-Trichophyton equinum(TEQG_03356) (0.0062)
araT-Trichophyton equinum(TEQG_03356) (0.0062)
araT-Trichophyton tonsurans (G_04876) (0.0006)
araT-Coccidioides immitis (CIMG_09387) (0.0012)
araT-Coccidioides posadasii(CPSG_03942) (0.0007)
araT-Coccidioides posadasii (CPC735_017640) (0.0011)
araT-Botryotinia fuckeliana (BC1G_08389) (0.1614)
araT-Pyrenophora tritici-repentis(PTRG_10527) (0.1731)
araT-Ustilago maydis(UM03895.1) (0.2292)
araT-Clavispora lusitaniae (CLUG_02297) (0.1162)
araT-Pichia guillermondii (LAT1) (0.0976)
araT-Pichia guillermondii(LAT2) (0.1324)
arat-Debaryomyces hansenii(DEHA2E01166g) (0.1547)
araT-pichia stipitis (0.1202)
glucT-candida albicans (0.1386)
araT-Debaryomyces hansenii(DEHA2B16082g) (0.1852)
AraT-kluveromyces marxianus (LAT1) (0.1386)
araT-Kluyveromyces lactis (KLLA-ORF10059) (0.1114)
araT-Lachancea thermotolerans(KLTH0H13728g) (0.1421)
araT-Vanderwaltozyma polyspora(Kpol_281p3) (0.1674)
araT-Zygosaccharomyces rouxii(ZYRO0E03916g) (0.1696)
araT-Pichia pastoris (0.1833)
araT- candida arabinofermentans (0.1378)
AraT-Ambrosiozyma monospora(LAT1) (0.1239)
araT-aspergillus clavatus(ACLA_044740) (0.0340)
araT-Neosartorya fischeri (NFIA_094320) (0.0303)
araT-Aspergillus flavus(AFLA_116400) (0.0731)
araT-Aspergillus terreus (ATEG_08609) (0.0740)
araT-aspergillus niger(ANI_1_1064034) (0.1016)
araT-telaromyces stipitatus(TSTA_124770) (0.1187)
araT-Penicillium chrysogenum (Pc20g01790) (0.0000)
araT-Penicillium chrysogenum (Pc20g01790)#2 (0.0000)
araT-Gibberella zeae (FG10921.1) (0.0631)
arat-nectria hematococco (0.0653)
araT-Glomerella graminicola (GLRG_10740) (0.1335)

FIG. 2

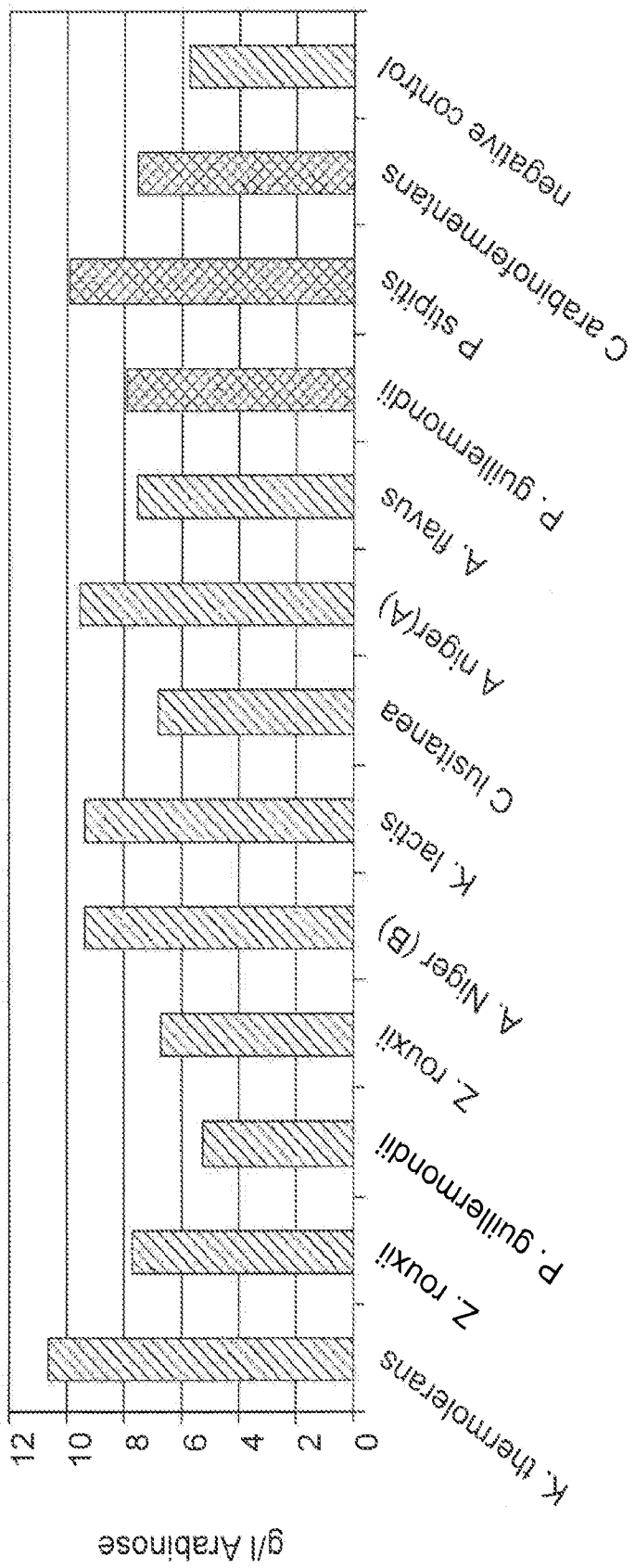
K. thermoterans
Z. rouxii
P. guilliermondii
Z. rouxii
A. Niger (B)
K. lactis
C lusitanea
A niger(A)
A. flavus
P. guilliermondii
P stipitis
C arabinofermentans
negative control
0
2
4
6
8
10
12
g/l Arabinose

FIG. 3

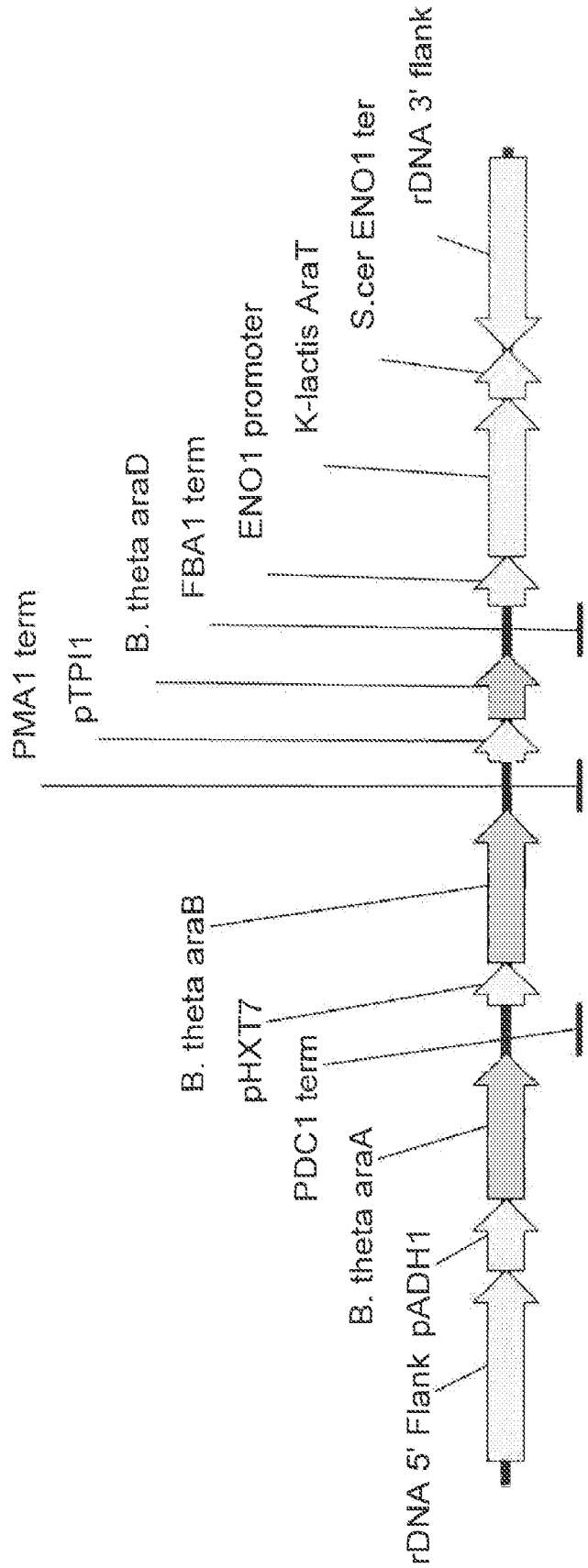
rDNA 5' Flank
pADH1
B. theta araA
PDC1 term
pHXT7
B. theta araB
PMA1 term
pTPI1
B. theta araD
FBA1 term
ENO1 promoter
K-lactis AraT
S.cer ENO1 ter
rDNA 3' flank
Arabinose gene assembly 1
14000 bp

FIG. 4

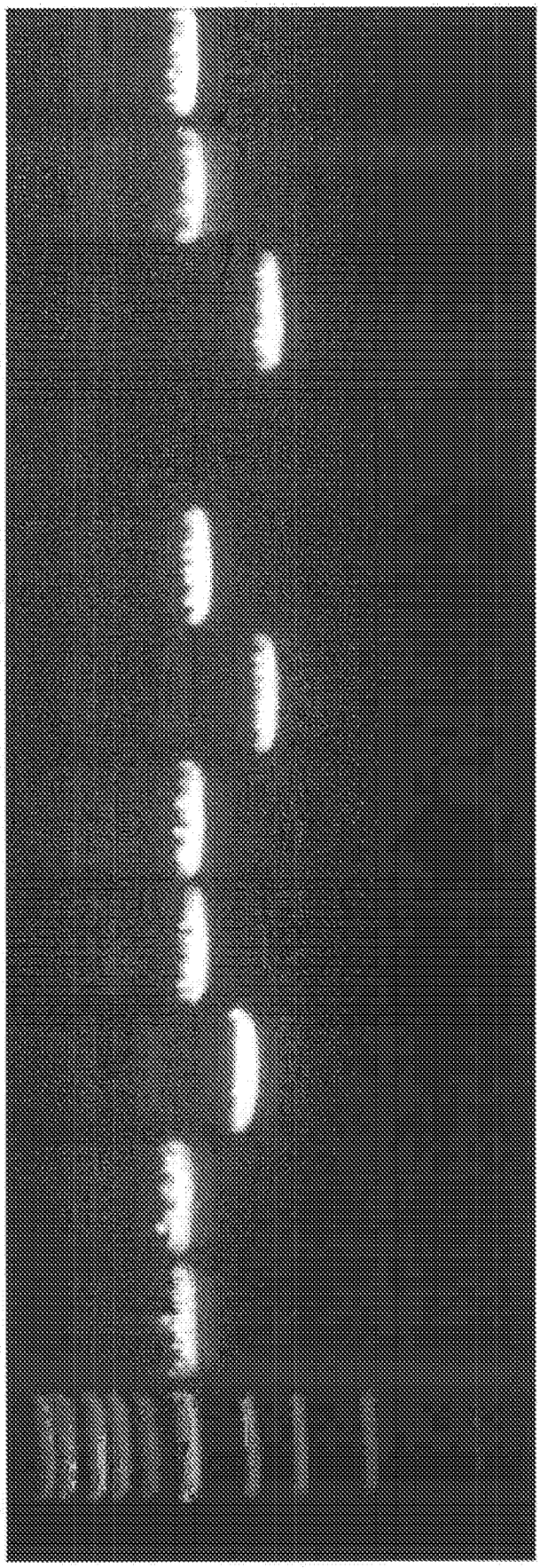

FIG. 5

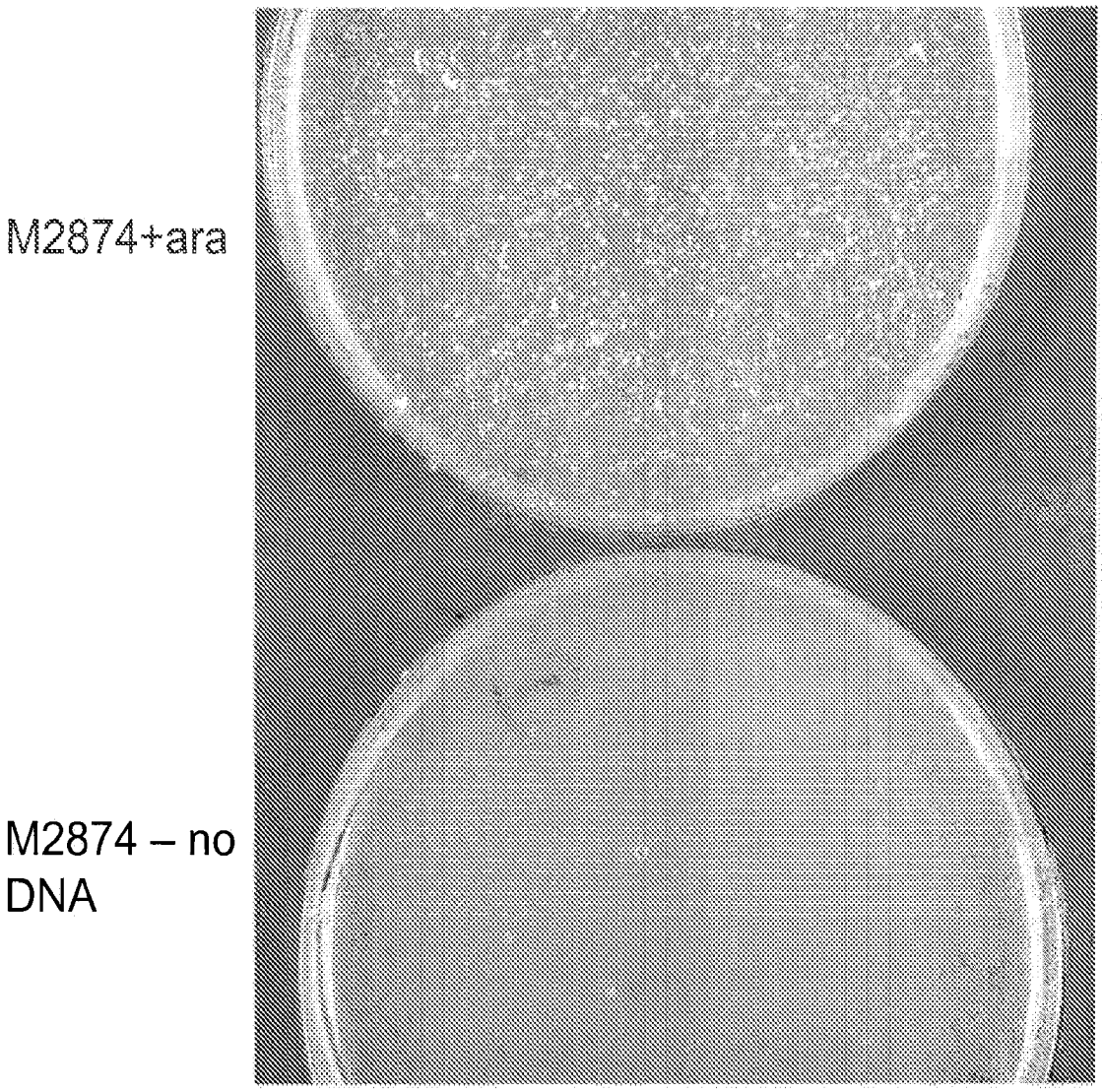

FIG. 6

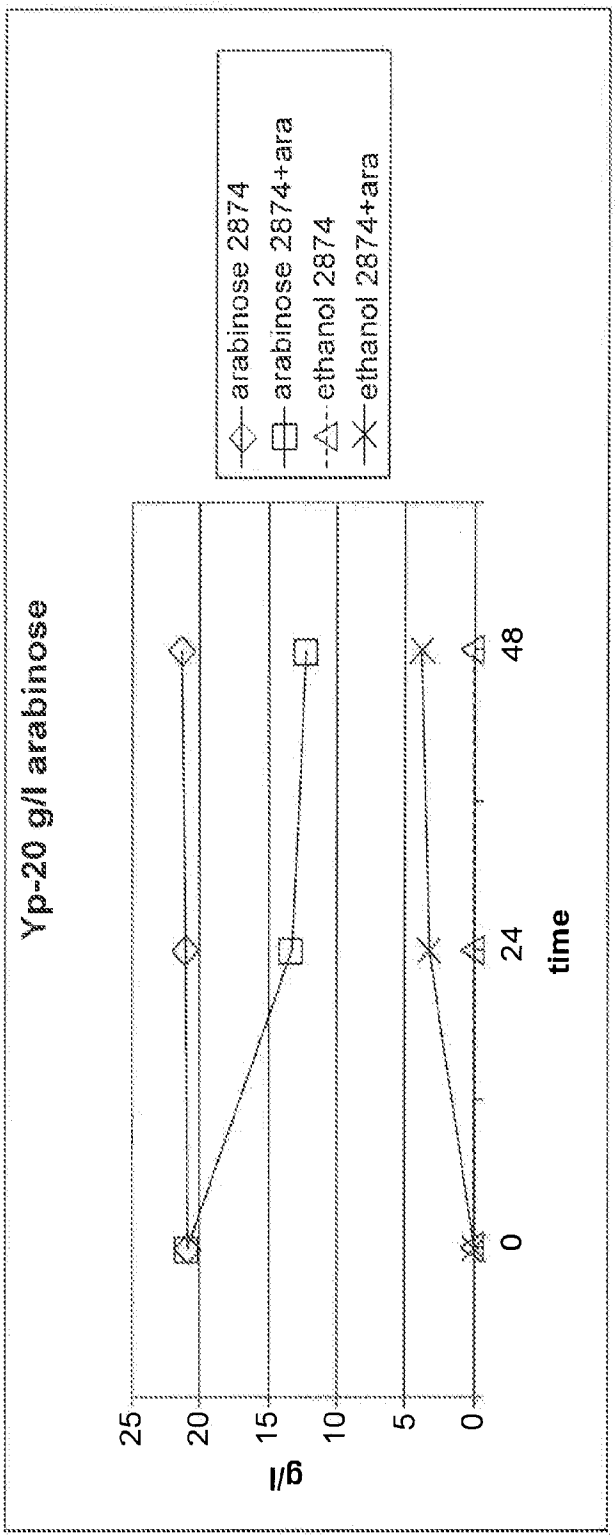
Yp-20 g/l arabinose
arabinose 2874
arabinose 2874+ara
ethanol 2874
ethanol 2874+ara
g/l
25
20
15
10
5
0
0
24
48
time

FIG. 7

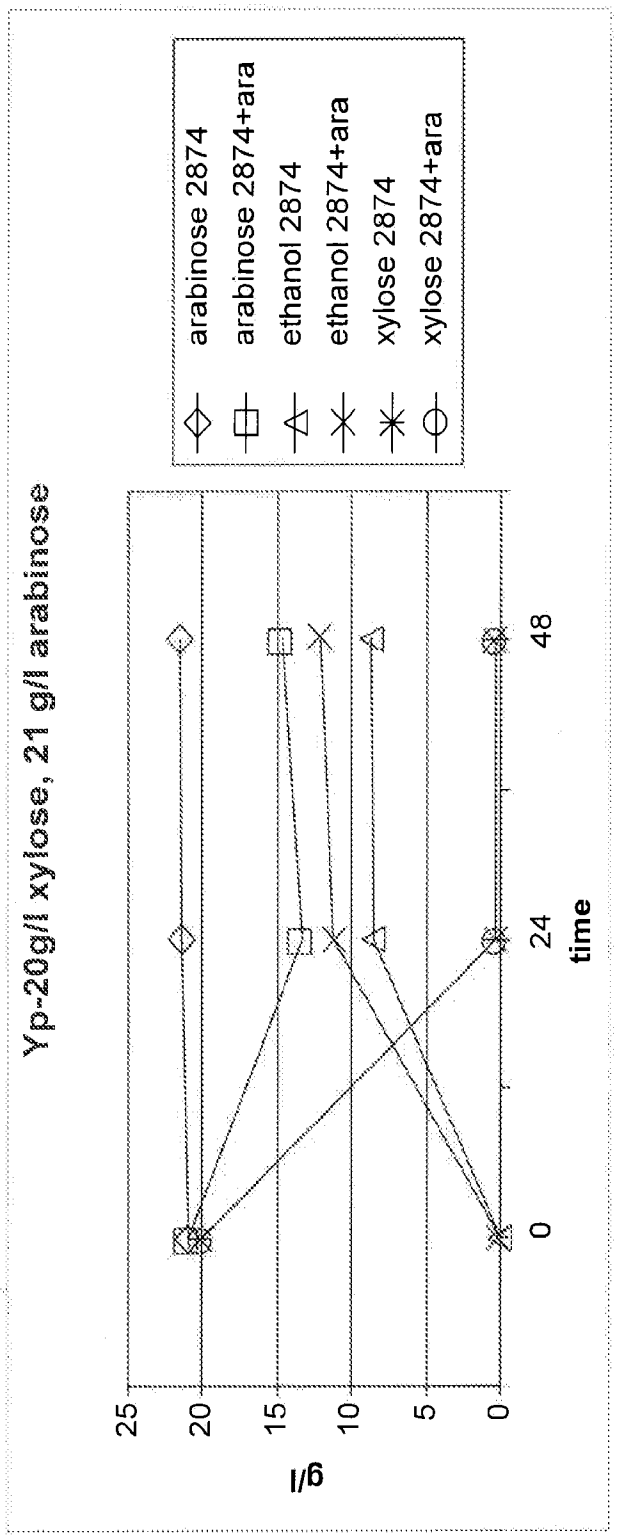
Yp-20g/l xylose, 21 g/l arabinose
arabinose 2874
arabinose 2874+ara
ethanol 2874
ethanol 2874+ara
xylose 2874
xylose 2874+ara
g/l
25
20
15
10
5
0
0
24
48
time

FIG. 8

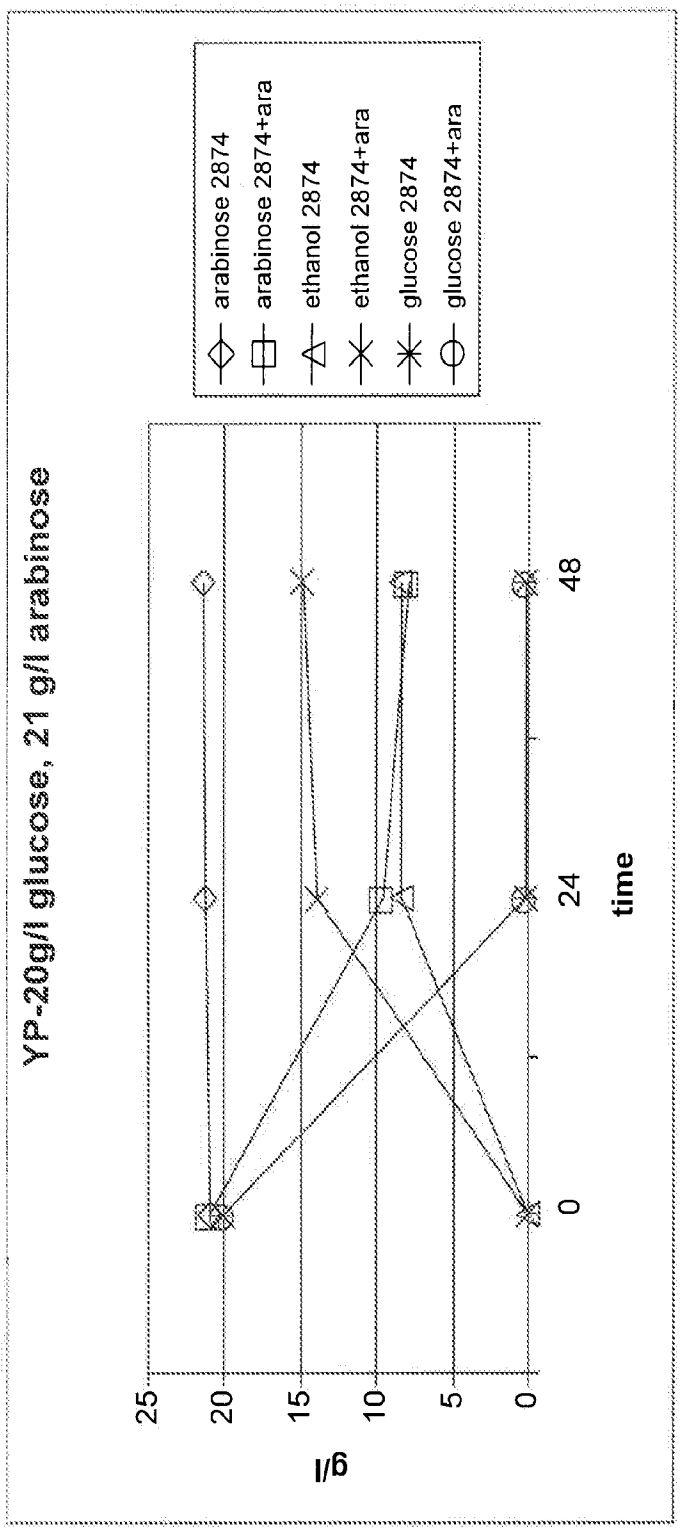
YP-20g/l glucose, 21 g/l arabinose
arabinose 2874
arabinose 2874+ara
ethanol 2874
ethanol 2874+ara
glucose 2874
glucose 2874+ara
g/l
25
20
15
10
5
0
0
24
48
time

FIG. 9

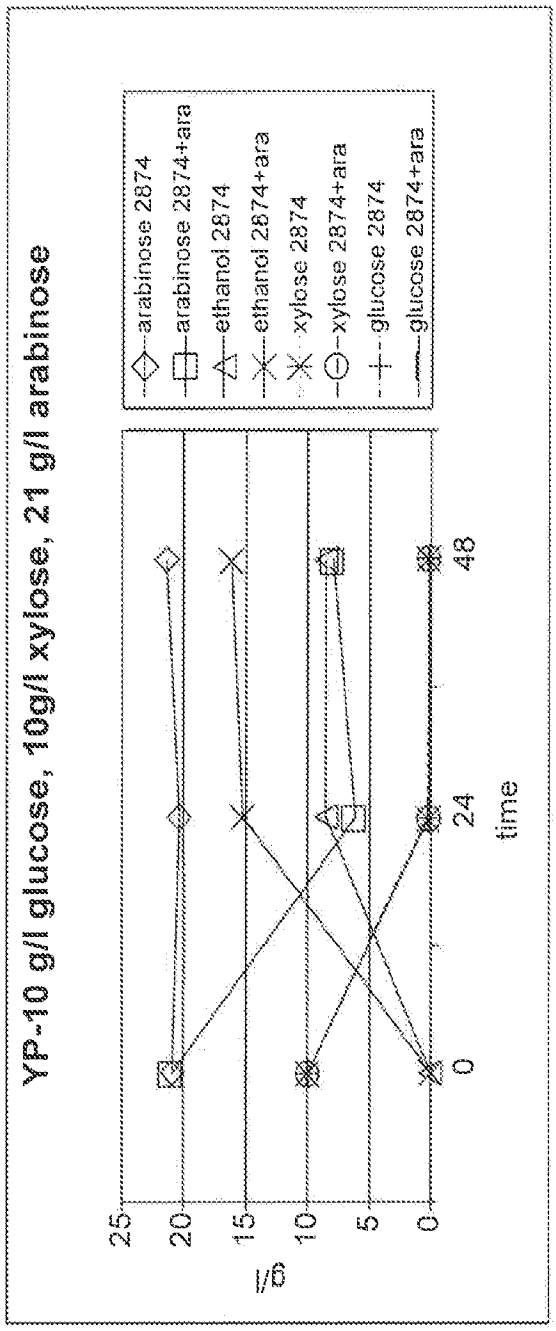
YP-10 g/l glucose, 10g/l xylose, 21 g/l arabinose
arabinose 2874
arabinose 2874+ara
ethanol 2874
ethanol 2874+ara
xylose 2874
xylose 2874+ara
glucose 2874
glucose 2874+ara
g/l
25
20
15
10
5
0
0
24
48
time

FIG. 10

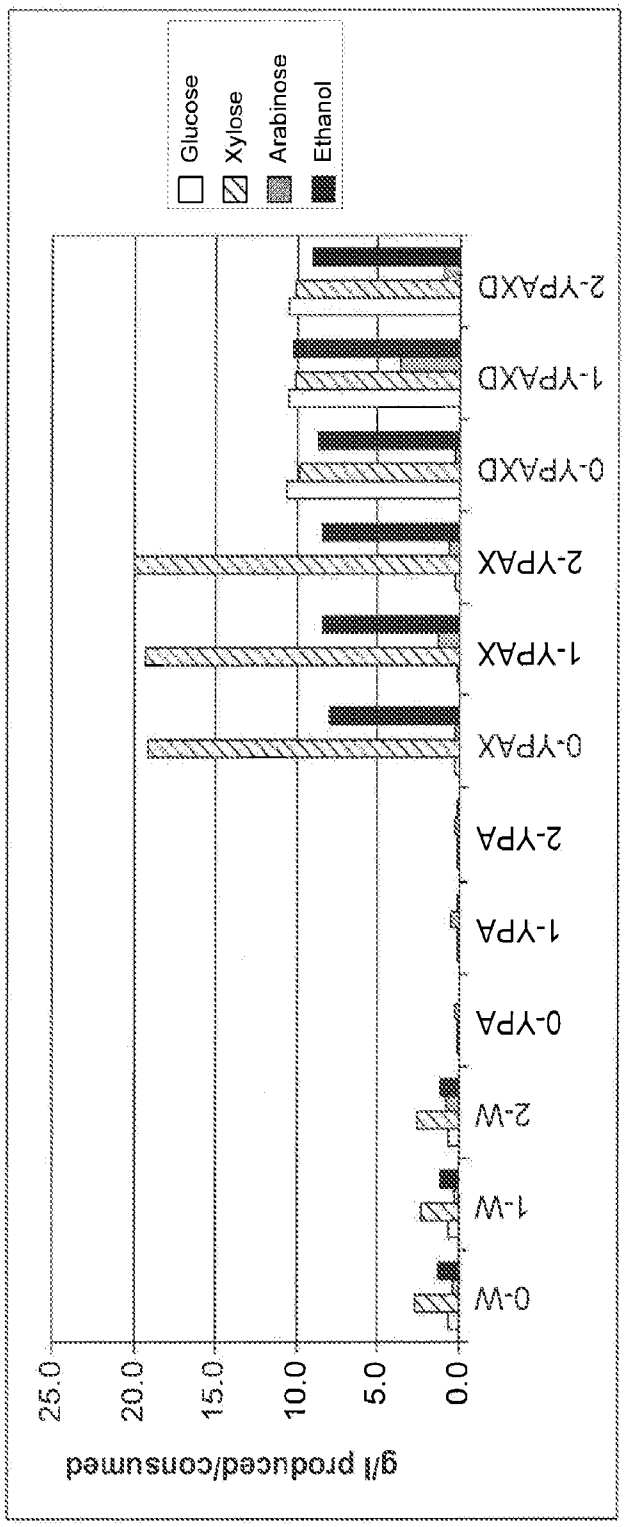
Glucose
Xylose
Arabinose
Ethanol
2-YPAXD
1-YPAXD
0-YPAXD
2-YPAX
1-YPAX
0-YPAX
2-YPA
1-YPA
0-YPA
2-W
1-W
0-W
25.0
20.0
15.0
10.0
5.0
0.0
g/l produced/consumed

FIG. 11

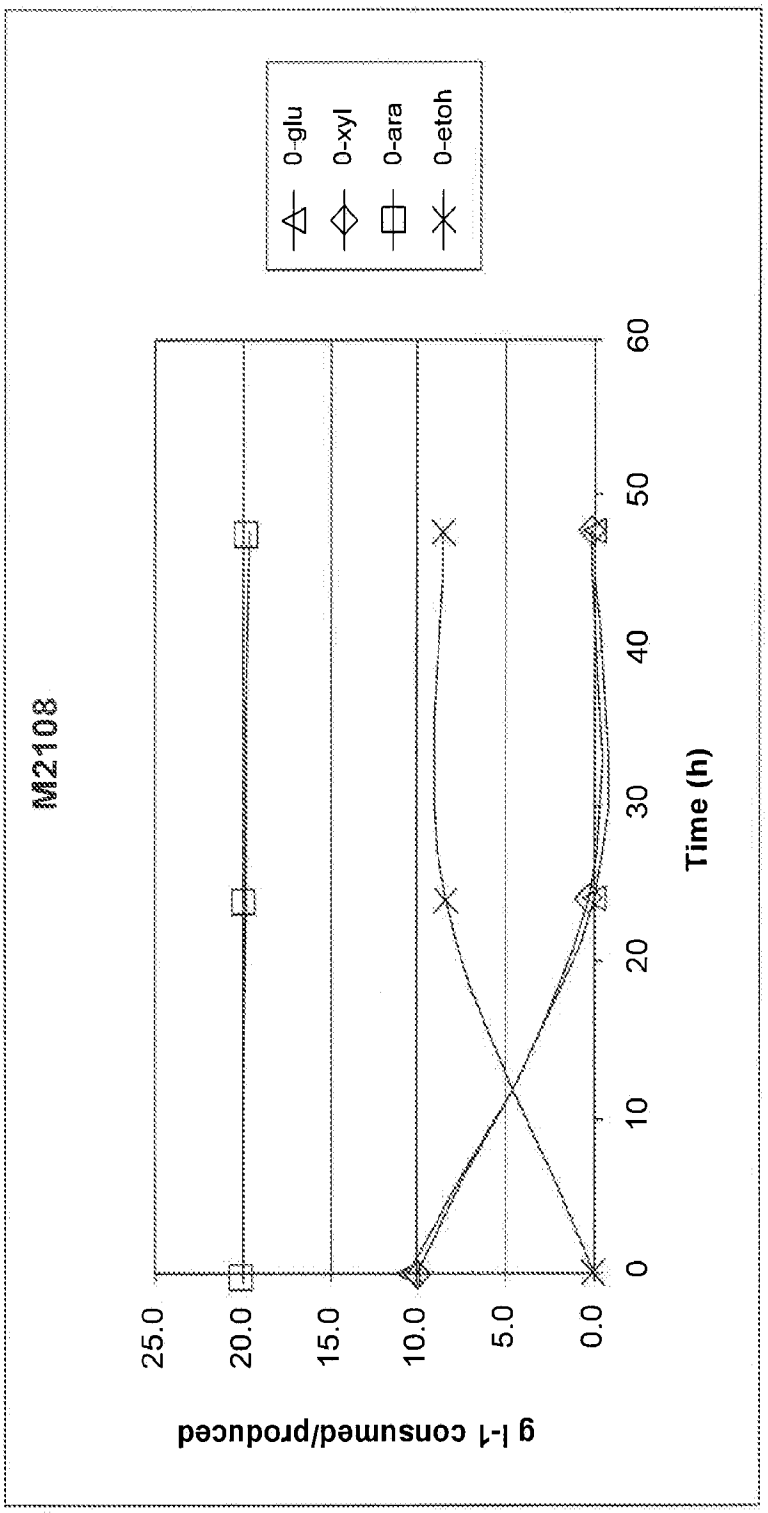
M2108
0-glu
0-xyl
0-ara
0-etoh
g l-1 consumed/produced
25.0
20.0
15.0
10.0
5.0
0.0
0
10
20
30
40
50
60
Time (h)

FIG. 12

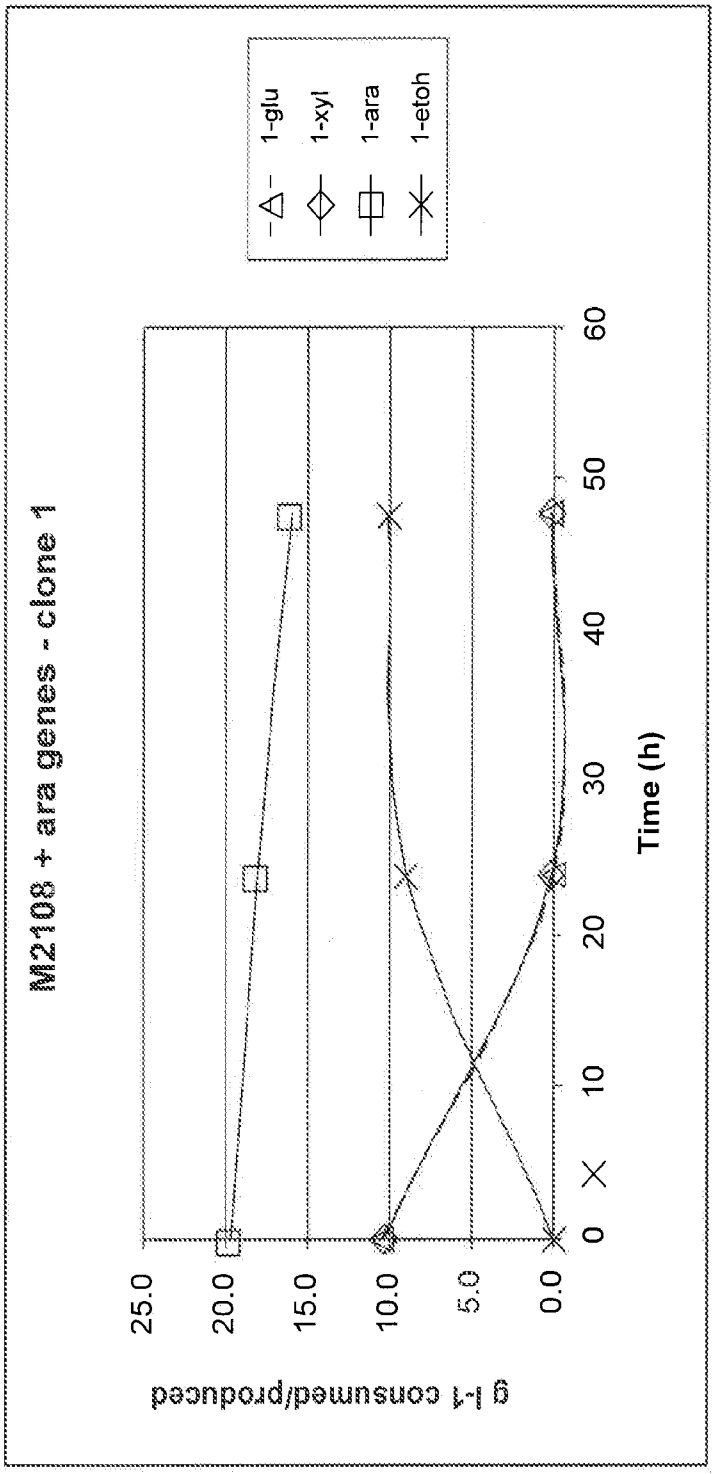
M2108 + ara genes - clone 1
1-glu
1-xyl
1-ara
1-etoh
g l-1 consumed/produced
25.0
20.0
15.0
10.0
5.0
0.0
0
10
20
30
40
50
60
Time (h)

FIG. 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 12/64457

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C07H 21/04; C12N 1/00; C12P 21/04 (2013.01)
USPC - 536/23.74
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 536/23.74

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 435/254.11, 435/69.9, 530/350 (keyword limited; terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (USPT, PGPB, EPAB, JPAB), Google Patents/Scholar: arabinose transporter, arabinose isomerase, ribulokinase, ribulose epimerase, yeast
GenCore 6.4: SEQ ID NO: 9

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y	US 2010/0143936 A1 (Boles, et al.) 10 June 2010 (10.06.2010) para [0010], [0020], [0026], [0028]	1-2 ---- 3
Y	US 2007/0118916 A1 (Puzio, et al.) 24 May 2007 (24.05.2007) para [5216], SEQ ID NO:9	3
A	US 2010/0304454 A1 (De Bont) 02 December 2010 (02.12.2010) whole doc.	1
A,P	US 2012/0129241 A1 (Zhang, et al.) 24 May 2012 (24.05.2012) whole doc.	1
A	WO 2009/008756 A2 (Da Fonseca, et al.) 15 January 2009 (15.01.2009) whole doc.	1
A	US 2010/0086965 A1 (Van Maris, et al.) 08 April 2010 (08.04.2010) whole doc.	1

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
06 March 2013 (06.03.2013)	15 MAR 2013
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 12/64457

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 4, 6-25, 29-33
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
Group I+: claims 1-3, drawn to a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose S-phosphate epimerase (RSPE). The first invention is restricted to SEQ ID NO: 9. Should an additional fee(s) be paid, Applicant is invited to elect an additional sequence(s) to be searched. The exact claims searched will depend on Applicant's election.

Group II: claims 5, drawn to a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (RSPE), wherein one or more of the AI, RK and RSPE is derived from an AI, RK and RSPE of B. thetaiotamicron.

- Please see extra sheet for Observations where unity of invention is lacking -

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-3 restricted to SEQ ID NO: 9

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

Continuation of:
Box NO III. Observations where unity of invention is lacking

Group III+: claims 26-28, drawn to a recombinant yeast cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), wherein the yeast cell is capable of taking up at least about 5 g/l of arabinose in 24 hours. The first invention is restricted to SEQ ID NO: 9. Should an additional fee(s) be paid, Applicant is invited to elect an additional sequence(s) to be searched. The exact claims searched will depend on Applicant's election.

The inventions listed as Groups I+ through III+ do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The inventions of Groups II do not include the inventive concept of a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), as required by Group I+ and III+.

The inventions of Groups III+ do not include the inventive concept of a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (RSPE), as required by Group I+ and II.

The inventions of Groups I+ and II share the technical feature of an arabinose isomerase (AI), a ribulokinase (RK) and a ribulose 5-phosphate epimerase (RSPE). The inventions of Groups I+ and III+ share the technical feature of an an arabinose transporter (AraT). However, this shared technical feature does not represent a contribution over prior art as being anticipated by US 2010/0143936 A1 to Boles et al. (hereinafter 'Boles'). Boles discloses Claim 1, a recombinant eukaryotic host cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT), a heterologous polynucleotide encoding an arabinose isomerase (AI), a heterologous polynucleotide encoding a ribulokinase (RK) and a heterologous polynucleotide encoding a ribulose 5-phosphate epimerase (RSPE) (para [0010], [0026] and [0028]). Further Boles teaches a recombinant yeast cell comprising a heterologous polynucleotide encoding an arabinose transporter (AraT) (para [0010] and [0021]), wherein the uptake rate of for L-arabinose in yeast cell is improved and substantially more efficient (para [0047] and [0153]), but does not specifically teach that the yeast cell is capable of taking up at least about 5 g/l of arabinose in 24 hours. Puzio et al. (US 2007/0118916 A1) teach a high-affinity L-arabinose transport protein (para [5216]) comprising SEQ ID NO:524 that is 100% identical to claimed SEQ ID NO:9. It would have been obvious to one of ordinary skill in the art to have applied the high-affinity L-arabinose transport protein of Puzio to the recombinant yeast cell of Boles, and thus to have obtained claimed specification wherein the yeast cell is capable of taking up at least about 5 g/l of arabinose in 24 hours, because the transport capacity of the AraT is an inherent property of it's structure. As said composition was known in the art at the time of the invention, this cannot be considered a special technical feature that would otherwise unify the groups.

Another special technical feature of the inventions listed as Groups I+ and III+ is the specific sequences recited therein. The inventions do not share a special technical feature, because no significant structural similarities can readily be ascertained among sequences. Without a shared special technical feature, the inventions lack unity with one another.

Groups I+ through III+ therefore lack unity under PCT Rule 13 because they do not share a same or corresponding special technical feature.